https://rgd.mcw.edu

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].

Rat Genome Database, Dept. of Physiology[1]; Clinical and Translational Science Institute[2]; Finance and Administration[3] Medical College of Wisconsin, Milwaukee, WI 53226 USA

Rat Genome Database

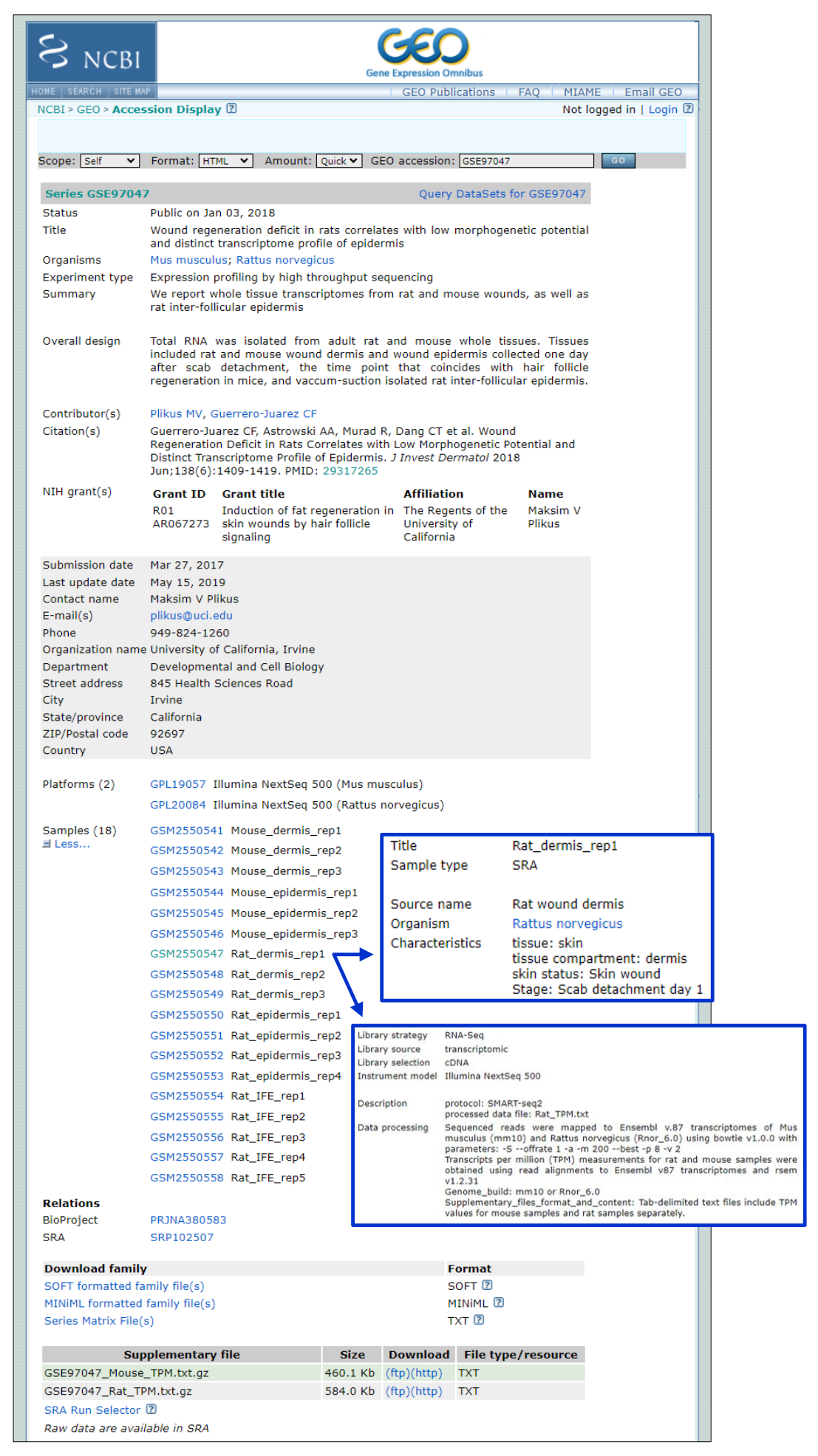
Abstract

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, https://www.ncbi.nlm.nih.gov/geo/) repository, with all data values converted to transcripts per million (TPM) data type. Integration of data submitted to GEO was designed in three phases as outlined below. Additional expression data, such as the baseline expression of strains related to the Hybrid Rat Diversity Program (https://rgd.mcw.edu/rgdweb/hrdp_panel.html) have been integrated into JBrowse2.

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 is 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 expression profiles 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.



Series GSE97047 Query DataSets for GSE97047

Status: Public on Jan 03, 2018
Title: Wound regeneration deficit in rats correlates with low morphogenetic potential and distinct transcriptome profile of epidermis
Organisms: Mus musculus; Rattus norvegicus
Experiment type: Expression profiling by high throughput sequencing
Summary: We report whole tissue transcriptomes from rat and mouse wounds, as well as rat inter-follicular epidermis
Overall design: Total RNA was isolated from adult rat and mouse whole tissues. Tissues included rat and mouse wound dermis and wound epidermis collected one day after scab detachment, the time point that coincides with hair follicle regeneration in mice, and vaccum-suction isolated rat inter-follicular epidermis.
Contributor(s): Plikus MV, Guerrero-Juarez CF
Citation(s): Guerrero-Juarez CF, Astrowski AA, Murad R, Dang CT et al. Wound Regeneration Deficit in Rats Correlates with Low Morphogenetic Potential and Distinct Transcriptome Profile of Epidermis. *J Invest Dermatol* 2018 Jun;138(6):1409-1419. PMID: 29317265

NIH grant(s):

Grant ID	Grant title	Affiliation	Name
R01 AR067273	Induction of fat regeneration in skin wounds by hair follicle signaling	The Regents of the University of California	Maksim V Plikus

Submission date: Mar 27, 2017
Last update date: May 15, 2019
Contact name: Maksim V Plikus
E-mail(s): plikus@uci.edu
Phone: 949-824-1260
Organization name: University of California, Irvine
Department: Developmental and Cell Biology
Street address: 845 Health Sciences Road
City: Irvine
State/province: California
ZIP/Postal code: 92697
Country: USA

Platforms (2): GPL19057 Illumina NextSeq 500 (Mus musculus)
GPL20084 Illumina NextSeq 500 (Rattus norvegicus)

