Rat gene expression data expansion at the Rat Genome Database, an update

Wendy M. Demos[1], Logan Lamers[1], Jeff L. De Pons[1], Adam C. Gibson[1], Varun Reddy Gollapally[1], G. Thomas Hayman[1], Mary L. Kaldunski[1], Akhilanand Kundurthi[1], Stanley J.F. Laulederkind[1], Jennifer R. Smith[1], Jyothi Thota[1], Marek A. Tutaj[1], Monika Tutaj[1], Mahima Vedi[1], Shur-Jen Wang[1], Kent C. Brodie[2], Stacy M. Zacher[3], Melinda R. Dwinell[1], Anne E. Kwitek[1].

[1]Rat Genome Database, Department of Physiology, [2]Clinical and Translational Science Institute, [3]Finance and Administration, Medical College of Wisconsin, Milwaukee, WI, USA.

The Rat Genome Database (RGD, https://rgd.mcw.edu) has been expanding and incorporating gene expression data content into the larger ecosystem of RGD. Researchers will soon be able to access gene expression values and sample metadata that were submitted to public resources such as the Gene Expression Omnibus (GEO) repository, with all data values converted to transcripts per million (TPM) data type.

In Phase One of the project, an expression curation tool was developed to aid in comprehensive Natural Language Processing (NLP) assisted manual curation of public datasets. To date, 2,109 GEO expression projects with focus on the *Rattus norvegicus* model have been reviewed and prioritized for curation. Of those projects, 307 met the primary prioritization criteria and metadata for 3,629 project samples have been loaded to the database.

The goal of Phase Two was to standardize the publicly available gene expression data by reprocessing samples in the curated GEO projects. RGD developed and is optimizing a bioinformatic pipeline that downloads and converts fastq files from the Sequence Read Archive, aligns to the most current and correct *R. norvegicus* genome assembly, and outputs TPM data type. This pipeline integrates quality control measures, verifies or calculates sample sex, produces alignments with the STAR aligner, and estimates gene and transcript level abundance with the RSEM software package.

Phase Three focuses on enhanced visualization of gene expression values. The tabular-based display of gene expression values has been updated on the RGD gene pages. Users can view data by anatomical system and download sample metadata, TPM values, and data sources. Evaluation of JBrowse2 visualizations for coverage and TPM values at the gene and transcript levels is underway. These curated and reprocessed expression data, download options and visualizations will provide standardization of publicly available data and continue to add value for RGD users.