Data mining of transcriptomes across vertebrate species is shedding light on "missing genes"

Lieven Thorrez LIEVEN.THORREZ@KULEUVEN.BE
KU Leuven, Dept. Development and Regeneration, Kortrijk, Belgium

Raf Huttener RAF.HUTTENER@KULEUVEN.BE
Thomas in't Veld INTVELD@KULEUVEN.BE
Leentje Van Lommel LEENTJE.VANLOMMEL@KULEUVEN.BE
Frans Schuit FRANS.SCHUIT@KULEUVEN.BE
KU Leuven, Dept. of Cellular and Molecular Medicine, Leuven, Belgium

Keywords: transcriptomes, species biology, missing genes

Abstract

Comparing genome and transcriptome sequences across species provides insights in differential use of genes and molecular pathways, invaluable to understand differences between species and the function of genes in animal human physiology. In this study, we mined transcriptome data from all sequenced species present at NCBI in order to map missing genes in species, pinpoint technical artifacts, improve gene annotation and generate novel biological insights. Human gene data of protein coding genes were imported with Biomart. For genes with multiple transcripts, only the entry with the longest transcript was retained. This resulted in 19204 unique human genes. From all species, genbank files containing transcriptome information were downloaded from NCBI and for each protein-coding gene the longest transcript was selected and percent GC and transcript length was calculated with EMBOSS infoseq. Also the number of transcript variants was calculated. Organisms were then compared by exact matching of gene names to human genes. Problems arise however when exact matching can not be performed (eg LOC names). Strategies to improve gene annotation are currently being explored and will be discussed. Comparison of genomewide transcript GC contents and transcript lengths across species or groups of vertebrate species (fish, birds, mammals) revealed several patterns. Regions where genes apparently are missing are clustered in peaks and certain peaks are specific for species groups. The species-specific buildup of GC content in certain regions of the transcriptome provides a window of opportunity to study gene evolution. Indeed, changes in GC content are associated with use of amino acids encoded by GC rich codons. These peaks co-occur with high GC-content regions. The missing status of several genes may due to limitations of current sequencing technologies, consistent with literature where sequencing difficulties of very low or high GC regions has been reported. Moreover, in these peaks, where transcripts are found or predicted, these transcripts are shorter than in human. Certain missing genes apparent in the peaks - although absent in the NCBI and Ensembl databases- could be identified by inspection of EST data. Identifying true missing genes is crucial to provide insights in the biology of species, for example pseudogenization of enamelin associated with the absence of teeth in birds. In conclusion, by means of mining large datasets of currently available transcriptome data, we were able to more precisely map missing genes in species, provide insights whether or not this can be related to technical artefacts and identify mechanisms driving speciation.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).