Samples (18) Less...
GSM2550541 Mouse_dermis_rep1
GSM2550542 Mouse_dermis_rep2
GSM2550543 Mouse_dermis_rep3
GSM2550544 Mouse_epidermis_rep1
GSM2550545 Mouse_epidermis_rep2
GSM2550546 Mouse_epidermis_rep3
GSM2550547 Rat_dermis_rep1
GSM2550548 Rat_dermis_rep2
GSM2550549 Rat_dermis_rep3
GSM2550550 Rat_epidermis_rep1
GSM2550551 Rat_epidermis_rep2
GSM2550552 Rat_epidermis_rep3
GSM2550553 Rat_epidermis_rep4
GSM2550554 Rat_IFE_rep1
GSM2550555 Rat_IFE_rep2
GSM2550556 Rat_IFE_rep3
GSM2550557 Rat_IFE_rep4
GSM2550558 Rat_IFE_rep5

Title: Rat_dermis_rep1
Sample type: SRA
Source name: Rat wound dermis
Organism: Rattus norvegicus
Characteristics: tissue: skin
tissue compartment: dermis
skin status: Skin wound
Stage: Scab detachment day 1

Library strategy: RNA-Seq
Library source: transcriptomic
Library selection: cDNA
Instrument model: Illumina NextSeq 500
Description: protocol: SMART-seq2
processed data file: Rat_TPM.txt
Data processing: Sequenced reads were mapped to Ensembl v.87 transcriptomes of Mus musculus (mm10) and Rattus norvegicus (Rnor_6.0) using bowtie v1.0.0 with parameters: -S --offrate 1 -a -m 200 --best -p 8 -v 2
Transcripts per million (TPM) measurements for rat and mouse samples were obtained using read alignments to Ensembl v87 transcriptomes and rsem v1.2.31
Genome_build: mm10 or Rnor_6.0
Supplementary_files_format_and_content: Tab-delimited text files include TPM values for mouse samples and rat samples separately.

Relations
BioProject: PRJNA380583
SRA: SRP102507

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE97047_Mouse_TPM.txt.gz	460.1 Kb	(ftp)(http)	TXT
GSE97047_Rat_TPM.txt.gz	584.0 Kb	(ftp)(http)	TXT

SRA Run Selector
Raw data are available in SRA

Figure 1: GEO Accession Display Interface. Inset shows sample specific details submitted to GEO.

Phase 1: GEO Metadata Curation

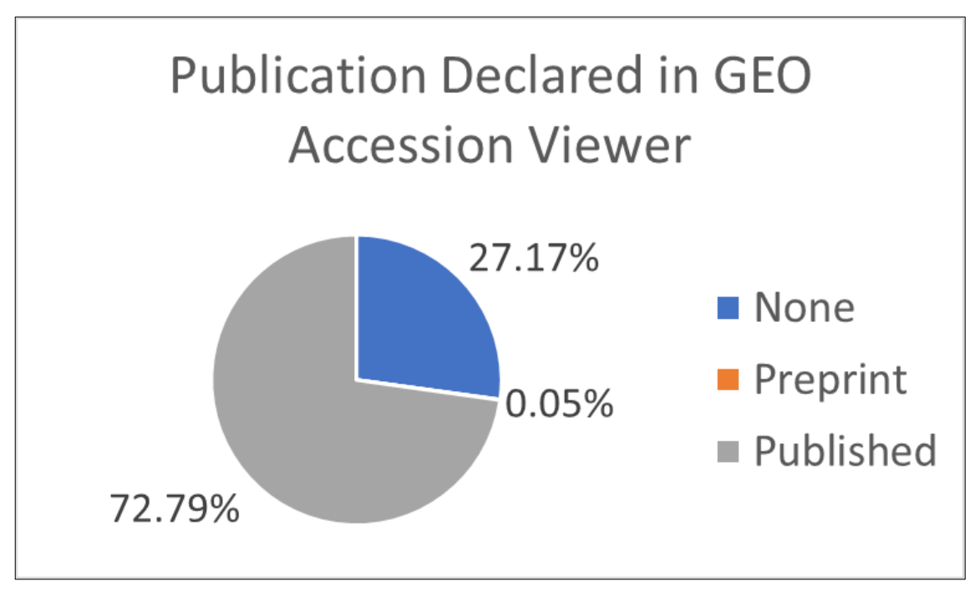


Figure 2: Publication rate of rat expression studies in GEO Accession Viewer. Publication is not necessary for curation if enough details are provided.

