Tissue Resolved, Gene Structure Refined Equine Transcriptome

Mansour,T.A.*[1,2], Scott, E.Y.*[3], Finno, C.J.[1], Bellone, R.R.[4], Mienaltowski, M.J.[3], Penedo, M.C.[4], Ross, P.J.[3], Valberg, S.J.[5], Murray, J.D.[1,3], Brown, C.T[1].

[1] Department of Population Health and Reproduction, University of California, Davis, [2] Department of Clinical Pathology, College of Medicine, Mansoura University, Egypt [3] Department of Animal Science, University of California, Davis, [4] Veterinary Genetics Laboratory, University of California, Davis, [5] Large Animal Clinical Sciences, Michigan State University, College of Veterinary Medicine

*Both authors contributed equally to this manuscript

To be submitted to Genome Research

Abstract

Transcriptome interpretation relies on a good-quality reference transcriptome for accurate quantification of gene expression as well as functional analysis of single nucleotide polymorphisms (SNPs). The current annotation of the horse genome lacks the specificity and sensitivity necessary to distinguish differential isoform expression, untranslated region (UTR) usage and, to some extent, sensitive differential gene expression. We built an annotation pipeline for horse and used it to integrate 1.9 billion reads from multiple RNA-seq data sets into a new refined transcriptome. This equine transcriptome integrates eight different tissues from over fifty individuals and improves gene structure, isoform resolution while providing considerable tissue-specific information. We utilized four levels of transcript filtration in our pipeline, aimed at producing several transcriptome versions that are suitable for different downstream analyses. Our most refined transcriptome includes 36,876 genes and 76,125 isoforms, with 6474 candidate transcriptional loci novel to the equine transcriptome. We have employed a variety of descriptive statistics and figures that demonstrate the quality and content of the transcriptome. The equine transcriptomes that are provided by this pipeline show the best tissue-specific resolution of any equine transcriptome to date. We encourage the integration of further transcriptomes with our annotation pipeline.

Introduction

Transcriptomics is rapidly evolving from a focus on novel gene identification to resolving structural gene details. The transcriptomes of better-studied organisms, such as Drosophila, mouse and human have been updated to accommodate for this transition (Okazaki et al. 2002; Brown et al. 2014; Mele et al. 2015). However, for less well characterized animals, such as the horse, there is often only annotation of a single variant of a gene and no inclusion of multiple splice variants, UTR extensions and non-protein coding RNA. This lack of information can challenge subsequent differential gene expression analyses and single nucleotide polymorphism (SNP) functional annotation. There have been several attempts to improve the equine transcriptome with single tissue transcriptomes from lamellar tissue (Holl et al. 2015) or peripheral blood mononuclear cells (Pacholewska et al. 2015) and from pooled composites of various tissues (Coleman et al. 2010; Hestand et al. 2015), however a broader effort defining and integrating many tissue-specific transcriptomes and obtaining the library depth and strand information required to capture gene complexity is still needed.

Current gene annotation pipelines differ in their reliance on the reference genome to construct the transcriptome. Reference genomes across species vary in coverage and accuracy, with different approaches in their functional annotation reflecting the state of the reference genome. ENSEMBL and NCBI provide publically available annotations for several vertebrate genomes including horses (Yandell and Ence 2012). Both underlying annotation pipelines integrate homology search and *ab initio* prediction however accurate UTR prediction and isoform recognition require species-specific transcriptional evidence (Kitts 2003; Curwen et al. 2004). For this equine transcriptome, the transcriptional evidence provided by total RNA sequencing (RNA-seq) was the basis of our gene annotation. This approach permits more reliable discovery of novel genes and isoforms, extension of UTRs and the flexibility necessary to establish a balance between sensitivity and specificity of gene detection for downstream applications.

Our annotation integrates the benefits of increased depth in reads and strand-specificity, for some tissues, as well as using a range of tissues from many horses, which allows tissue-specific transcriptomes to be extracted. We have incorporated RNA-seq from a diverse set of 8 tissues ranging from the central nervous system (CNS), skin and skeletal muscle tissues in adults to the inner cell mass

(ICM) and trophectoderm (TE) in embryonic tissues (Table 1). The diversity in age, sex and tissue of the samples included in our assembly supply the equine transcriptome with its best spatiotemporal resolution and most complete gene UTR definition to date.

We recognize that availability of annotation criteria and integration of transcriptome data is paramount for systematically improving the equine transcriptome. Our goal is to encourage equine researchers to incorporate their transcriptomic data using our pipeline as the common annotation pipeline and our initial transcriptomes as a reference framework. We intend to continue improving equine gene annotation through better UTR definition, isoform splicing characterization and novel gene identification. The annotation presented in this paper will improve the gene structure definition in current databases and the accuracy of downstream analyses, including both differential gene expression analysis and SNP annotation in the horse.

Results

Overall Mapping Statistics and Gene Counts After Filtration

RNA-seq of 59 samples in 12 libraries from 8 different horse tissues provided 1917.7 million fragments and 364 Gb of reads. A summary of the library preparation, number of horses per library and total number of fragments and bases provided by each tissue library can be found in Table 1. The overall average mapping rate for Tophat2 was ~83% with concordance rates ranging from 29% to 89% (average 75%) for paired end libraries. Concordance rates seem to be affected by the type of library preparation, where polyA selected and strand-specific libraries have the best rates. Library specific mapping rates can be found in Supplementary Table 1. The initial Cufflinks assembly identified 117,019 genes/211,562 transcripts. After this initial analysis we applied four grades of filtration (Figure 1). Primary filtration of transcripts removed pre-mRNA fragments by eliminating single exon transcripts that were present within introns or overlapping with exons of other multi-exon transcripts. After primary filtration there were 75,102 genes/162,261 transcripts. The second filter was implemented to remove isoforms likely to be experimental artifacts by excluding low abundant transcripts with less than 5% of total expression for their locus. The remaining 114,830 transcripts represented 75,375 genes. In the third filter, non-coding transcripts that lack any supporting evidence from NCBI or ENSEMBL annotations, non-horse gene models ("Other RefSeq" and "TransMap RefGene" UCSC tracks) or *ab initio* predictions ("Augustus", "Geneid", "Genscan" and "N-SCAN" UCSC gene prediction tracks) were excluded. This third filtered version of the transcriptome has 76,323 transcripts in 37,062 genes. The last filter was for removing likely erroneous transcripts. The mtDNA in mammals is known for gene overlapping and polycistronic expression (Taanman 1999), permitting inaccurate prediction of mitochondrial transcripts by Cufflinks; we therefore excluded the mitochondrial contigs from our filtered assembly. Also, short transcripts less than 201bp (192 transcripts in 184 genes) were removed because they are more likely to represent repetitive sequences or incomplete gene fragments. Once erroneous transcripts were removed, our final refined version of the transcriptome contained 36,876 genes (76,125 transcripts) including 15,343 single exon transcripts, 8,808 two-exon transcripts, and 51,974 transcripts with three or more exons. A version of our refined transcriptome that is merged with the NCBI and ENSEMBL annotations, with redundant transcripts removed, is also available. This is the most comprehensive product of our pipeline and is valuable for differential gene expression analysis in tissues other than those provided in our assembly. Summary statistics including N50, number of genes and Mb and average length of fragment for all six versions of the transcriptome can be found in

Supplementary Table 2.

Comparison between Our Transcriptome and Currently Available Equine Transcriptomes

We performed a comparison between our transcriptome and gene models from NCBI, ENSEMBL and two published equine transcriptomes, that we refer to as Hestand (Hestand et al. 2015) and ISME (Pacholewska et al. 2015) (Table 2 and Supplementary Table 3). In our comparisons, transcripts sharing one or more splice junctions are considered similar but only those with identical intron chains are matching. The comparison reveals that the matching transcripts between our refined transcriptome and NCBI annotation are greater than 2.5-fold those matching the ENSEMBL annotation. However the highest number of matching transcripts occurred with the ISME transcriptome with 12,849 transcripts (Figure 2A). About 50% of the refined transcripts have a similar match in all the public transcriptomes. Evidence of improvements to the annotation of genes with a similar match to other assemblies can be found in genes such as *MUTYH*, where the three major isoforms annotated in humans (Plotz et al. 2012) are now distinguishable in the horse (Figure 2C). The gene *CYP7A1* is another example where a novel first exon has been annotated and extended in our version of the transcriptome (Finno et al. 2016) (Figure 2D). About 20% and 28% of the refined transcripts are novel when compared to NCBI and ENSEMBL annotations respectively. Combined, there are 22,641 transcripts in candidate novel loci. Our approach of applying four successive steps of filtration strictly qualifies our novel genes as transcripts with ORFs or exonic overlap with candidate gene models. Mainly, novel transcripts contained within introns of other genes were excluded to avoid the artifacts of retained intronic reads, common in rRNA depleted libraries. Using the NCBI as a reference for comparison, our novel transcripts from the refined transcriptome have no bias towards any particular chromosome after accounting for chromosome size (Supplementary Figure 1). In order to calculate the gene and isoform detectability of our transcriptome compared to current annotation, we calculated sensitivity and specificity (Burset and Guigo 1996) between our transcriptome and a reference and found that, using NCBI as the reference, our transcriptome had a 78.8% sensitivity and 23.8% specificity at the base level and a 32% sensitivity and 21.1% specificity at the locus level. Detailed pairwise assessment for all equine annotations can be found in Supplementary Table 4. We developed a statistic to assess the conflict between different assemblies, termed “complex loci”, which refer to the loci that represent one gene locus in one transcriptome and two or more gene loci in another. Our transcriptome has 1355 and 997 transcripts that were considered complex loci between our transcriptome and NCBI and ENSEMBL, respectively. The Hestand transcriptome, however, has less with 660 and 798 complex loci against the NCBI and ENSEMBL, respectively. The ISME transcriptome has substantially more, with 1546 and 1226 complex loci when compared to NCBI and ENSEMBL, respectively.

UTR extension

To test the effect of the new assembly on the UTRs of known genes, we identified the protein coding isoforms sharing the exact intron chain with NCBI isoforms, which yielded 9736 isoforms from 7419 genes. The difference in the total length of each transcript was then calculated and we found that we extended the length of 8899 isoforms (6817 genes) by 29.7 Mb in total. 831 isoforms (718 genes) lost 0.3 Mb in total with an average of 0.4 kb per isoform, while 6 isoforms did not change.

