# Expansion and Standardization of Rat Expression Data at the Rat Genome Database

Wendy M. Demos[1*], Kent C. Brodie[2], Jeffrey L. De Pons[1], Adam C. Gibson[1], Varun Reddy Gollapally[1], G. Thomas Hayman[1], Mary L. Kaldunski[1], Akhilanand Kundurthi[1], Logan Lamers[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], 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*

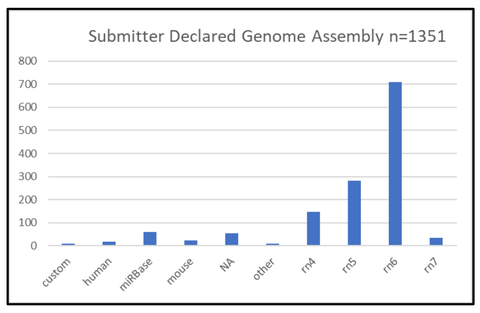
https://rgd.mcw.edu

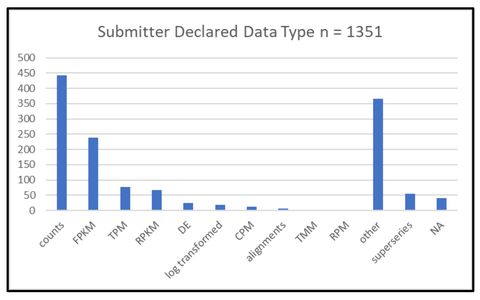
## Abstract

The Rat Genome Database (RGD) is expanding and incorporating expression data content into the larger ecosystem of RGD so users can seamlessly query for coherent gene information across portals. Researchers will be able to access expression metadata and values that were submitted to public resources such as the Gene Expression Omnibus (GEO) repository, with all expression values converted to transcripts per million (TPM). 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. The expression curation tool relies on a pipeline that imports metadata from the GEO Accession Display and utilizes NLP to match ontology terms to GEO series attributes. Curators can enter missing terms, confirm the predicted term, or provide a more specific term when appropriate. Fields for descriptors such as tissue type, vertebrate trait, clinical measurement, strain, cell type, experimental condition, etc. are built into the user interface. To enhance operational efficiency, the curation process begins at the project level interface, where ontology terms are entered a single time and then propagated across all applicable samples. Curators conduct a sample level review and edit terms on a per-sample basis as needed. When the metadata are correct and as complete as possible, they are loaded into the appropriate tables in RGD's relational database. The expression values submitted for the curated GEO series will be loaded for all genes and transcripts in the corresponding files. Currently, RGD has imported 1,859 GEO series related to *Rattus norvegicus* expression studies. Of those, 1,351 have been reviewed and prioritized for curation. To date, metadata for 165 GEO series have been uploaded. Expression value types submitted to the repository represent a wide range of analysis outputs (i.e., FPKM, counts, $log_2FC$). The submitter declared genome assemblies in the reviewed GEO series include versions RGSC3.4 -mRatBN7.2 as well as custom and non-rat references. The lack of standardization in the repository makes it difficult to identify rat data and furthermore, correlate expression values across studies. The goal of Phase Two is to standardize the expression values by developing and evaluating 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. This pipeline integrates quality control measures, alignment with the STAR[1] aligner, and abundance estimation with the RSEM[2] software package. Phase Three will focus on enhanced visualization of expression values. The current tabular-based view will be updated and new graphical visualizations at the gene and transcript levels are planned.