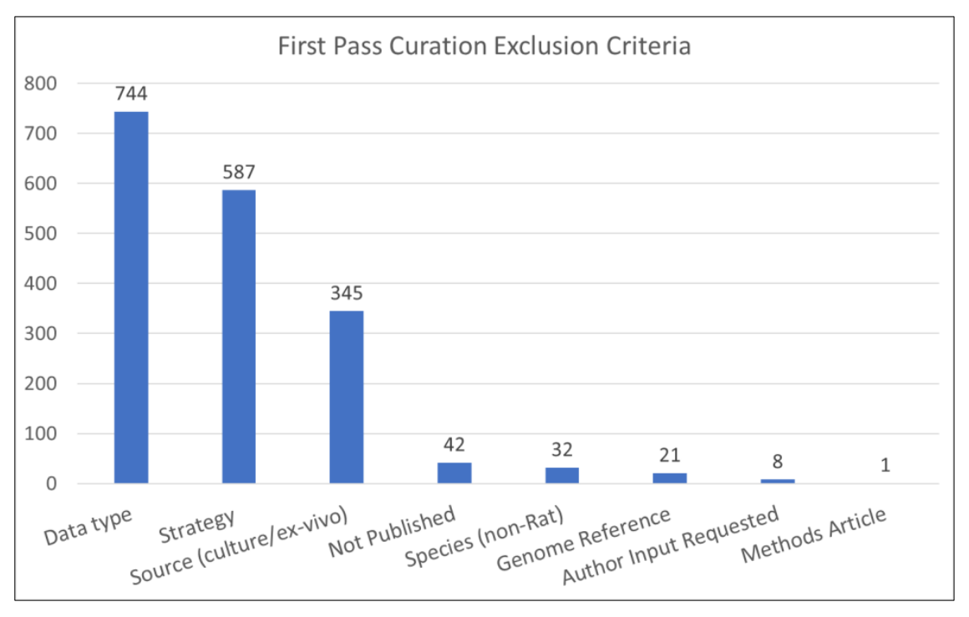
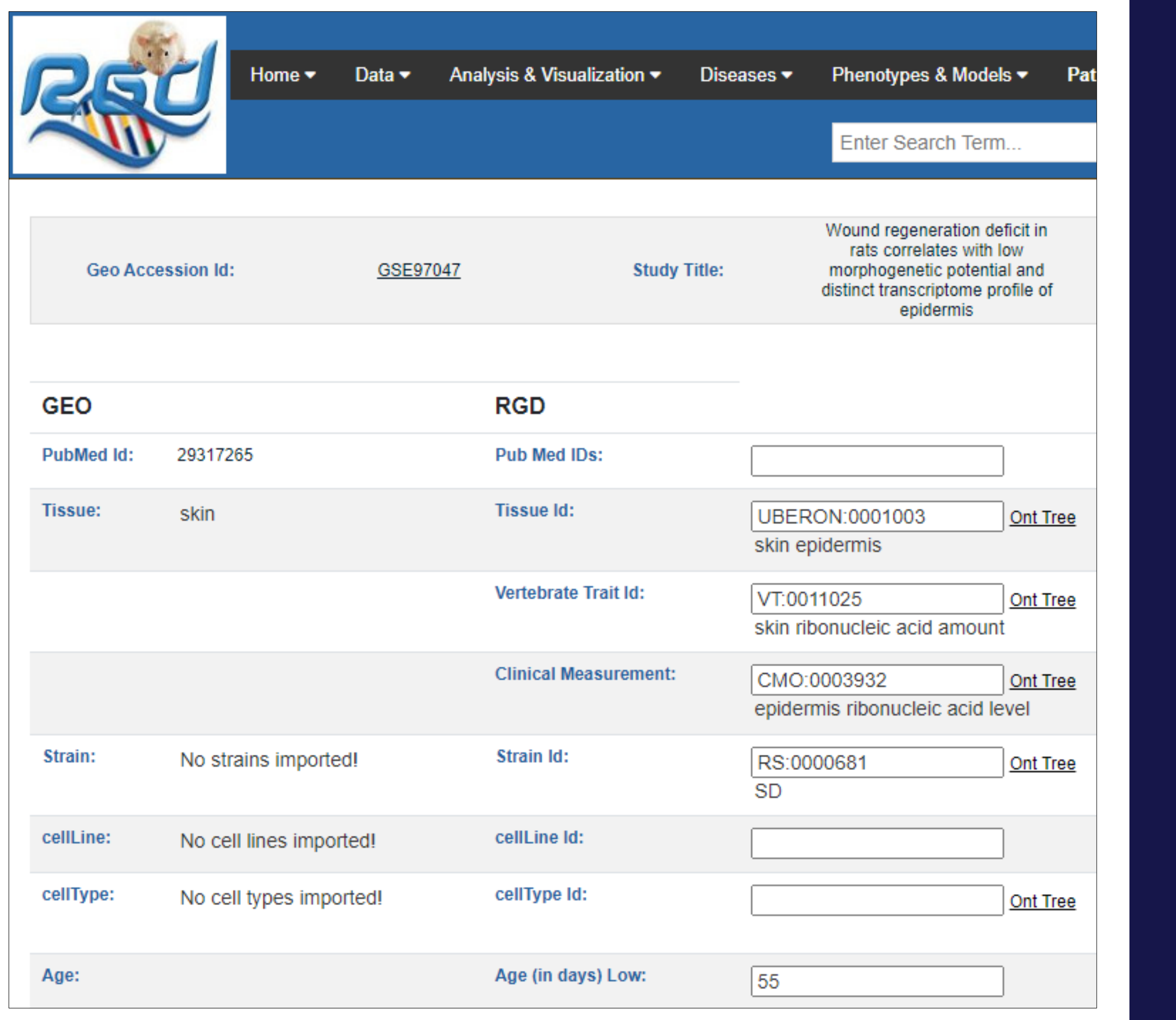


Figure 3: Criteria for exclusion from first-pass curation that are grounds to mark projects for future or no curation.

Curators leverage information submitted to the GEO repository (**Figure 1**) and associated publications, if available, (**Figure 2**) and prioritize curation based on data type FPKM / TPM (for data import) and bulk RNA sequencing strategy. Most projects earmarked for future curation are due to the data type submitted to GEO followed by the sequencing strategy (**Figure 3**). The Expression Curation Tool (**Figure 4**) relies on a pipeline that imports data from the GEO and utilizes Natural Language Processing to match ontology terms to GEO accession attributes. This interface provides fields for 7 ontologies, descriptors, and experimental conditions. Curators utilize built-in ontology browsers to enter missing terms, confirm the predicted terms, or provide a more specific term when appropriate.



Geo Accession Id: GSE97047
Study Title: Wound regeneration deficit in rats correlates with low morphogenetic potential and distinct transcriptome profile of epidermis

GEO		RGD		
PubMed Id:	29317265	Pub Med IDs:		
Tissue:	skin	Tissue Id:	UBERON:0001003 skin epidermis	Ont Tree
		Vertebrate Trait Id:	VT:0011025 skin ribonucleic acid amount	Ont Tree
		Clinical Measurement:	CMO:0003932 epidermis ribonucleic acid level	Ont Tree
Strain:	No strains imported!	Strain Id:	RS:0000681 SD	Ont Tree
cellLine:	No cell lines imported!	cellLine Id:		
cellType:	No cell types imported!	cellType Id:		Ont Tree
Age:		Age (in days) Low:	55	

Figure 4: Expression curation tool interface. Terms are first added at the project level and propagated across all applicable samples. Curators can then modify individual sample terms and attributes as needed based on the reference, within the next window of the interface prior to loading metadata to the underlying data tables.