Gene and Isoform Distinctions between Tissue-Specific Transcriptomes

We selected genes with high expression in at least one tissue and substantial expression differences across tissues. A heatmap grouped genes that may be co-expressed as well as illustrating the relationship between the tissue-specific transcriptomes. As expected, the transcriptomes from the three central nervous system (CNS) tissues clustered together, as did the two embryonic tissues, with the skin and skeletal muscle furthest from these clusters (Figure 3A). Blocks of genes showing uniquely high expression in a given tissue were further annotated with NCBI gene names and then summarized with Panther biological processes annotations to reveal gene expression patterns relating to the specific function of the tissue. The top tissue-specific hits are reported in the text below, with the full Panther annotation tables detailed in Supplementary Table 5. The CNS cluster contained overrepresented processes regarding brain function and development: nervous system development ($p=2.10E^{-03}$, fold-enrichment=7.12) and its regulation ($p=1.43E^{-03}$, fold-enrichment=15.98) and , neurogenesis ($p=9.36E^{-05}$, fold-enrichment=7.73). The retina contained processes consisting of photoreception and development: phototransduction ($p=3.52E^{-08}$, fold-enrichment=80.75), photoreceptor cell maintenance ($p=4.53E^{-08}$, fold-enrichment=43.84), visual perception ($p=3.69E^{-18}$, fold-enrichment=37.64) and visual perception ($p=3.69E^{-18}$, fold-enrichment=37.64). The skeletal muscle encompassed genes pertaining to muscle physiology and regulation: transition between slow and fast fiber ($p=6.25E^{-03}$, fold-enrichment >100), skeletal muscle contraction ($p=2.03E^{-07}$, fold-enrichment > 100), sarcomere organization ($p=1.11E^{-06}$, fold-enrichment=80.89), and myofibril assembly ($p=6.05E^{-11}$, fold-enrichment=72.8). The embryonic tissues have the most general processes assigned to their distinct clusters: translation ($p=1.15E^{-11}$, fold-enrichment=16.35) and peptide biosynthetic processes ($p=1.95E^{-11}$, fold-enrichment=15.8). And finally, the skin consisted of processes concerning epithelial organization and production: intermediate filament organization ($p=1.69E^{-07}$, fold-enrichment > 100), skin development ($p=1.89E^{-09}$, fold-enrichment=22.49), epidermis development ($p=8.02E^{-04}$, fold-enrichment=13.96) and hair follicle morphogenesis ($p=6.23E^{-03}$, fold-enrichment=55.85).

When attention is given to the isoforms showing unique presence or sole absence in a tissue, the cerebellum and retina possess the most isoforms that are uniquely present, with the retina also containing the largest amount of solely absent isoforms (Figure 3B). The transcripts solely absent from the retina pertain mainly to positive regulation of DNA replication ($p=2.49E^{-03}$, fold-enrichment=3.53) and anatomical structure development ($p=2.72E^{-19}$, fold-enrichment=1.48). Utility of these isoforms, in terms of expression in TPM, is strongest in the skin, retina, skeletal muscle and to a small extent the cerebellum (Figure 3B). Despite these differences in unique isoforms, multi-exons transcripts and multi-transcript loci, the splicing rate across tissues, as calculated by Cuffcompare (Trapnell et al. 2013), ranges from 1.7 to 1.9 (Table 3).

Nuclear coding versus mitochondrial encoded genes were parsed out per tissue to determine how much of the sequencing resources are allocated to genes of the mitochondria (Figure 3C), with the conclusion that the brainstem, spinal cord, embryonic tissues and skeletal muscle exhibit the largest proportions of transcriptional output devoted to mitochondrial genes.

Classification and Annotation of Novel Genes

Classification of our novel genes was necessary to better represent how our transcriptome contributed to novel gene identification. Three categories of novel genes based upon the supportive evidence within and across species were made with each successive category being less supported by equine or orthologous gene models. Our first category of novel genes hosts those missing from NCBI and/or ENSEMBL annotation, but supported by either NCBI, ENSEMBLE, Hestand or ISME

annotations (Category I). The second group of novel genes were novel to all public equine annotations, but conserved by means of orthologous gene similarity or supported by possible gene prediction (Category II). The third category of novel genes were unsupported by any candidate gene models, but had an ORF (Category III). Category I has a total of 8459 transcripts, with 2/3 of these transcripts originating from the ENSEMBL annotation (Figure 4A). Another 1849 transcripts in this category are novel to both NCBI and ENSEMBL annotations, yet supported by Hestand or ISME annotations. Homology with the SWISS-prot database identified at least one significant ($p<1E-10$) hit for almost half the transcripts in this Category I (Supplementary Table 6). Category I novel genes not only feature independent evidence for gene models missed by NCBI and/or ENSEMBL annotations, but they are also enriched for improved gene models. The second category has 7494 transcripts that – unless on the opposite strand - do not overlap with known gene models in public annotations. Annotation of these transcripts was performed partially by testing overlap with non-horse gene models and also by homology search, resulting in 16% of the transcripts showing significant hits against the SWISS-prot database (Supplementary Table 4). The third category of novel genes includes 6687 transcripts with an ORF as the only functional support for these transcripts. The first category of novel genes shows the most diverse distribution of exon numbers comprising the genes (ranging from 1 to 28), whereas the unsupported genes contained mainly single exon genes (Figure 4B). The expression of the three categories of novel transcripts in relation to the number of exons comprising these genes demonstrates not only the presence of these genes, but also the expression patterns in relation to tissue specificity (Figure 4C). Supported novel genes (Category I) had the highest expression in the cerebellum and spinal cord, which consisted mainly of genes with up to three exons (Figure 4C). However, when looking at only the second category of novel genes, the embryo contributed the highest expression of novel transcripts, which mainly consisted of transcripts with two exons (Figure 4C). Category III novel transcripts mainly consisted of single exon transcripts and showed similarly low expression across all tissues (Figure 4C).

Discussion

Using RNA-seq from fifty-nine horses across eight tissues has allowed us to capture transcriptome complexity and provide spatial resolution in terms of tissue-specificity in manner that exceeds any current equine annotations. Our descriptive statistics and accessible pipeline make this project open to modifications and further integration of transcriptomes.

RNA-seq based transcriptomes are prone to false inflation of total transcript and gene numbers for several reasons culminating in misassembly of transcripts. Technical limitations such as limited sequencing read length and amplification errors, false splicing events, and assembler deficiencies are among several reasons of misassembly. Pervasive transcription is another predominant source of such inflation (Bertone et al. 2004; Schadt et al. 2004; Khaitovich et al. 2006). Some types of sequencing libraries increase the problem as well; for example rRNA depletion inflates the assembly with primary transcripts and false isoforms exhibiting intronic retention (Sultan et al. 2014). Our pipeline takes these factors into account and runs unguided by a reference transcriptome with several transcript filtration steps aimed to reduce inclusion of inaccurate transcripts, while retaining the sensitivity for novel transcript detection. The effect of this procedure can be seen by comparing the gene numbers between our initial unfiltered and final filtered transcriptomes, where gene inflation was reduced by 68% (Table 2) and our final refined transcriptome contained 36,876 genes and 76,125 isoforms.

Although not indicative of transcriptome quality, we calculated specificity, as a measure of difference between our transcriptome and other annotations, and sensitivity, which indicates how our

transcriptome covers another annotation. These parameters demonstrate that our aggressive filtering does sacrifice sensitivity at the locus level only by a margin of approximately 5%, and increases our specificity often by more than 10%, relative to NCBI and ENSEMBL (Supplementary Table 3). We have a comparable sensitivity to the Hestand transcriptome. However, the numbers of unstranded and multi-exon transcripts in the Hestand transcriptome relative to our refined version serve as the more informative statistics. We have approximately six fold less unstranded transcripts and more than double the multi-exon transcripts (Table 2). Regarding how our transcriptome compares to the other recent equine ISME assembly (Pacholewska et al. 2015), which is ENSEMBL annotation guided, we have three times more matching transcripts to the ISME assembly than to the ENSEMBL annotation itself (Figure 2A), suggesting significant improvement made by ISME annotation. However their improvements are impaired by false inflation in the number of genes identified due to presenting most of the transcripts in two copies representing the forward and reverse strands. This, as well as the strict filtering done by Hestand et al (2015), could be why our transcriptome statistics follow trends seen in the Hestand transcriptome more closely. Hestand et al (2015) also observed a bias towards single exon genes, which represented approximately 55% of their whole transcriptome (Hestand et al. 2015). However, after our filtration steps, our annotation has single exon transcripts representing only 20% of our whole transcriptome. We also illustrate their function by assessment of their transcription in different tissues relative to the number of exons, which indicates that a majority of transcripts consisting of one or two exons occupy a large proportion of transcriptional output (Supplementary Figure 2A), with genes containing three or more exons having a similar distribution of expression across tissues (Supplementary Figure 2B). Our statistic, complex loci, also highlights a level of sensitivity as well as area for further investigation in our transcriptome. We have more than two times more complex loci, using NCBI as a reference, than Hestand, with the inflated ISME complex loci numbers being attributable to the double reporting of their transcripts. Awareness of these complex loci allows for discovery of novel genes with complete overlap with known reference genes as well as further refinement of transcriptome-wide gene structure and lncRNA identification. A pipeline to appropriately process these loci has yet to be established, with most loci being handled on a case-to-case basis. The evaluation of these alternative descriptive statistics emphasize the deficiencies of commonly used transcriptome statistics and demonstrates how they are prone to concealing improvements in gene structure and novel gene identification, while attempting to draw comparisons with transcriptomes that have alternative pipeline-specific limitations such as gene number inflation or failure to remove erroneous transcripts.