## Background

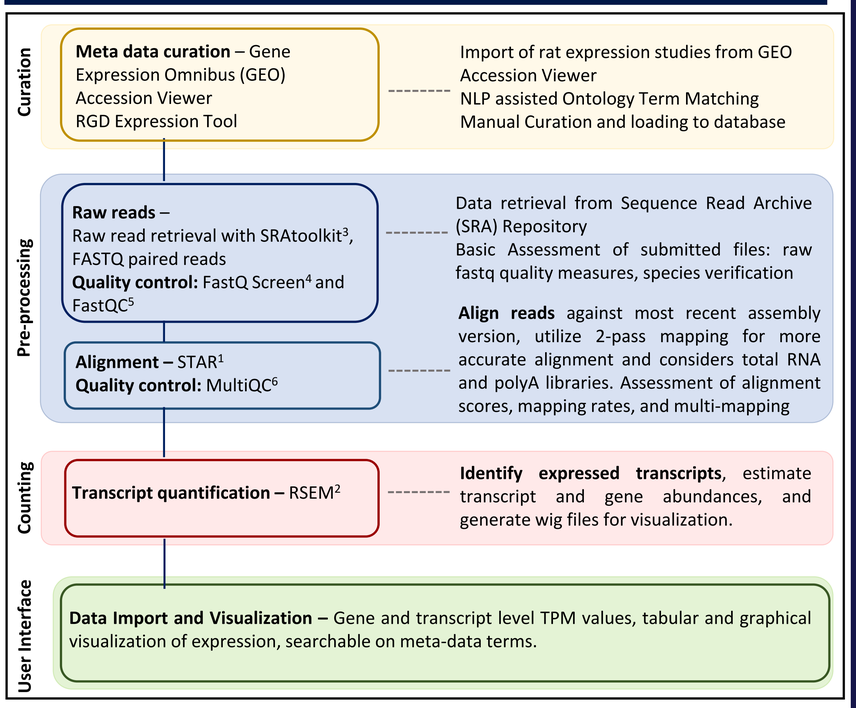
1,351 studies were reviewed at the time of writing the abstract for prioritization of meta-data curation. 58% are tagged for future curation due to stringent first pass curation criteria (data type = TPM, FPKM, source = non-culture, strategy = bulk RNA, illumina sequencing). 22% of the reviewed studies are not associated with published articles A variety of genome assemblies were reported by GEO Submitters for each GEO Accession. (Assembly names are simplified for ease of viewing in the graph and are not indicative of Ensembl or NCBI resources).



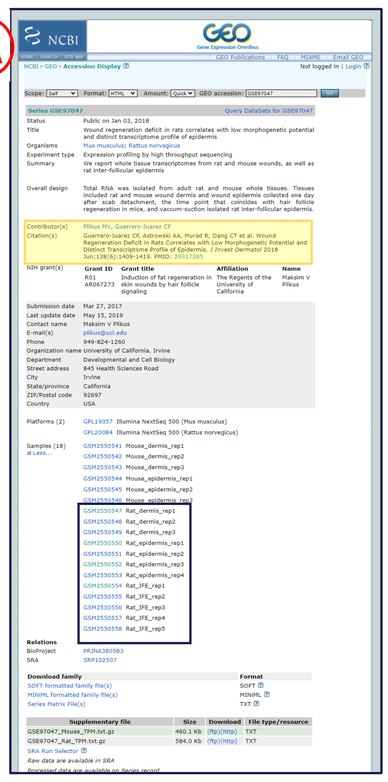




Submitters to GEO declare the data format type for the provided results files. The most common data type in the 1,351 reviewed studies fall into the O*ther* category. The *Other* category is non-FPKM or TPM data such as peak data, alignment files, UMI counts, etc. *SuperSeries* category may have multiple data types as a SuperSeries accession may have many SubSeries each with a different data type. At times, a data type will be provided, but upon review, the submitted data do not reflect the type, or there is a discrepancy between the filename and declared type (For example RPKM data are provided but the filename is ACCESSION.fpkm.txt).