Phase 2: Data Analysis

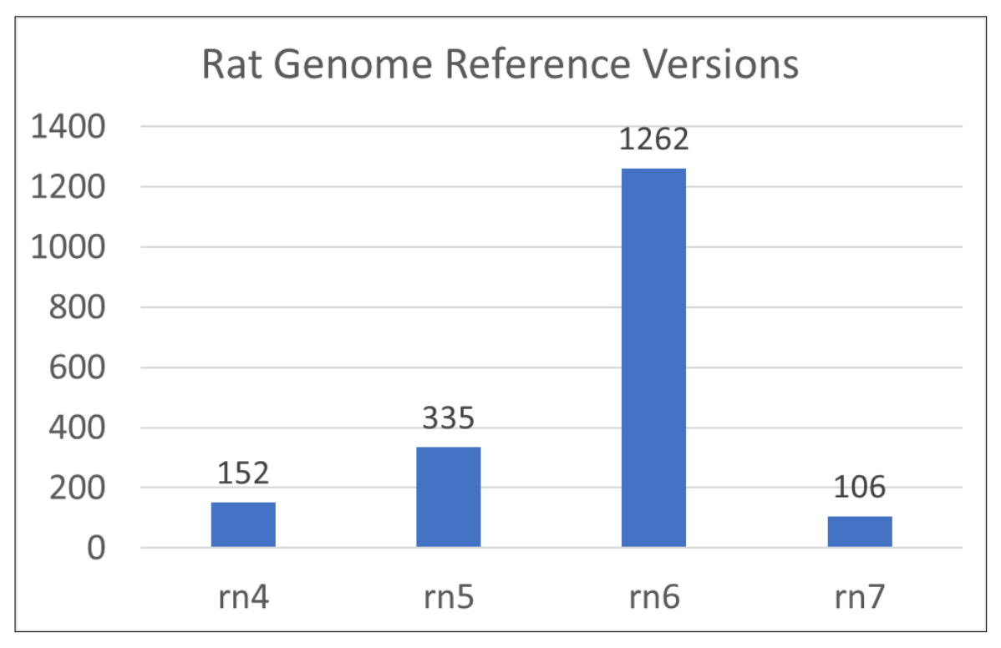


Figure 5: *R. norvegicus* genome reference versions declared in GEO for rat expression studies. This graph does not discern between RefSeq or Ensembl versions.

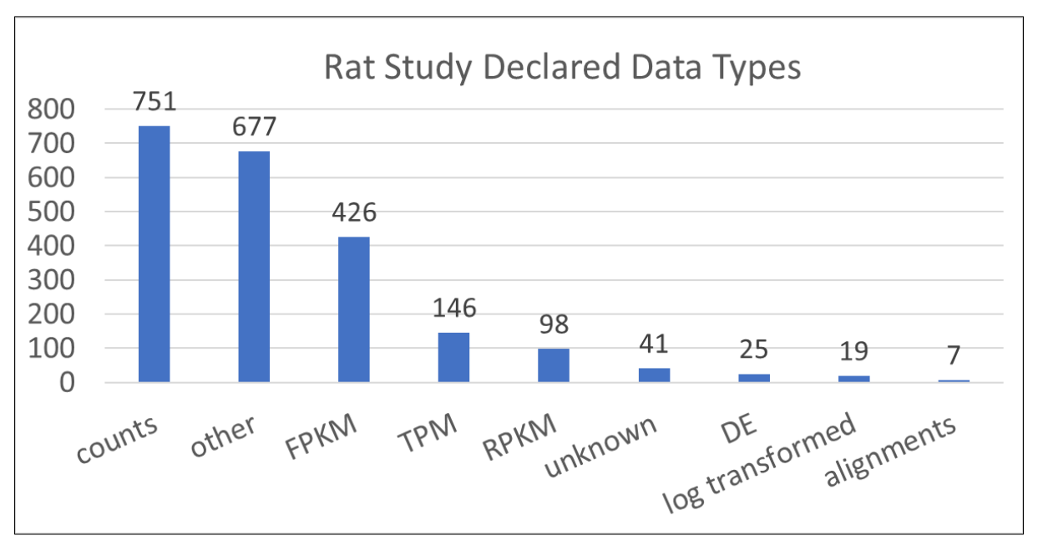
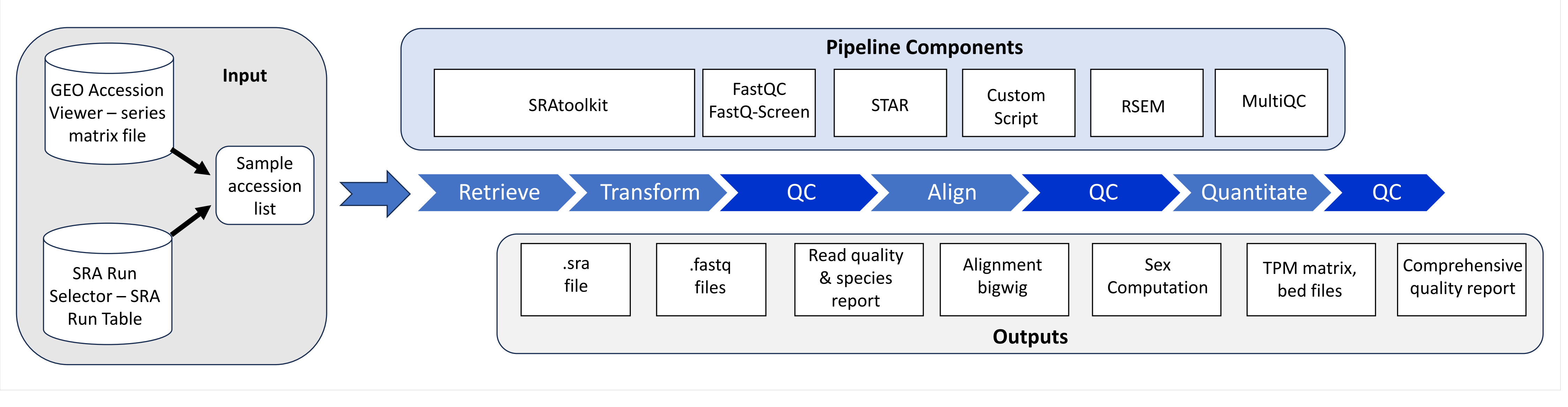


Figure 6: Processed or raw data are submitted as supplemental files to GEO. Submitters may declare the data type in the data processing field at the time of submission.

Submission to GEO does not require a standard reference **(Figure 5)** or data type (**Figure 6**). The RGD RNAseq pipeline (**Figure 7**) uses a sample accession list generated from information provided in the GEO series matrix file and the SRA run table. The workflow is launched from a bash menu and requires the accession list as input. Data are retrieved from the cloud as .sra files with the SRAtoolkit to a local server and transformed to fastq files. Multiple quality control steps occur throughout the process to evaluate sequencing quality measures, confirm species and sex (**Table 1**, **Figure 8**). Alignment to the rat genome reference is completed by the STAR aligner and abundance measures are calculated by RSEM. Verification of the workflow included comparing data submitted to GEO with generated data output. Expression thresholds as defined by Expression Atlas were used to verify similarity between results (**Figure 9**).