Accurate identification of UTRs is often difficult for *ab initio* programs and requires sufficient support of transcription evidence. Transcripts from the CNS in the *Drosophila* (Smibert et al. 2012), developing *Caenorhabditis elegans* (Mangone et al. 2010) and zebrafish (Ulitsky et al. 2012) show extended 3'UTRs, often associated with alternative polyadenylation or cleavage sites of pre-mRNA. Because of the large proportion of CNS, as well as embryonic, tissues used for our transcriptome, we were able to expand UTR annotation by an average of 3.3 kb per transcript. Further improvements to this transcriptome would include providing tissue-specific UTR lengths and allowing for a more clear depiction of differences in gene structure between tissues. The improved UTR structure provided by our transcriptome has already shown its utility in the horse community by defining isoform and gene boundaries of *MUTYH* and *TOE1* (Scott et al. 2016) as well as providing an alternative start exon for *CYP7A1* (Finno et al. 2016)

Our final version of the transcriptome represents a collection of genes provided by eight distinct tissue-specific RNA-seq libraries. This feature allows us to extract inherent tissue-specific characteristics from the transcriptome regarding gene expression, mitochondrial gene expression and

isoform usage between the tissues. When Spearman correlation was applied to the tissues and the genes, clusters of genes corresponding and exclusive to the inherent functions of the tissue were revealed (Figure 3A). This indicates that above any noise that was created by the different methods of cDNA library preparation or sequencing, a transcriptomic signature with biological relevance to the tissue can be deciphered. Additional to the nuclear gene expression signature, the amount of transcription occurring from the mitochondrial chromosome can stipulate how much of the sequencing resources are being allocated to mitochondrial originated genes. Across the eight tissues, one would expect the tissue with the largest numbers of mitochondria would have the largest proportion of transcriptional output allocated to mitochondrial genes, as seen with the human transcriptome (Mele et al. 2015). Our data demonstrates these trends in the brainstem, skeletal muscle and spinal cord, however the cerebellum and retina do show an unexpectedly low mitochondrial gene load (Figure 3C). Further research establishing the relationship between the amount of mitochondria processed in a sample for RNA-seq and the resulting mitochondrial expression loads would be beneficial to understanding how much of the transcriptional output is dominated by an individual mitochondrion. This information would also allow researchers to extract more information pertaining to mitochondria from RNA-seq data.

In addition, to gene expression patterns, isoform usage further distinguished tissue specificity. In humans, these eight tissues exhibit distinct transcriptome profiles, with the CNS displaying the most complex splicing activity (Mele et al. 2015), along with embryonic tissues, as seen in *Drosophila* (Brown et al. 2014). The differentiation status of the cells has been suggested to increase the transcriptome complexity in terms of splicoforms (Wu et al. 2010), which is a trend seen between our embryonic ICM and TE, where the embryonic ICM exhibits more uniquely present isoforms than the embryonic TE (Figure 3B). On a similar note, the skin and retina demonstrate large numbers of tissue-specific transcripts with concordant expression, suggesting the influence of these unique isoforms on tissue-specific functions. Although the skeletal muscle has a relatively low amount of unique tissue-specific isoforms, it shows utility of these isoforms with relatively high cumulative expression values (Figure 3B) as well as comparable splicing rates (Table 3). In agreement with previous retina transcriptome work (Farkas et al. 2013), the retina displayed the most unique splicing with 2962 uniquely present isoforms (876 genes) and 8202 uniquely absent isoforms (5256 genes), along with a relatively high cumulative expression, highlighting its transcriptome specificity in the form of splicing. Reinforcing this specificity, Panther annotations of these unique isoforms are related to phototransduction (p=6.62E-11) and photoreceptor cell maintenance (p=2.48E-9). Three tissues: retina, skin, and embryo, had shorter read lengths and were not prepared as stranded libraries and thus these data may be artificially understated in terms of transcriptome complexity.

We identified 7494 candidate novel transcripts. These novel transcripts are selected based on having no overlap with genes in current equine annotation and authenticated by their protein-coding ability and overlap with aligned non-horse genes or *ab initio* gene predictions. Additionally, our novel transcripts have a diversity of coding exons with a particular expression bias towards the embryonic tissues used in this assembly, in which a majority of these novel transcripts contain two exons (Figure 4C). The Category I novel transcripts highlight the deficient equine ENSEMBL annotation, the need to pool the databases to get the most transcriptome coverage and the ability of our transcriptome to capture the potentially rare novel gene models (Figure 4A). Despite the ORF requirement for Category III novel transcripts, there is an obvious enrichment for single exon transcripts and a marked reduction of total transcription level (Figure 4C), which is indicative of non-coding RNA. Our novel gene analysis also produced a category of novel transcripts that were removed due to not having ORFs and were presumed to represent noisy transcription relating to primary transcripts, repetitive elements, sequencing errors and genome-based errors. The collection of Category III and these excluded novel transcripts may represent

a repository of non-annotated non-coding RNA, which is an area that needs further annotation in the horse genome.

This transcriptome assembly pipeline not only produces flexible incorporation of additional transcriptomes, it also provides several products regarding levels of transcript filtering and appropriateness for downstream analysis. The different extractable transcriptome versions include transcriptomes after each individual filter, a transcriptome merged with NCBI and ENSEMBL annotations to achieve breadth not covered by our tissues and the final refined transcriptome containing only genes with complete ORFs and genes aligning with other non-horse genes or *ab initio* gene predictions. These transcriptomes as well as the pipeline to make each of these transcriptomes are publically available on our GitHub repository. By making the workflow public and easy to execute and manipulate, we aim to expand the spectrum of tissues embodying this transcriptome and eliminate biases in annotated genes and thus downstream differential gene analysis. Increasing and providing the option for tissue-specific transcriptomes, allows for more targeted and refined usage of a certain tissue's transcriptome for differential gene analysis, resulting in downstream analysis of more significant differentially expressed genes. As stated in our overall goals of this project, we have provided a framework for further improving the equine transcriptome and produced an equine transcriptome that expands on current equine annotations in the manner of UTR extension, isoform detection and novel gene identification.

Methods

RNA-seq library preparation

A total of twelve RNA-seq libraries in 8 tissues from 59 individuals (20 female, 27 males and 12 embryos) were used to prepare our transcriptome. The brainstem, spinal cord and cerebellum were strand-specific 100 bp paired-end (PE) libraries. The skeletal muscle tissues were strand-specific PE125 bp libraries. A subset of the embryo ICM (3 samples) and TE (3 samples) were unstranded PE100 bp libraries, the other subset (3 ICM and 3 TE) consisted of single end (SE) 100 bp reads. The retina RNA-seq libraries were unstranded SE80 bp libraries. And the skin libraries were all unstranded and consisted of PE80 bp, SE80bp and SE95 bp reads. The brainstem, spinal cord and cerebellum RNA libraries were all rRNA-depleted, the skin, retina and skeletal muscle libraries were poly-A captured. The embryonic libraries were neither poly-A selected nor rRNA depleted, they were prepared with the Ovation® RNA-seq System V2 (NuGEN, San Carlos, CA, USA), which aims to amplify mRNA as well as non-polyadenylated transcripts. Table 1 summarizes the tissue-specific RNA-seq library parameters.

Trimming and Mapping of Reads

The Illumina adaptors as well as the reads were trimmed with the sliding window quality trimmer Trimmomatic (Bolger et al. 2014) with a window size of 4 and a softer quality threshold of 2 (Macmanes 2014). Mapping of the trimmed reads was done with Tophat2 (Kim et al. 2013) to EquCab2.0, 2007 (ftp://hgdownload.cse.ucsc.edu/goldenPath/equCab2). Cufflinks (Trapnell et al. 2013) was used to assemble transcripts from the aligned RNA-Seq reads. Two cerebellar samples failed assembly due to computation limitations (8 CPUs, 250 Gb RAM and 7 days) and required digital normalization (Brown 2013) to 200X coverage before mapping with Tophat2.

Filtering Transcripts

Four categories of filters were used to remove likely pre-mRNA and artifactual transfrags, as summarized in Figure 1, resulting in six versions of the transcriptome. Primary transcript filtration was done using Cuffcompare (Trapnell et al. 2013) between our assembly and a version of our assembly containing only multi-exon transcripts and removing transcripts overlapping with intronic regions and with class codes "i", "e" and "o". The input trimmed RNA-seq reads were then back-mapped to the pre-mRNA free transcriptome using the quasi-mapping based software package Salmon (Rob Patro 2015). While back-mapping, a second filtration step was implemented: low abundance transcripts in every locus were excluded with the lower threshold of a TPM (normalized read count standing for transcripts per million) less than 5% of the total TPM per locus. For the third filter, Transdecoder (Haas et al. 2013) was used to predict the ORFs and Cuffcompare (Trapnell et al. 2013) to determine any exonic overlap with any candidate gene locus using the class codes "j", "o", "x" and "c". In the Transdecoder analysis, the longest open reading frames were extracted as well as any sequences having significant homology to the Pfam and Swissprot protein databases. Finally, the removal of likely erroneous mitochondrial and short transcripts was done by a homemade script.

Transcriptome Comparisons

Comparisons of our refined transcriptome to the four public horse transcriptomes were done using Cuffcompare (Trapnell et al. 2013). In any pairwise comparison, two transcripts are considered matching if they have the exact intron chain, despite differing terminal exons (class code "="). If the transcripts are not matching but sharing one or more splice junctions (class code "j"), these would be considered similar transcripts. A transcript is considered novel if it does not overlap with any gene model in the 2nd reference assembly (class code "u"). All other class codes including any kind of overlap with a reference annotation on the opposite strand were considered as "other". For more detailed descriptions of the class codes provided by Cuffcompare, please see their manual (Trapnell et al. 2013). Complex loci were flagged if a gene model of one assembly overlapped with 2 different gene models in the other assembly. Sensitivity and specificity relative to a given reference transcriptome were calculated per base, intron and locus for each transcriptome and reference combination as described by Burset and Guigó (Burset and Guigo 1996).

Novel Gene Prediction

Any transcript in our final *refined* transcriptome is defined as novel if it does not overlap with a gene model in at least one of the two public equine assemblies, NCBI and ENSEMBL (Cuffcompare class code "u"). Transcripts considered novel were divided into three groups according to the degree of supportive evidence. Transcripts novel to either the NCBI or ENSEMBL assemblies with transcriptional supportive evidence from the other or any other public assembly (Hestand et al. ; Pacholewska et al. 2015) were in the first category of novel transcripts. Supportive evidence is defined as any overlapping with exon sequence (Cuffcompare class code "=","j","o","x" or "c"). The second and third categories of novel transcripts required that the transcript be absent in all current equine transcriptomes. Transcripts in the second category have supportive evidences in non-horse alignment gene models or *ab initio* gene prediction tracks from the UCSC genome browser. The third category of novel transcripts included

transcripts that lack such evidence but have ORFs.

Tissue-Specific Characterization of The Transcriptome

Tissue-specific transcriptomes were generated by back-mapping the input trimmed RNA-seq reads with Salmon (Rob Patro 2015) to the refined version of the transcriptome to obtain expression information on a tissue-specific level. A transcript is considered expressed in a given tissue if it has a TPM more than 5% of the total TPM per locus calculated from the tissue specific libraries only. Biological processes identified within the tissue-specific gene blocks were annotated with Panther (Mi et al. 2013) and reported if the p-values were below the Bonferroni-corrected threshold (5% experiment-wide).