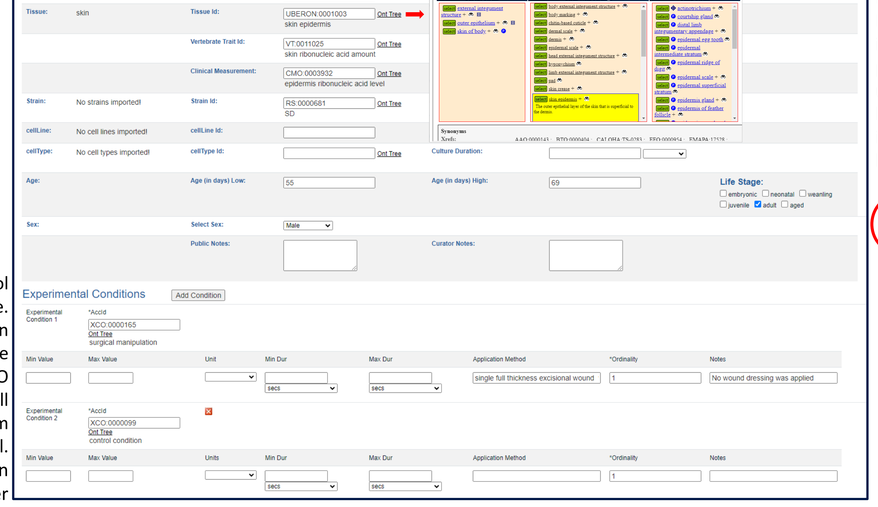
## Methods



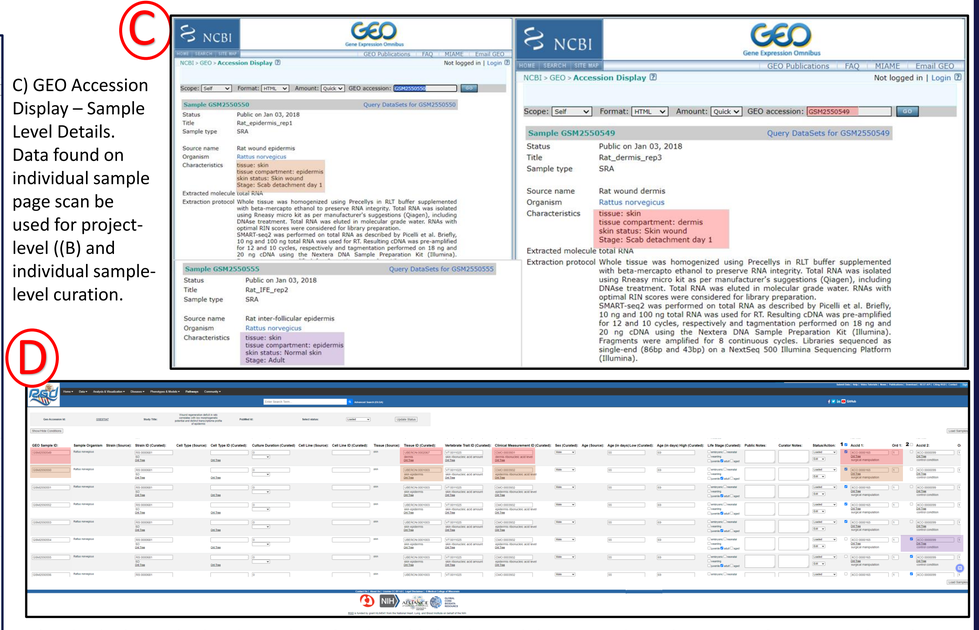

## Phase 1: Metadata Curation




A) GEO Accession Display – Landing page for a given Geo Series. Experiment summary, submitter contact information, and publication ID (if available) are provided. Links to the individual sample detail web pages are provided (Boxed in blue.). Near the bottom of the page, download links for the entire series may be available, or data may be provided on a per-sample basis (https://www.ncbi.nlm.nih.gov/geo/query/ ).




Wound regeneration deficit in rats correlates with low morphogenetic potential and distinct transcriptome profile of epidermis

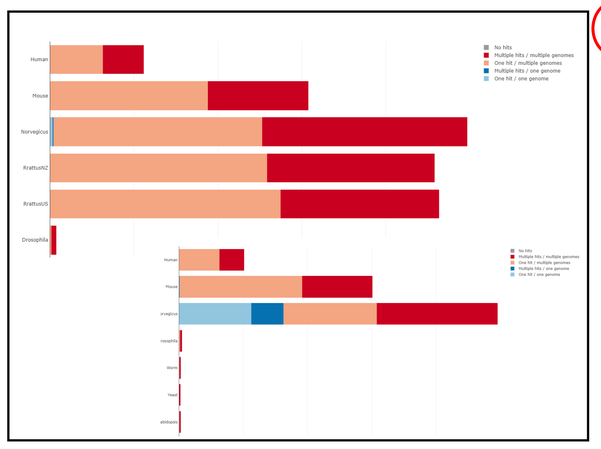



C) GEO Accession Display – Sample Level Details. Data found on individual sample page scan be used for project-level ((B) and individual sample-level curation.

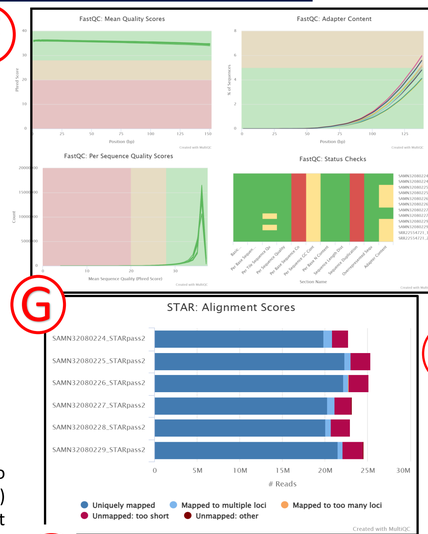



D) Individual sample details can be refined. Highlighted in red, brown, and purple are specific terms and conditions related to samples GSM2550549, GSM2550550, and GSM2550555 as displayed on the GEO Accession display (C). This differs from what was entered on the project-level page in the curation tool. Once the curation is complete, the Curator will select the "Load Samples" button in the upper right corner and meta-data are loaded to RGD's relational database.