Input: GEO Accession Viewer – series matrix file; SRA Run Selector – SRA Run Table → Sample accession list

Pipeline Components: SRAtoolkit; FastQC FastQ-Screen; STAR; Custom Script; RSEM; MultiQC

Retrieve → Transform → QC → Align → QC → Quantitate → QC

Outputs: .sra file; .fastq files; Read quality & species report; Alignment bigwig; Sex Computation; TPM matrix, bed files; Comprehensive quality report

Figure 7: Schematic of the RGD RNAseq workflow highlighting input, software components, and outputs. Sample-specific attributes are gathered from the GEO Accession Viewer and the SRA Run Selector and compiled into a text file. The attributes are leveraged in various stages of the process. The workflow pauses prior to each QC step, ensuring curator review prior moving forward to the next stage. BigWig files are intended for visualization in JBrowse2, matrix files will be imported into the database for display on the gene pages.

Sample ID	Input Sex	Computed Sex	Agreement
GSM4874160	M	M	Agree
GSM4874161	F	F	Agree
GSM4874162	F	F	Agree
GSM4874165	F	F	Agree
GSM4874166	F	M	Conflict
GSM4874168	M	M	Agree
GSM4874169	M	M	Agree

Table 1: Example of sample sex quality control output. In this case, the sex declared of sample GSM4874166 in the GEO accession viewer does not agree with the computed sex.

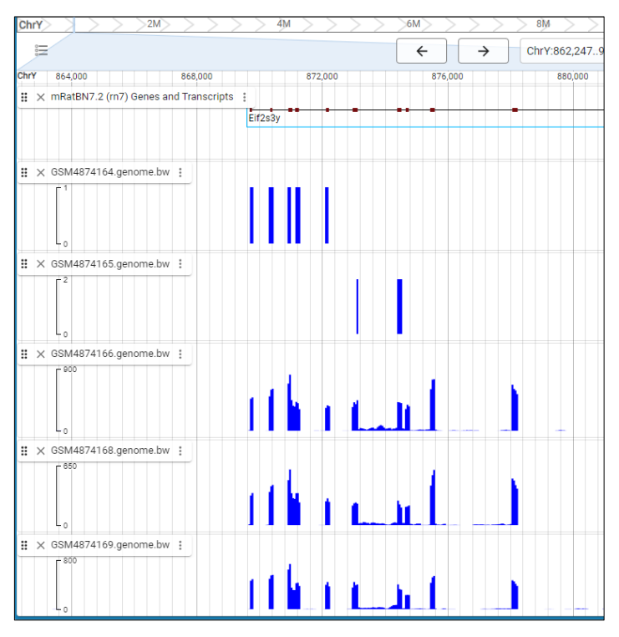


Figure 8: Sample sex can be verified by viewing the bigwig files in JBrowse2. In this example, the alignment supports the computed sex of male is accurate.

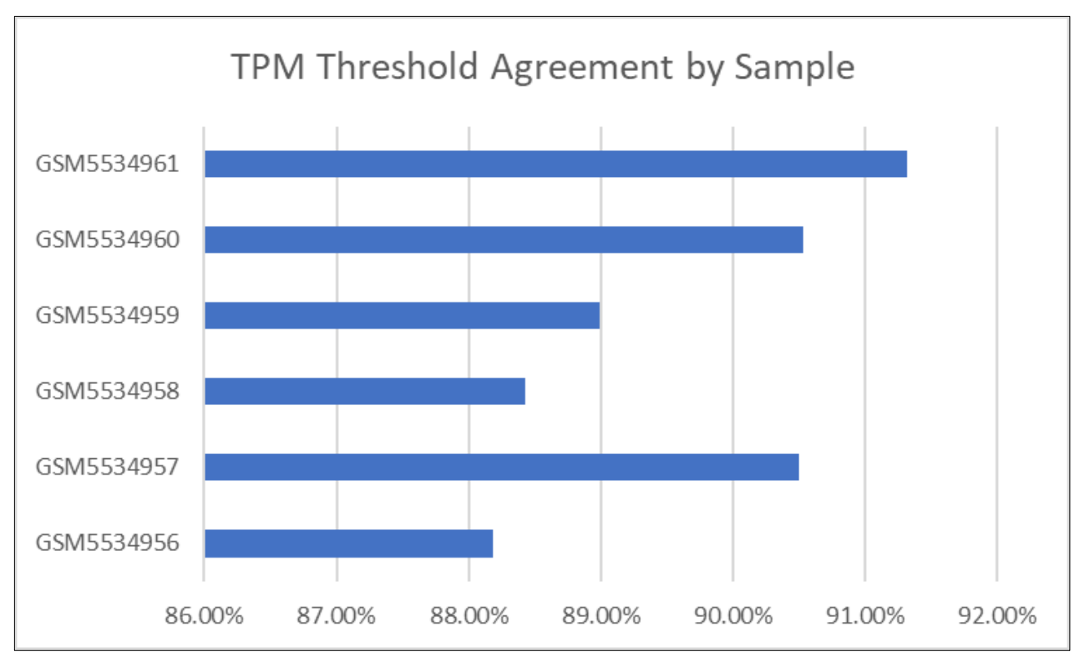
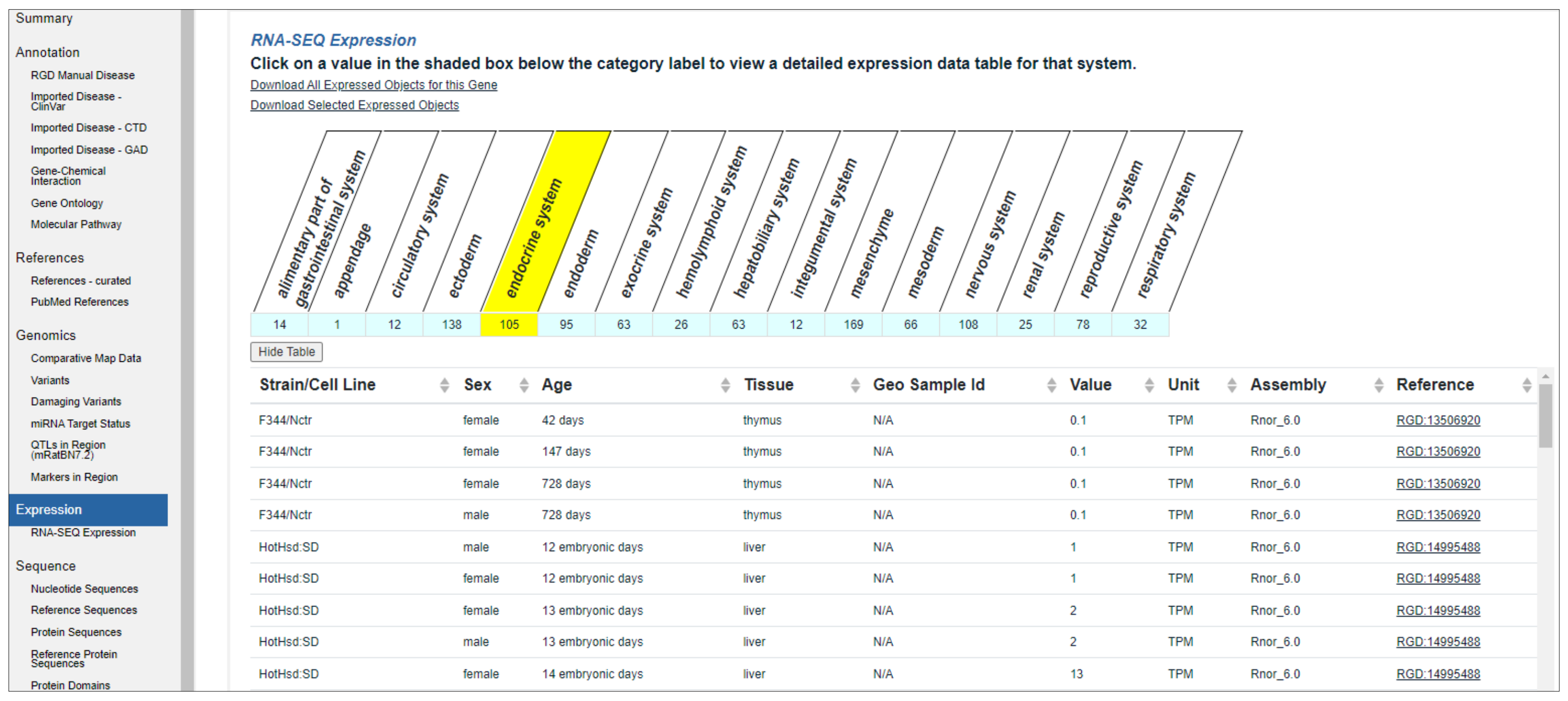