UCSC Track hubs

Gene Annotation Format (GTF) files were converted into the binary bigbed files (Kent et al. 2010) using UCSC kintUtils (https://github.com/ENCODE-DCC/kentUtils). The track hub directory structure was designed as recommend by UCSC genome browser (Raney et al. 2014). Tracks were constructed using "bigBed 12" format and multiple libraries of the same tissue were organized in composite tracks. The hub files are hosted on a github server as a part of the horse_trans repository (https://github.com/dib-lab/horse_trans).

Data Access

Scripts for the pipeline as well as the GTF files for the transcriptome can be found at: https://github.com/dib-lab/horse_trans. (This repository is archived by Zenodo at 10.5281/zenodo.56934)

References

Bellone RR, Holl H, Setaluri V, Devi S, Maddodi N, Archer S, Sandmeyer L, Ludwig A, Foerster D, Pruvost M et al. 2013. Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. *PLoS One* **8**: e78280.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242-2246.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Brown C, Howe, A., Zhang, Q., Pyrkosz, A.B. & Brom, T.H. 2013. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**: 393-399.

Burset M, Guigo R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34**: 353-367.

Coleman SJ, Zeng Z, Wang K, Luo S, Khrebtukova I, Mienaltowski MJ, Schroth GP, Liu J, MacLeod JN. 2010. Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Anim Genet* **41 Suppl 2**: 121-130.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14**: 942-950.

Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA. 2013. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC Genomics* **14**: 486.

Finno CJ, Bordbari M, Monsour T, Bannasch DL, Mickelson DB, Valberg SJ. 2016. Spinal cord transcriptome profiling of equine vitamin E deficient neuroaxonal dystrophy identifies dysregulation of cholesterol homeostasis with upregulation of liver X receptor target genes. *In Review, Free Rad Biol Med*.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494-1512.

Hestand MS, Kalbfleisch TS, Coleman SJ, Zeng Z, Liu J, Orlando L, MacLeod JN. 2015. Annotation of the Protein Coding Regions of the Equine Genome. *PLoS One* **10**: e0124375.

Holl HM, Brooks SA, Archer S, Brown K, Malvick J, Penedo MC, Bellone RR. 2016. Variant in the RFWD3 gene associated with PATN1, a modifier of leopard complex spotting. *Anim Genet* **47**: 91-101.

Holl HM, Gao S, Fei Z, Andrews C, Brooks SA. 2015. Generation of a de novo transcriptome from equine lamellar tissue. *BMC Genomics* **16**: 739.

Iqbal K, Chitwood JL, Meyers-Brown GA, Roser JF, Ross PJ. 2014. RNA-seq transcriptome profiling of equine inner cell mass and trophectoderm. *Biol Reprod* **90**: 61.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, Paabo S. 2006. Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genet* **2**: e171.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kitts P. 2003. The NCBI Handbook: Chapter 14. Genome Assembly and Annotation Process. In *National Center for Biotechnology*. McEntyre, J. & Ostell, J.

Macmanes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front Genet* **5**: 13.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V et al. 2010. The landscape of C. elegans 3'UTRs. *Science* **329**: 432-435.

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**: 660-665.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8**: 1551-1566.

Okazaki Y Furuno M Kasukawa T Adachi J Bono H Kondo S Nikaido I Osato N Saito R Suzuki H et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-573.

Pacholewska A, Drogemuller M, Klukowska-Rotzler J, Lanz S, Hamza E, Dermitzakis ET, Marti E, Gerber V, Leeb T, Jagannathan V. 2015. The transcriptome of equine peripheral blood mononuclear cells. *PLoS One* **10**: e0122011.

Plotz G, Casper M, Raedle J, Hinrichsen I, Heckel V, Brieger A, Trojan J, Zeuzem S. 2012. MUTYH gene expression and alternative splicing in controls and polyposis patients. *Hum Mutat* **33**: 1067-1074.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**: 1003-1005.

Rob Patro GD, Carl Kingsford. 2015. Accurate, fast, and model-aware transcript expression quanitification with Salmon. *bioRxiv* doi:http://dx.doi.org/10.1101/021592.

Schadt EE, Edwards SW, GuhaThakurta D, Holder D, Ying L, Svetnik V, Leonardson A, Hart KW, Russell A, Li G et al. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol* **5**: R73.

Scott EY, Penedo MC, Murray JD, Finno CJ. 2016. Defining Trends in Global Gene Expression in Arabian Horses with Cerebellar Abiotrophy. *In Review, Cerebellum*.

Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB et al. 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* **1**: 277-289.

Sultan M, Amstislavskiy V, Risch T, Schuette M, Dokel S, Ralser M, Balzereit D, Lehrach H, Yaspo ML. 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**: 675.

Taanman JW. 1999. The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta* **1410**: 103-123.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46-53.

Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP. 2012. Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**: 2054-2066.

Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, Hutchison S, Raha D, Egholm M, Lin H, Weissman S et al. 2010. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc Natl Acad Sci U S A* **107**: 5254-5259.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329-342.

Tables

Table 1. Sample and library preparations used as input for our equine transcriptome

Tissue	Library Preparation	Library Characteristics	#Samples	#Frag(M)	#bp(Gb)	Reference
BrainStem	RiboRNA-depleted	PE100bp, stranded	8*	166.73	33.68	Finno et al, 2016
Cerebellum	RiboRNA-depleted	PE100bp, stranded	12	411.48	82.3	Scott et al, 2016
Muscle	Poly-A capture	PE125bp, stranded	12	301.94	76.08	
Retina	Poly-A captured	PE80bp unstranded	2	20.3	3.28	Bellone et al, 2013
SpinalCord	RiboRNA-depleted	PE100bp, stranded	16*	403	81.4	Finno et al, 2016
Skin	Poly-A captured	PE80bp, unstranded	2	18.54	3	Holl et al, 2016
	Poly-A captured	SE80bp, unstranded	2	16.57	1.34	Holl et al, 2016
	Poly-A captured	SE95bp unstranded	3	105.51	10.02	Bellone et al, 2013
Embryo ICM	Ovation RNA-seq	PE100bp, unstranded	3	126.32	25.26	Iqbal et al, 2014
	Ovation RNA-seq	SE100bp, unstranded	3	115.21	11.52	Iqbal et al, 2014
Embryo TE	Ovation RNA-seq	PE100bp, unstranded	3	129.84	25.96	Iqbal et al, 2014
	Ovation RNA-seq	SE100bp, unstranded	3	102.26	10.23	Iqbal et al, 2014
Total			69	1917.7	364.07	

*Seven individuals had both brainstem and spinal cord tissue collected from them. Seven of the skin samples were taken from 5 individuals and one individual had both retina and skin sampled, bringing our total number of individuals to 59.

Table 2. Comparison of current public equine annotations to six versions of our transcriptome (bolded and outline in red) in terms of gene numbers and composition

	Unfiltered	**Mature**	**High-exp**	**Supported**	**Refined**	**Merged**	Hestand	ISME	NCBI	ENSEMBL
Genes (super-loci)	117019	75102	75375	37062	36876	47760	56495	42654	24342	26962
Transcripts	211562	162261	114830	76323	76125	121997	68594	285538	43417	29196
Multi-transcript loci	17136	15430	14602	14511	14505	17835	8465	23833	7257	1592
Multi-exon transcripts	108985	108985	61570	60839	60782	97654	30949	259556	39272	19805
Redundant transcripts	0	0	0	0	0	2	3	12578	141	79
Unstranded transcripts	46928	35881	35872	6676	6618	5705	37673	4732	0	0
Single exon transcripts	102577	53276	53260	15484	15343	24341	37642	13404	3723	6862
Two exons transcripts	11114	11114	9449	8857	8808	11350	4410	13092	2425	2972
Many exons transcripts	97871	97871	52121	51982	51974	86306	26542	259042	37269	19362

Table 3. Tissue-specific splicing rate as calculated by Cuffcompare, with relevant number of multi-exonic transcripts and multi-transcript loci per tissue.

	Embryo ICM	Embryo TE	Skin	Brainstem	Cerebellum	Retina	Spinal cord	Muscle
Genes	33998	32050	30003	34792	36139	26733	34980	29549
transcripts	57400	54424	51995	62993	66364	47095	66001	52000
multi-exon transcripts	44069	42433	42432	49346	51640	39420	52175	42483
multi-transcript loci	11938	11461	11797	13066	13334	10866	13352	11560
Splicing rate	1.7	1.7	1.7	1.8	1.8	1.8	1.9	1.8

