BMC Bioinformatics

ORAL PRESENTATION Open Access

Analysis of equine protein-coding gene structure and expression by RNA-sequencing

Stephen J Coleman[1]*, Zheng Zeng[2], Jinze Liu[2], James N MacLeod[1]

From UT-ORNL-KBRIN Bioinformatics Summit 2010
Cadiz, KY, USA. 19-21 March 2010

Background

RNA-sequencing (RNA-seq) data from eight equine tissue samples (34-day whole embryo, full term placental villous, adult testes, adult cerebellum, adult articular cartilage, adult LPS-stimulated articular cartilage, adult synovial membrane, and adult LPS-stimulated synovial membrane) were used to refine the structural annotation of protein-coding genes in the horse and for a preliminary assessment of tissue-specific expression patterns.

Materials and methods

A consensus set of equine protein-coding gene structures was defined by consolidation of gene sets predicted by Ensembl and NCBI (containing 20,322 and 17,610 genes respectively) and structural annotation derived from the RNA-seq experiments. First, the genomic locus data and intervals for the genes predicted by Ensembl and NCBI were combined. Overlapping loci from the two gene sets were analyzed by calculating the intersection and union of their respective coding (exon) sequences to determine if they represented the same gene. Next, structural boundaries for the resulting loci were compared to 75,116 expressed structures defined by the RNA-seq tag alignments. Experimentally derived RNA-seq annotation superceded the *in silico* predictions in reaching consensus gene models. For loci where RNA-seq derived structures could not be generated, the longest consensus model from the *in silico* predictions was used. In the opposite situation, RNA-seq data identified 215 transcriptional units with strong homology to known mammalian gene sequences, but not included in the *in silico* equine gene sets. Gene symbols were assigned to the consensus models based on the established Ensembl and NCBI annotations. The resulting consensus gene set currently contains 20,302 protein-coding genes.

Results and conclusion

Relative expression levels between tissues were determined for 17,270 of the consensus genes that do not structurally overlap with other protein-coding genes in the equine genome. Expression values were calculated by taking the sum of hits (individual basepair coverage) generated by the tags that aligned to a gene's coding region and dividing it by the cDNA length. All values were normalized to the total number of hits generated for that tissue sample. The number of genes expressed by individual tissues ranged from 9,716 (56.3%) to 12,038 (69.9%) [mean = 11,163 (64.6%)]. Gene ontology annotation was used to evaluate the functional and structural categories of genes expressed in either a stable or tissue-restricted pattern.

Author details

[1]Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA. [2]Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA.

Published: 23 July 2010

doi:10.1186/1471-2105-11-S4-O8
Cite this article as: Coleman *et al.*: **Analysis of equine protein-coding gene structure and expression by RNA-sequencing.** *BMC Bioinformatics* 2010 **11**(Suppl 4):O8.

* Correspondence: sjcole0@uky.edu
[1]Department of Veterinary Science, University of Kentucky, Lexington, KY 40546, USA