Figure 9: Agreement of expression levels between sample transcript TPM data submitted for GSE182706 and RGD generated data. Data were normalized to common transcripts.

Phase 3: GEO Data Visualization

To date, a total of 31 studies encompassing about 22,000 records with an average of approximately 32,000 individual expression values per record are viewable in the database. Initial loads were imported from Expression Atlas, leveraging their standardization of the metadata fields, and the mapping of the sequences to the Rnor6.0 assembly. Additionally, 3 rat studies and one study with data for both rat and dog were imported from GEO. In all cases, samples are untreated controls and expression values are in TPM. As curated GEO projects are subjected to reprocessing by the RGD RNAseq workflow, those data will be loaded to the database. Data are viewable and downloadable from any gene page within the RGD (**Figure 10**). Publication information for each data point is available from the 'Reference' link on the expanded expression table. (**Figure 11**).



RNA-SEQ Expression

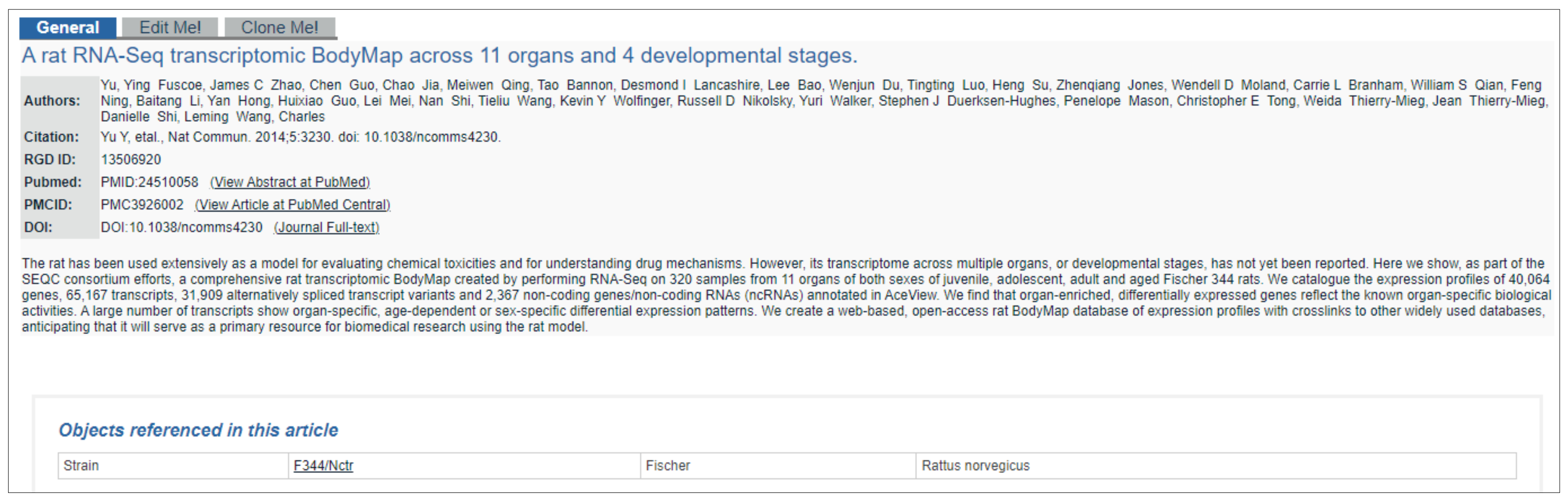
Click on a value in the shaded box below the category label to view a detailed expression data table for that system.