Figures

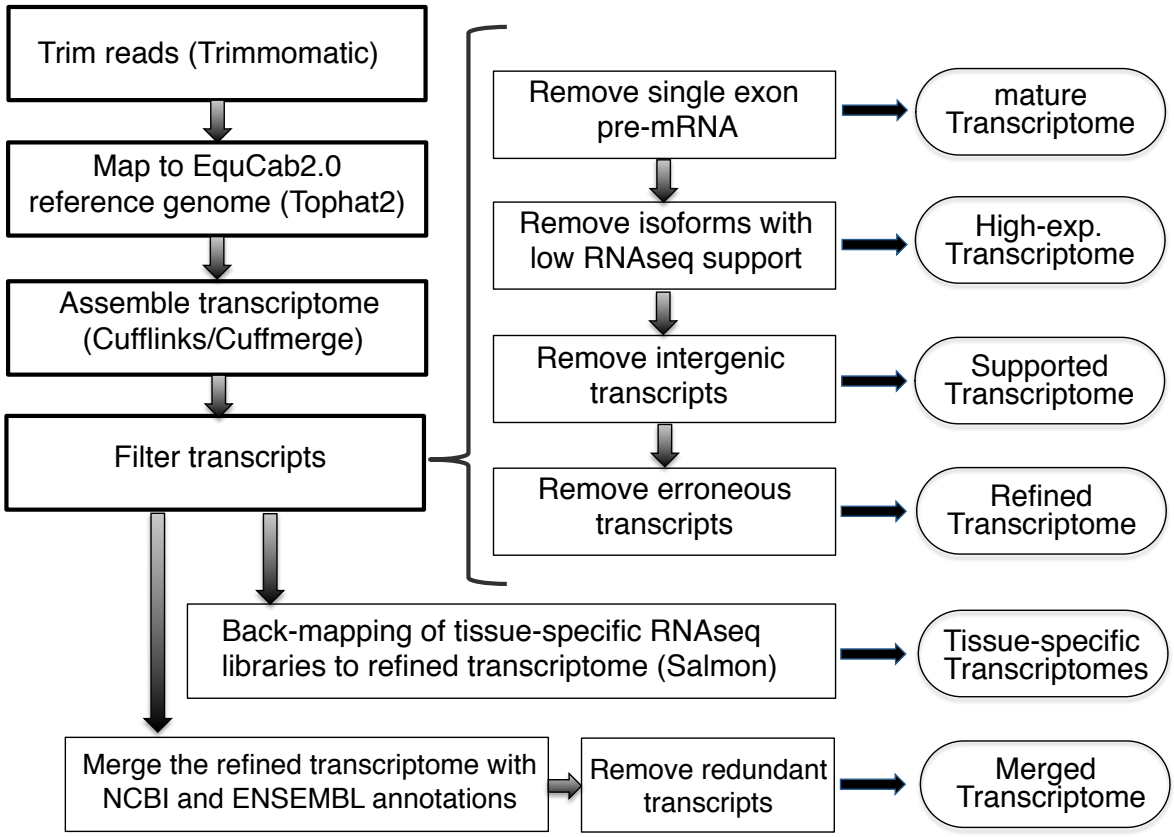
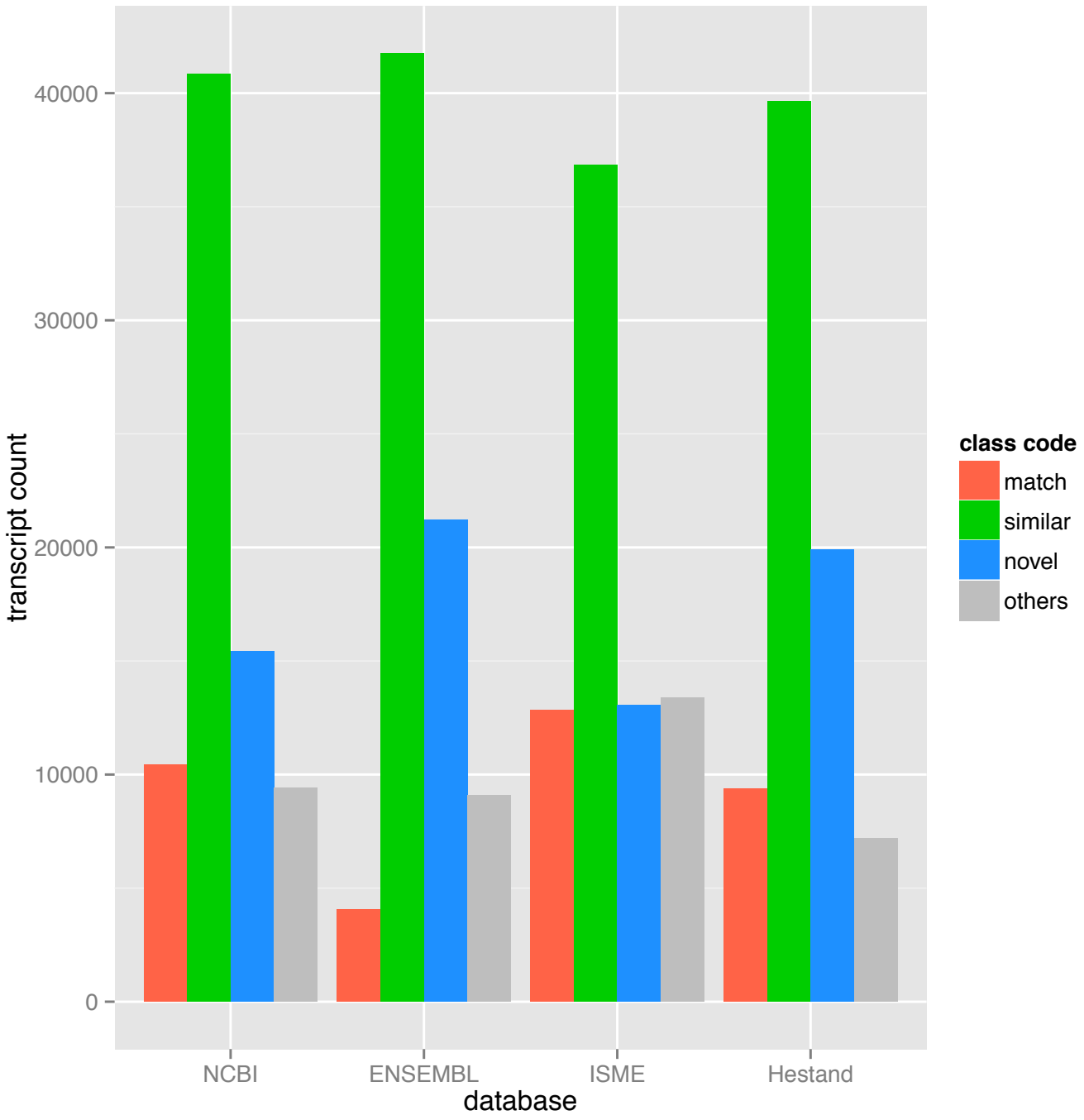
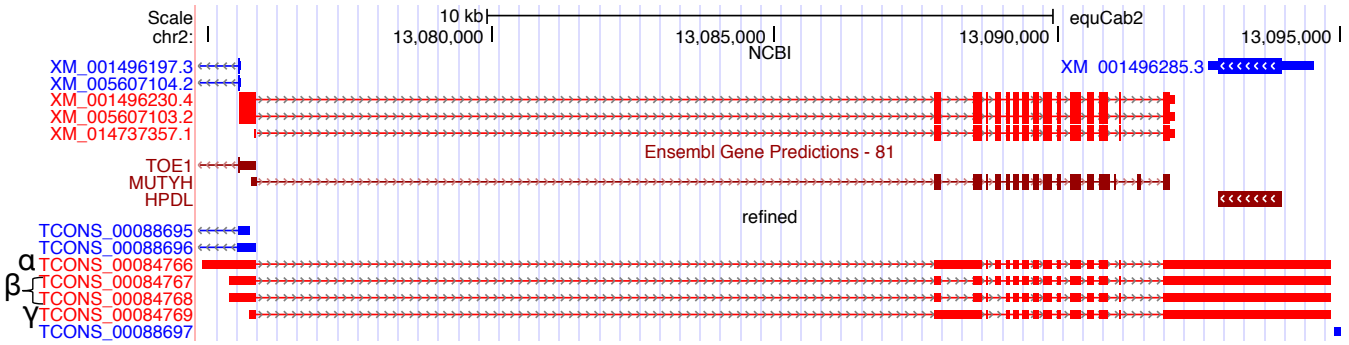


Figure 1. An outline of the workflow used to generate each version of the transcriptome. Transcriptome products are in ovals. Programs used to perform various steps are indicated in parentheses. All transcriptomes are publically available.

A)


40000
30000
20000
10000
0
transcript count
NCBI
ENSEMBL
ISME
Hestand
database
class code
match
similar
novel
others

B)



C)

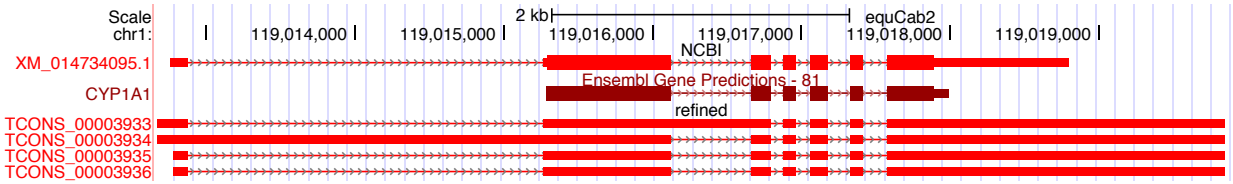
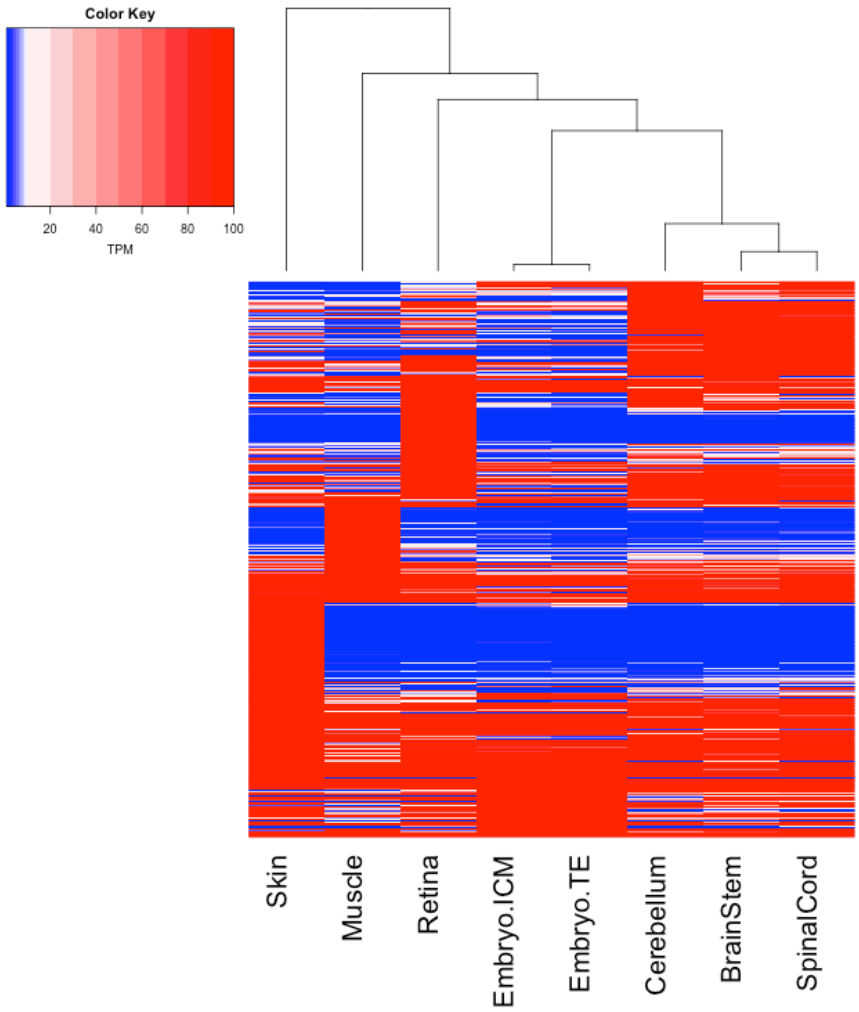
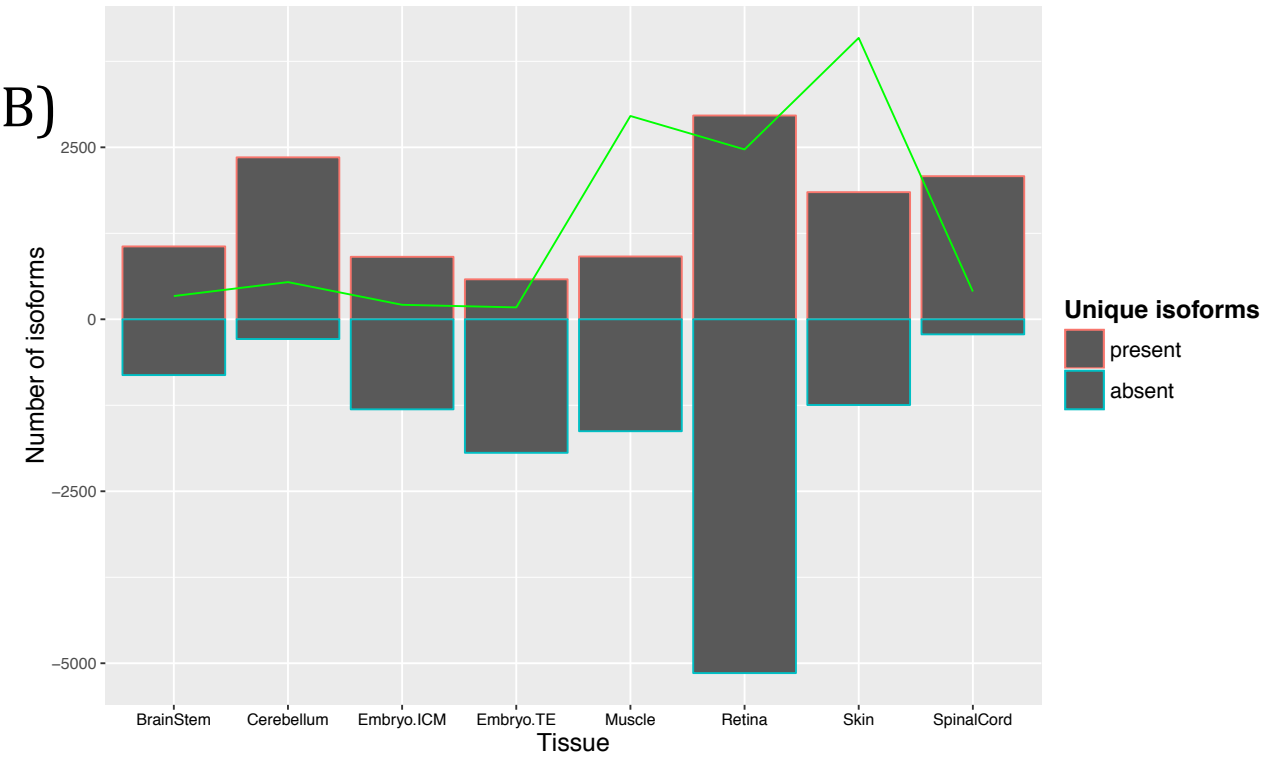