Download All Expressed Objects for this Gene
Download Selected Expressed Objects

alimentary part of gastrointestinal system	appendage	circulatory system	ectoderm	endocrine system	endoderm	exocrine system	hemolymphoid system	hepatobiliary system	integumental system	mesenchyme	mesoderm	nervous system	renal system	reproductive system	respiratory system
14	1	12	138	105	95	63	26	63	12	169	66	108	25	78	32

Strain/Cell Line	Sex	Age	Tissue	Geo Sample Id	Value	Unit	Assembly	Reference
F344/Nctr	female	42 days	thymus	N/A	0.1	TPM	Rnor_6.0	RGD:13506920
F344/Nctr	female	147 days	thymus	N/A	0.1	TPM	Rnor_6.0	RGD:13506920
F344/Nctr	female	728 days	thymus	N/A	0.1	TPM	Rnor_6.0	RGD:13506920
F344/Nctr	male	728 days	thymus	N/A	0.1	TPM	Rnor_6.0	RGD:13506920
HotHsd:SD	male	12 embryonic days	liver	N/A	1	TPM	Rnor_6.0	RGD:14995488
HotHsd:SD	female	12 embryonic days	liver	N/A	1	TPM	Rnor_6.0	RGD:14995488
HotHsd:SD	female	13 embryonic days	liver	N/A	2	TPM	Rnor_6.0	RGD:14995488
HotHsd:SD	male	13 embryonic days	liver	N/A	2	TPM	Rnor_6.0	RGD:14995488
HotHsd:SD	female	14 embryonic days	liver	N/A	13	TPM	Rnor_6.0	RGD:14995488

Figure 10: RNA-seq expression value display from a RGD gene page. Users can navigate to the expression section by using the navigation tools on the left-hand side of the browser window (highlighted in blue). Data are presented by organ system within the interface. Users can download all data points as a table or limit the table data to a single system.



A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages.

Authors: Yu, Ying Fuscoe, James C Zhao, Chen Guo, Chao Jia, Meiwen Qing, Tao Bannon, Desmond I Lancashire, Lee Bao, Wenjun Du, Tingting Luo, Heng Su, Zhenqiang Jones, Wendell D Moland, Carrie L Branham, William S Qian, Feng Ning, Baitang Li, Yan Hong, Huixiao Guo, Lei Mei, Nan Shi, Tieliu Wang, Kevin Y Wolfinger, Russell D Nikolsky, Yuri Walker, Stephen J Duerksen-Hughes, Penelope Mason, Christopher E Tong, Weida Thierry-Mieg, Jean Thierry-Mieg, Danielle Shi, Leming Wang, Charles
Citation: Yu Y, etal., Nat Commun. 2014;5:3230. doi: 10.1038/ncomms4230.
RGD ID: 13506920
Pubmed: PMID:24510058 (View Abstract at PubMed)
PMCID: PMC3926002 (View Article at PubMed Central)
DOI: DOI:10.1038/ncomms4230 (Journal Full-text)

The rat has been used extensively as a model for evaluating chemical toxicities and for understanding drug mechanisms. However, its transcriptome across multiple organs, or developmental stages, has not yet been reported. Here we show, as part of the SEQC consortium efforts, a comprehensive rat transcriptomic BodyMap created by performing RNA-Seq on 320 samples from 11 organs of both sexes of juvenile, adolescent, adult and aged Fischer 344 rats. We catalogue the expression profiles of 40,064 genes, 65,167 transcripts, 31,909 alternatively spliced transcript variants and 2,367 non-coding genes/non-coding RNAs (ncRNAs) annotated in AceView. We find that organ-enriched, differentially expressed genes reflect the known organ-specific biological activities. A large number of transcripts show organ-specific, age-dependent or sex-specific differential expression patterns. We create a web-based, open-access rat BodyMap database of expression profiles with crosslinks to other widely used databases, anticipating that it will serve as a primary resource for biomedical research using the rat model.

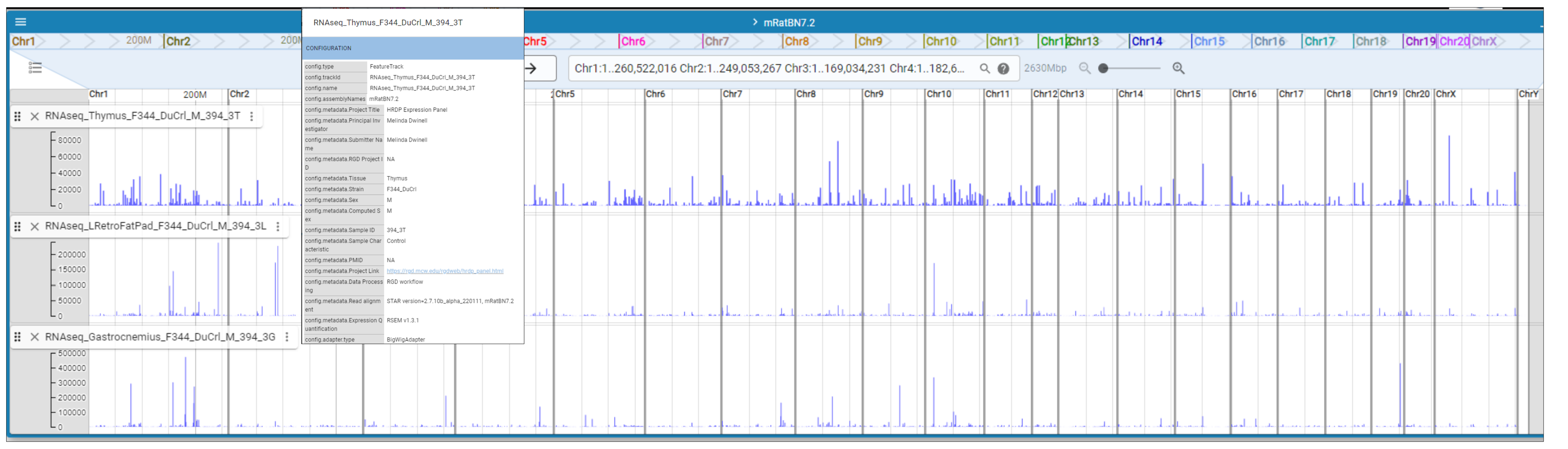
Objects referenced in this article

Strain	F344/Nctr	Fischer	Rattus norvegicus

Figure 11: Reference information. Links located under the 'Reference' column heading of the expression table will bring users to a new window that provides information regarding the origin of the data and related objects.