Figure 2. Comparison of our refined transcriptome to current equine annotations (A). The annotation of *MUTYH* in the refined version of the transcriptome shows the addition of several isoforms, α, β, and γ, as seen in the human, of *MUTYH* (B). The gene annotation of *CYP7A1* in the refined transcriptome also shows the inclusion of an extended alternative first exon not seen in other species (C).

A)


Color Key
20
40
60
80
100
TPM
Skin
Muscle
Retina
Embryo.ICM
Embryo.TE
Cerebellum
BrainStem
SpinalCord

B)


2500
0
−2500
−5000
Number of isoforms
BrainStem
Cerebellum
Embryo.ICM
Embryo.TE
Muscle
Retina
Skin
SpinalCord
Tissue
Unique isoforms
present
absent

C)

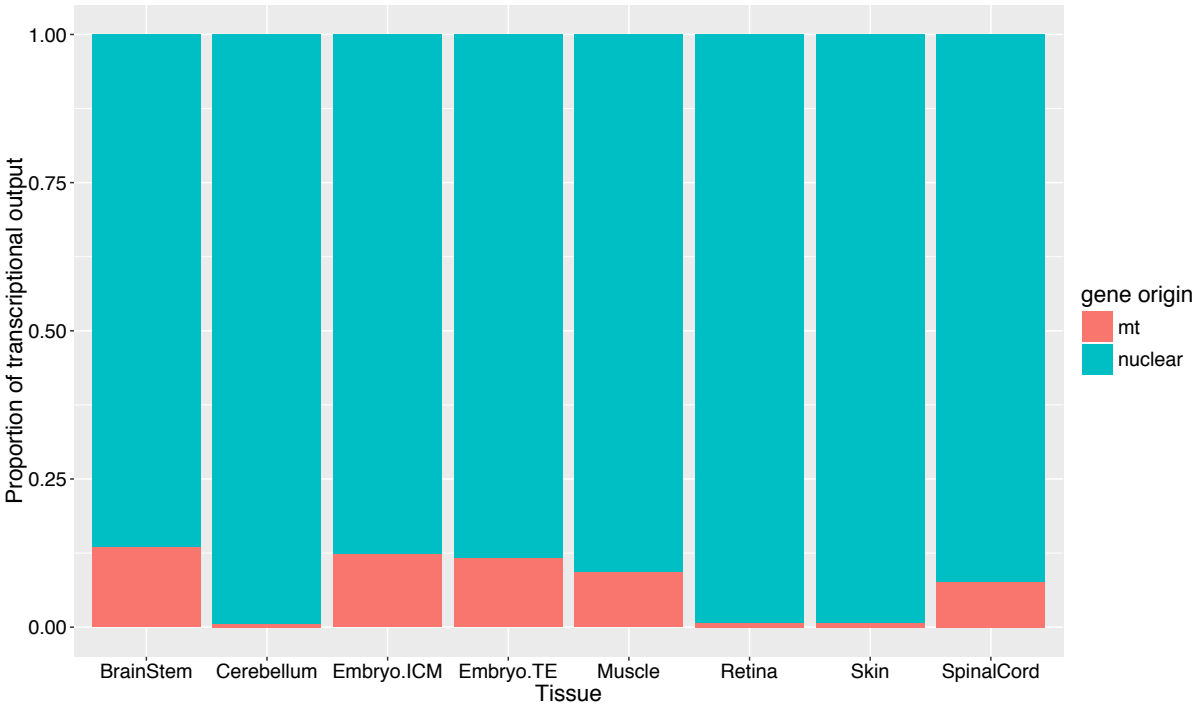
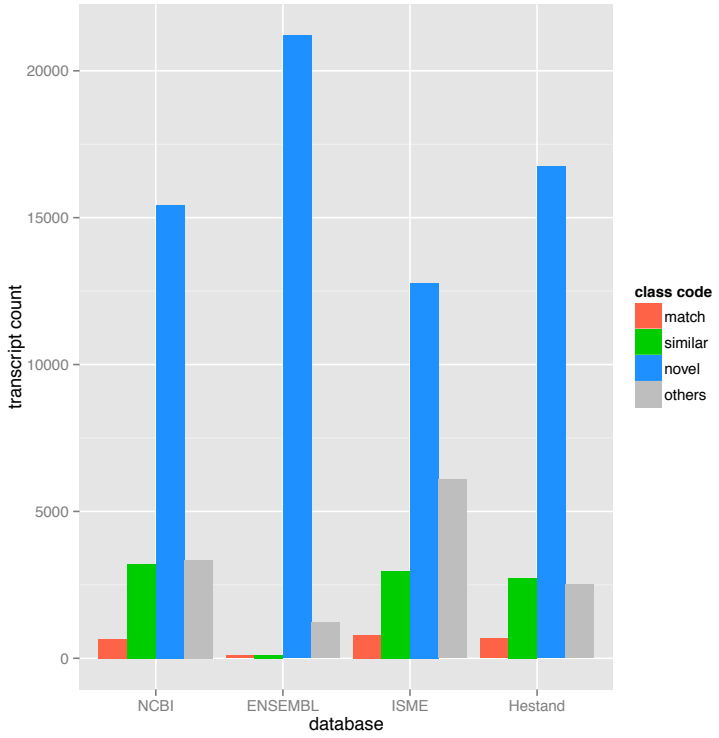
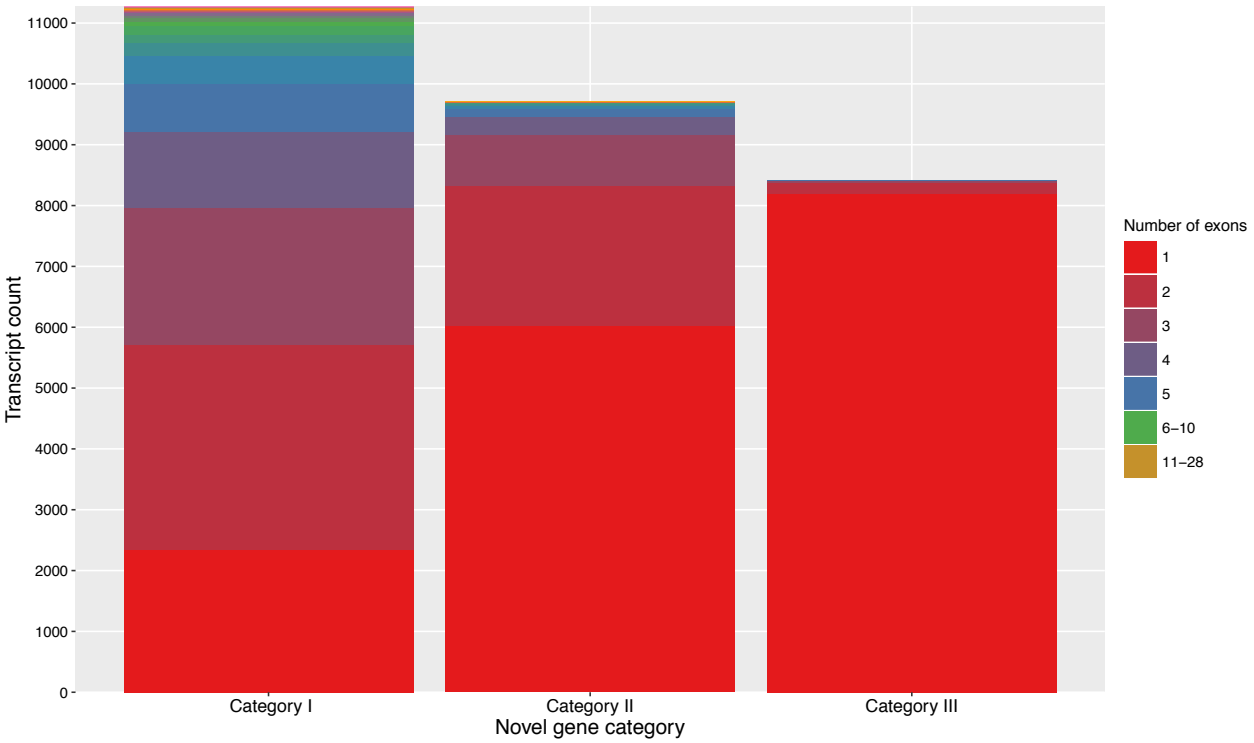


Figure 3. Tissue-specific gene and isoform composition of the transcriptome. A heatmap looking at a subset of genes which had a sum of TPMs across all tissues above 200 and a standard deviation above 200 were hierarchically clustered based on Pearson correlation between gene expression and a Spearman correlation between the tissue-specific expression profiles. Blocks of genes showing exclusive abundance in that tissue-specific transcriptome were further annotated and a Panther statistical overrepresentation test was done to provide a functional summary of the GO biological processes corresponding to these genes (A). Focusing on tissue specificity of isoforms, isoforms uniquely present (the only tissue possessing a TPM of at least 5) or solely absent (TPM= 0) show that the retina has the most solely absent and uniquely present transcripts (B). The number of transcripts expressed in a unique tissue-specific manner is depicted by the bar outlined in red above the x-axis, while the number of transcripts that are absent in a unique tissue-specific manner is denoted by the blue outlined bars extending below the x-axis. The green trendline corresponds to the cumulative TPM of the uniquely present transcripts. Regarding transcription of mitochondrial genes versus nuclear encoded genes, we provided the proportion of the transcriptional output for each tissue that was devoted to mitochondrial transcripts by calculating the percentage of expression originating from the mitochondrial contigs (C).

A)


20000
15000
10000
5000
0
transcript count
NCBI
ENSEMBL
ISME
Hestand
database
class code
match
similar
novel
others

B)


Transcript count
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Category I
Category II
Category III
Novel gene category
Number of exons
1
2
3
4
5
6–10
11–28

C)

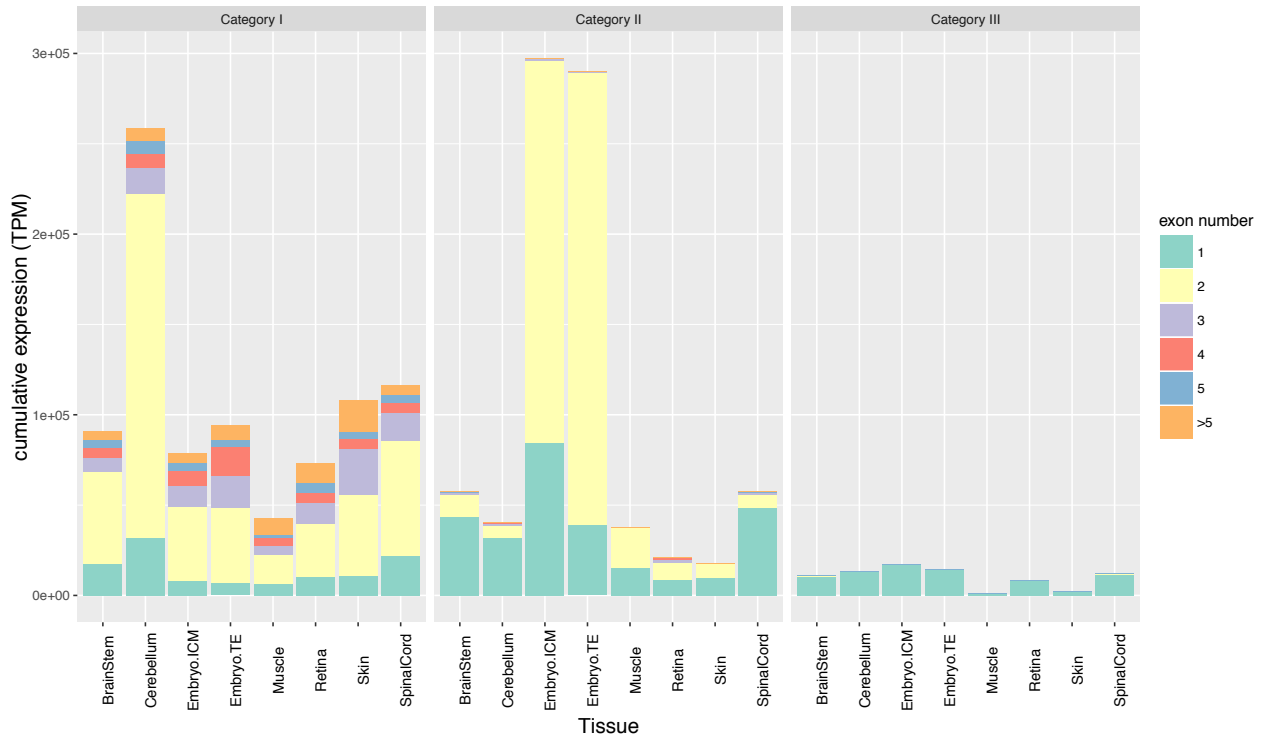


Figure 4. Novel gene analysis and classification into three different categories. The three categories of novel genes were supported novel genes (Category I), unsupported, but conserved, novel genes (Category II) and the unsupported, un-conserved, but complete ORF novel genes (Category III). Several of the novel genes that were supported by another equine assembly (Category I) originated from the ENSEMBL assembly and indeed had supporting gene models in another equine assembly (A). All three categories of novel genes showed varying exonic composition with the more supported novel genes (Category I and II) showing greater exon composition diversity (B). The novel genes and their cumulative TPM in each tissue, with regard to exon number (B), show a bias in expression of 1 and 2 exon genes in the cerebellum and embryonic tissues, with exon composition per category of novel genes reflecting what was found in (B).

Supplementary Tables

Supplementary table 1 overall mapping statistics for trimmed reads with Tophat2

Tissue/Library	Mapping	Concordance
BrainStem/PE_100_fr.firststrand	91.11%	85.51%
Cerebellum/PE_100_fr.firststrand	85.87%	78.19%
Embryo.ICM/PE_100_fr.unstranded	50.45%	29.48%
Embryo.ICM/SE_100_fr.unstranded	78.42%	--
Embryo.TE/PE_100_fr.unstranded	51.78%	31.51%
Embryo.TE/SE_100_fr.unstranded	77.95%	--
Muscle/PE_125_fr.firststrand	92.87%	89.10%
Retina/PE_81_fr.unstranded	72.49%	68.77%
Skin/PE_81_fr.unstranded	74.42%	69.49%
Skin/SE_81_fr.unstranded	50.91%	--
Skin/SE_95_fr.unstranded	84.98%	--
SpinalCord/PE_100_fr.firststrand	91.33%	85.46%
Total	**82.83%**	**74.99%**

Supplementary Table 2. General statistics regarding each version of the transcriptome.

	Unfiltered	mature	High-exp	Supported	Refined	Merged
Number of Transcripts	211,562	162,261	114,830	76,323	76,125	121,997
Average length	5272	6055	4183	5780	5793	4542
Maximum length	301,449	299,228	299,228	299,228	299,228	299,228
Minimum length	21	21	33	83	201	27
N50	11,185	11,608	8736	9624	9622	7767
Mb altogether	1115.4	982.4	480.3	441.1	441	554.2

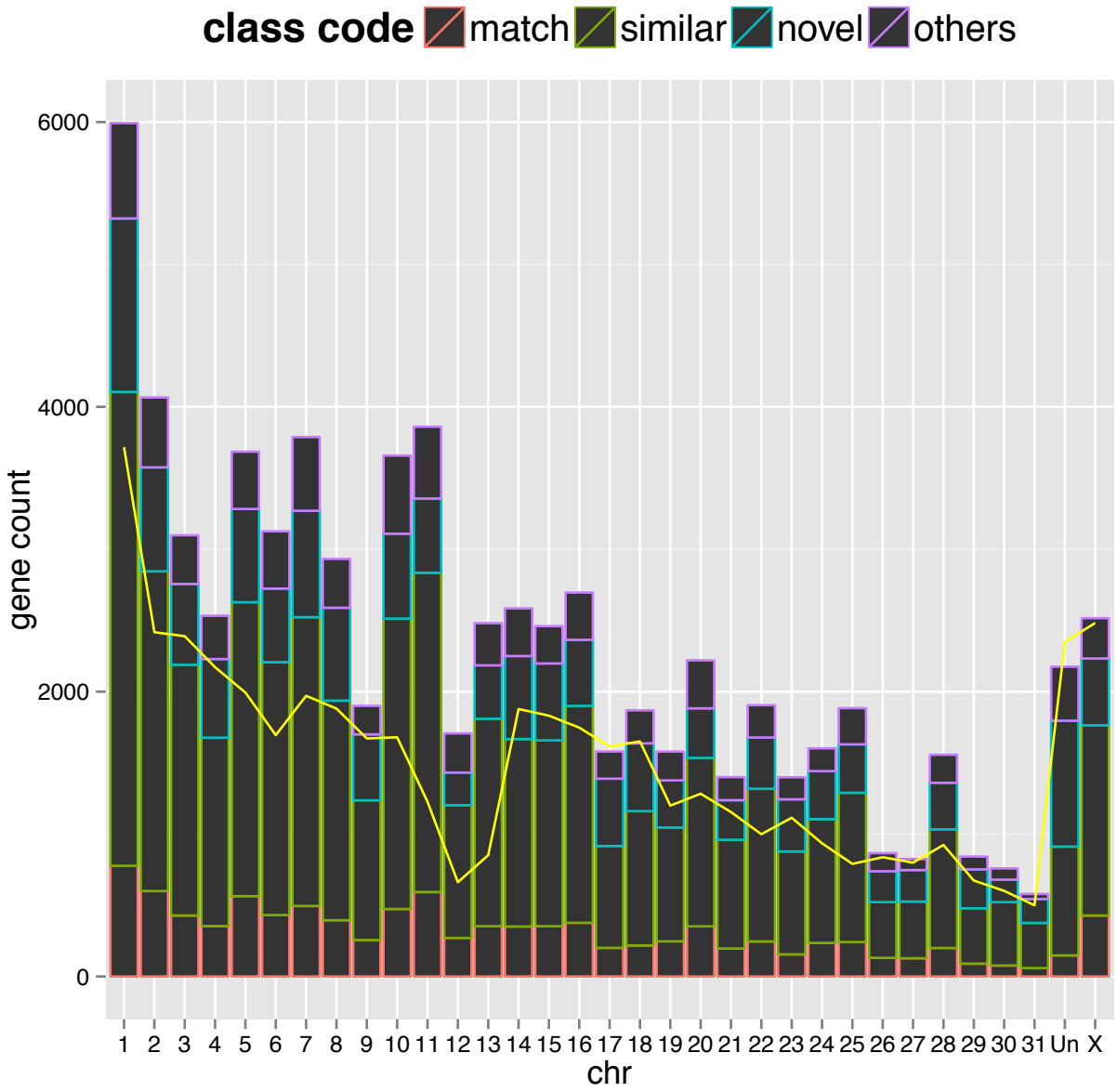
Supplementary Table 4. Sensitivity (sn) and specificity (sp) analysis of all equine annotations.