Expanded Content

In addition to providing standardized publicly available data, RGD recently released baseline expression profiles of classic inbred strains of the of the Hybrid Rat Diversity Panel. The first release includes gastrocnemius, left retroperitoneal fat pad and thymus of the F344/DuCrl and LE/StmMcwi strains (**Figure 12**). Further development is underway to provide gene and transcript level expression visualizations. TPM values as displayed in **Figure 13** may be an available option for data submitted to the Community Projects Portal.

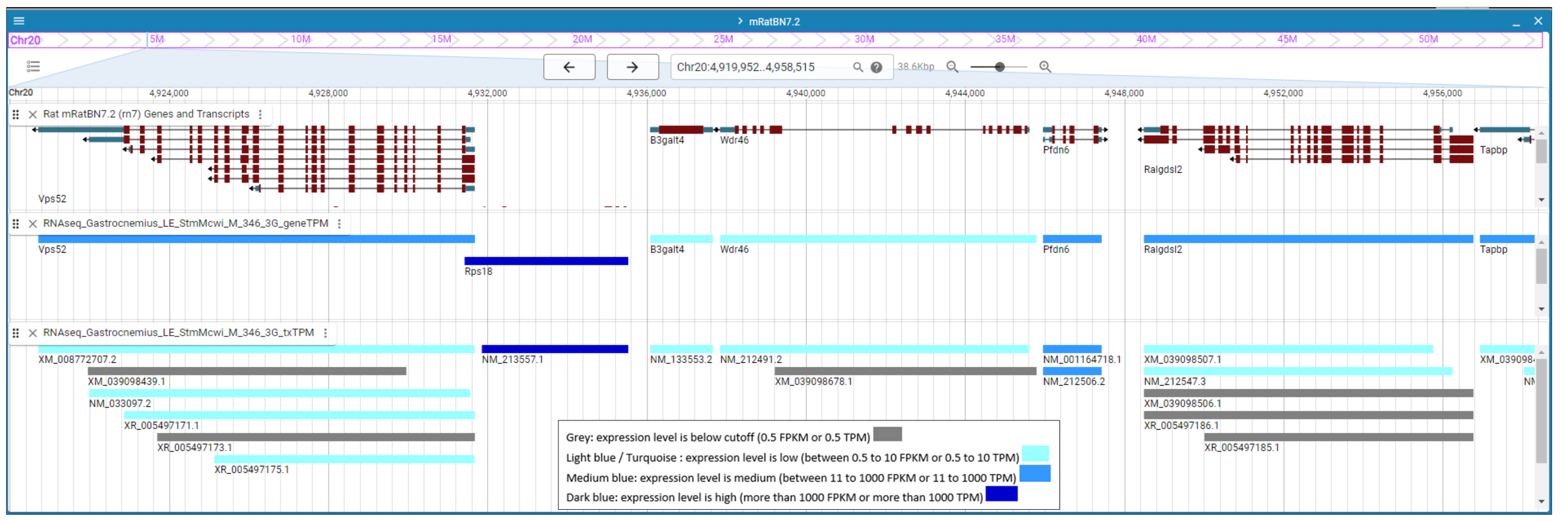

mRatBN7.2
Chr1
200M
Chr2
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
ChrX
ChrY
Chr1:1..260,522,016 Chr2:1..249,053,267 Chr3:1..169,034,231 Chr4:1..182,6...
2630Mbp
RNAseq_Thymus_F344_DuCrl_M_394_3T
80000
60000
40000
20000
0
RNAseq_LRetroFatPad_F344_DuCrl_M_394_3L
200000
150000
100000
50000
0
RNAseq_Gastrocnemius_F344_DuCrl_M_394_3G
500000
400000
300000
200000
100000
0

RNAseq_Thymus_F344_DuCrl_M_394_3T

CONFIGURATION

config.type: FeatureTrack
config.trackId: RNAseq_Thymus_F344_DuCrl_M_394_3T
config.name: RNAseq_Thymus_F344_DuCrl_M_394_3T
config.assemblyNames: mRatBN7.2
config.metadata.Project Title: HRDP Expression Panel
config.metadata.Principal Investigator: Melinda Dwinell
config.metadata.Submitter Name: Melinda Dwinell
config.metadata.RGD Project ID: NA
config.metadata.Tissue: Thymus
config.metadata.Strain: F344_DuCrl
config.metadata.Sex: M
config.metadata.Computed Sex: M
config.metadata.Sample ID: 394_3T
config.metadata.Sample Characteristic: Control
config.metadata.PMID: NA
config.metadata.Project Link: https://rgd.mcw.edu/rgdweb/hrdp_panel.html
config.metadata.Data Processing: RGD workflow
config.metadata.Read alignment: STAR version=2.7.10b_alpha_220111, mRatBN7.2
config.metadata.Expression Quantification: RSEM v1.3.1
config.adapter.type: BigWigAdapter

Figure 12: RNAseq profiles for gastrocnemius, left retroperitoneal fat pad and thymus of a male F344/DuCrl rat aligned to mRatBN7.2 in JBrowse2. Data have recently been released for public viewing from the track menu at https://rgd.mcw.edu/jbrowse2 . Profiles generated with alignment to GRCr8 and detailed track information (inset) will be available in a future release of JBrowse2.



Grey: expression level is below cutoff (0.5 FPKM or 0.5 TPM)
Light blue / Turquoise : expression level is low (between 0.5 to 10 FPKM or 0.5 to 10 TPM)
Medium blue: expression level is medium (between 11 to 1000 FPKM or 11 to 1000 TPM)
Dark blue: expression level is high (more than 1000 FPKM or more than 1000 TPM)

Figure 13: Gene (center row) and transcript (bottom row) level TPM tracks in JBrowse2. Expression levels are color-coded as below cutoff, low, medium, and high as defined by Expression Atlas (threshold ranges displayed in the inset). Transcripts or genes with a TPM value equal to zero are not displayed.

Acknowledgements: We gratefully acknowledge our funding support from the National Institutes of Health (R01HL064541, U24HG010859, R24OD024617) and the researchers who faithfully use our website and data!

MEDICAL COLLEGE OF WISCONSIN

GLOBAL CORE BIODATA RESOURCE

ALLIANCE of GENOME RESOURCES FOUNDING MEMBER

This and other recent RGD presentations will be freely available for viewing and download in RGD's Presentations Archive. https://rgd.mcw.edu/wg/com-menu/poster_archive/