quary ID	reference ID	base		intron_chain		locus		missed exons	missed introns	missed loci	novel loci	Complex loci
		sn(%)	sp(%)	sn(%)	sp(%)	sn(%)	sp(%)					
Hestand_2014	NCBI	78.6	46.8	26.1	33.2	36.8	15.8	40305(17.4%)	33356(16.3%)	5151(21.2%)	33010(58.3%)	660
ISME.PBMC	NCBI	87.5	24.4	45.5	6.9	52.6	29.4	14807(6.4%)	7915(3.9%)	2690(11.1%)	21561(50.5%)	1546
ensGTF_file	NCBI	48.8	81.3	14.3	28.4	25.3	22.9	48574(21.0%)	34379(16.8%)	5268(21.6%)	7039(26.1%)	389
UnflitTrans	NCBI	81.4	14.5	33.8	12.2	36.8	7.5	35952(15.5%)	28504(13.9%)	5199(21.4%)	73188(62.5%)	1543
fl1_NoIntronicFrag	NCBI	80.5	19.3	33.8	12.2	36.8	11.7	36833(15.9%)	28504(13.9%)	5258(21.6%)	53222(70.9%)	1534
fl2_hiExp	NCBI	79.1	20.3	25.9	16.6	32.1	10.3	40183(17.3%)	30973(15.1%)	5485(22.5%)	53277(70.7%)	1360
fl3_supported	NCBI	78.8	23.8	25.9	16.8	32.1	21	41041(17.7%)	30975(15.1%)	5734(23.6%)	16678(45.0%)	1357
RNAseqSupTrans	NCBI	78.8	23.8	25.9	16.8	32	21.1	41117(17.7%)	30996(15.1%)	5774(23.7%)	16540(44.9%)	1355
mergedTrans	NCBI	99.8	27.6	94.6	38.2	93.9	45.5	0(0.0%)	212(0.1%)	0(0.0%)	23051(48.3%)	1520
Hestand_2014	ensGTF_file	80.6	28.8	24.2	15.5	17.8	8.5	32861(16.3%)	27604(15.8%)	6715(24.9%)	32987(58.3%)	798
ISME.PBMC	ensGTF_file	96.4	16.1	73.1	5.6	69	42.1	5235(2.6%)	3258(1.9%)	2605(9.7%)	19096(44.8%)	1226
NCBI	ensGTF_file	81.3	48.8	28	14.1	22.6	25	21592(10.7%)	9170(5.3%)	7039(26.1%)	5268(21.6%)	413
UnflitTrans	ensGTF_file	73.5	7.9	23	4.2	16.7	3.8	33579(16.6%)	24365(14.0%)	9103(33.8%)	83909(71.7%)	1164
fl1_NoIntronicFrag	ensGTF_file	73.2	10.5	23	4.2	16.7	6	34275(17.0%)	24365(14.0%)	9141(33.9%)	56888(75.7%)	1146
fl2_hiExp	ensGTF_file	72.3	11.1	19.1	6.2	14	5	36375(18.0%)	26235(15.0%)	9372(34.8%)	57022(75.7%)	999
fl3_supported	ensGTF_file	72.1	13.1	19.1	6.2	14	10.1	36934(18.3%)	26239(15.0%)	9493(35.2%)	19515(52.7%)	999
RNAseqSupTrans	ensGTF_file	72.1	13.1	19.1	6.2	14	10.2	37017(18.3%)	26254(15.1%)	9551(35.4%)	19376(52.5%)	997
mergedTrans	ensGTF_file	99.9	16.6	86.5	17.6	83.6	45.8	0(0.0%)	1657(0.9%)	0(0.0%)	21312(44.6%)	1241

Supplementary Table 6. Statistics for annotation of novel genes with Blastp and Blastx.

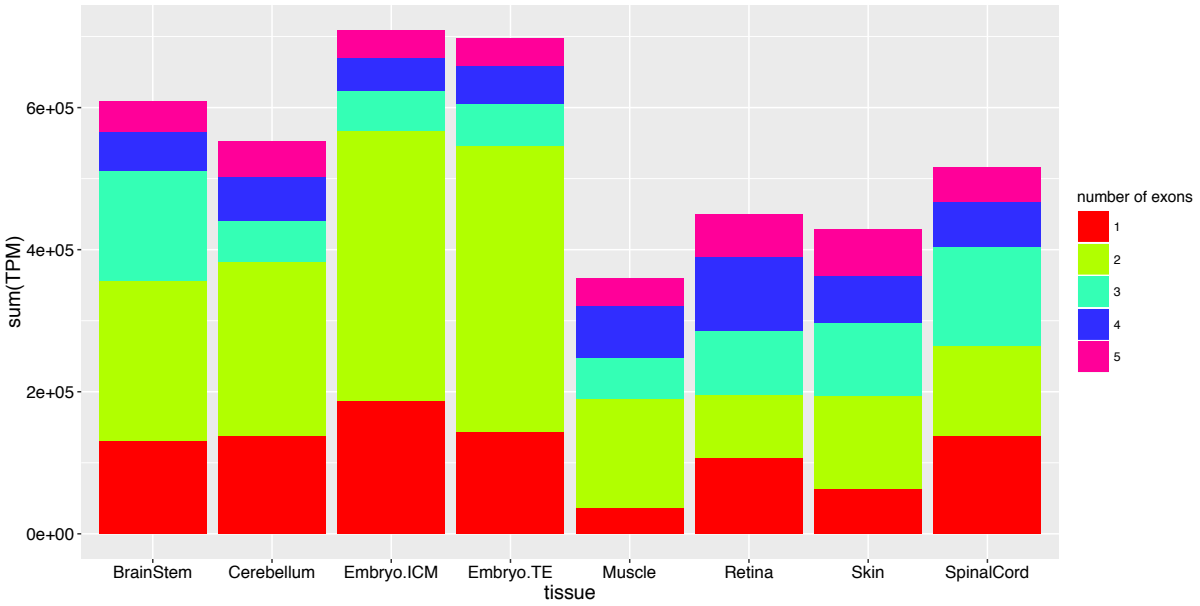
Group	Isoforms						Genes					
	All novel	Blastx hits	Blastp hit	Hits in both	Sum	Frequency	All novel	Blastx	Blastp	Hits in both	Sum	Frequency
Category I	8459	2237	59	1874	4170	49.2%	5136	1373	45	1072	2490	48.4%
Category II	7494	941	16	247	1204	16.0%	6474	795	13	223	1031	15.9%
Category III	6687	392	11	252	655	9.7%	6657	389	11	252	652	9.7%

Supplementary Figures

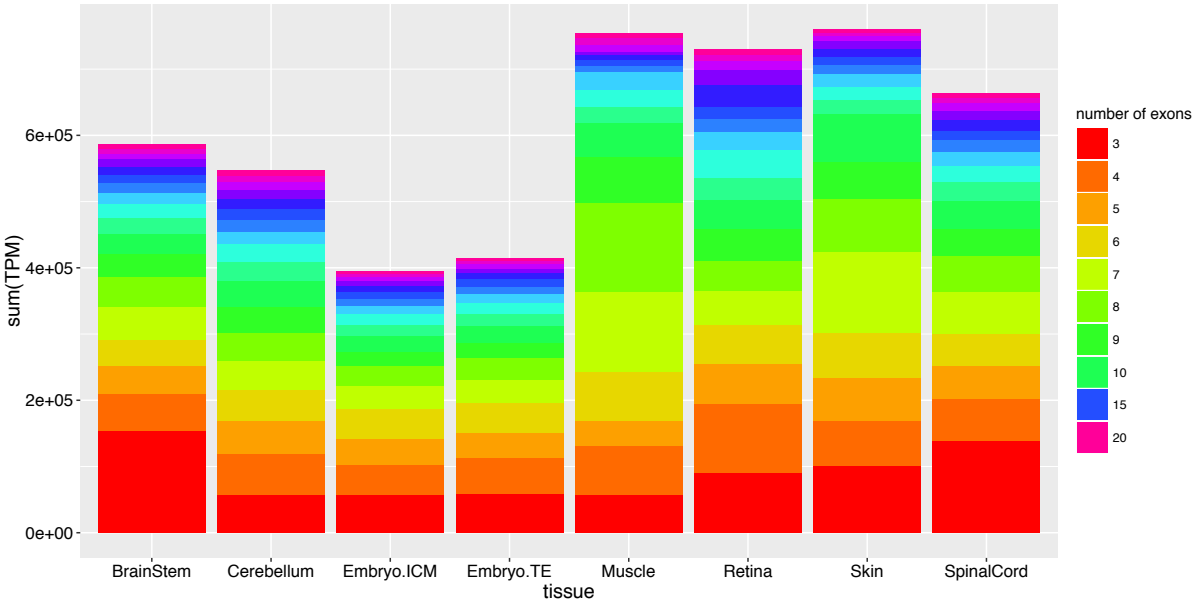


Supplementary Figure 1. The dispersion of the comparison annotation as they relate to the chromosomes, with reference solely to the NCBI database, demonstrates no bias in a certain comparison annotation to any particular chromosome in the refined version of the transcriptome, the yellow line represents a scale of the size of the chromosomes in Mb (B).

A)


6e+05
4e+05
2e+05
0e+00
sum(TPM)
BrainStem
Cerebellum
Embryo.ICM
Embryo.TE
Muscle
Retina
Skin
SpinalCord
tissue
number of exons
1
2
3
4
5

B)



Supplementary Figure 2. Transcript expression patterns in all tissues relative to the number of exons comprising the transcript in transcripts with up to 5 exons (A) and in transcripts with over 3 exons (B).