## Phase 2: Data Re-analysis

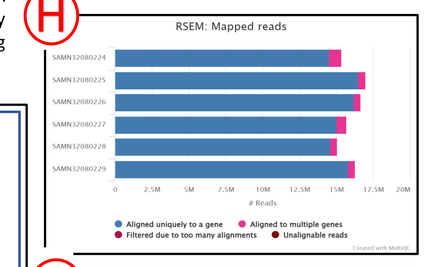


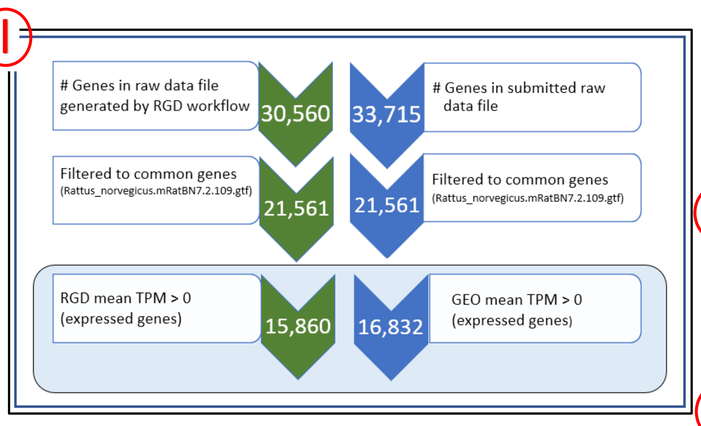

E) The Pre-processing stage of the data re-analysis pipeline utilizes FastQ Screen[4] to verify the organism matches what is provided in the GEO Accession Viewer. (Inset) Both *R. norvegicus* and *R. rattus* assemblies are included in the tool to verify rat species. description and F) FastQC[5] assesses the general quality of the SRA downloaded and converted FASTQ files (Data for all samples in series compiled by MultiQC[6]) . MultiQC[6] is used to assess the G) STAR[2] alignment and H) RSEM[1] mapping qualities.

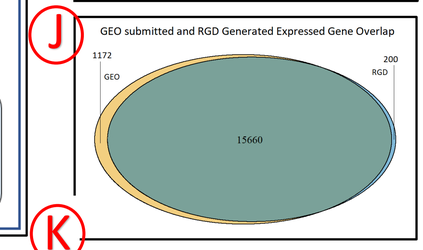


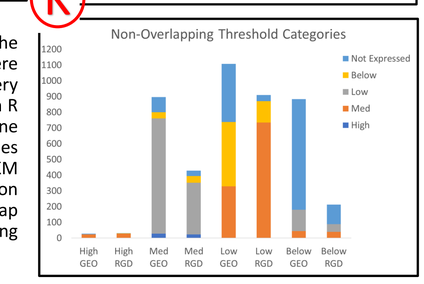






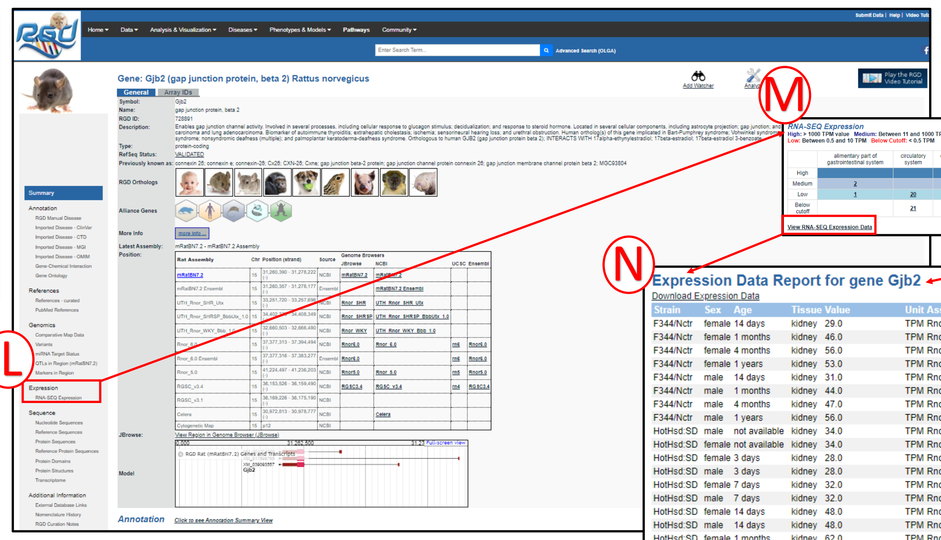



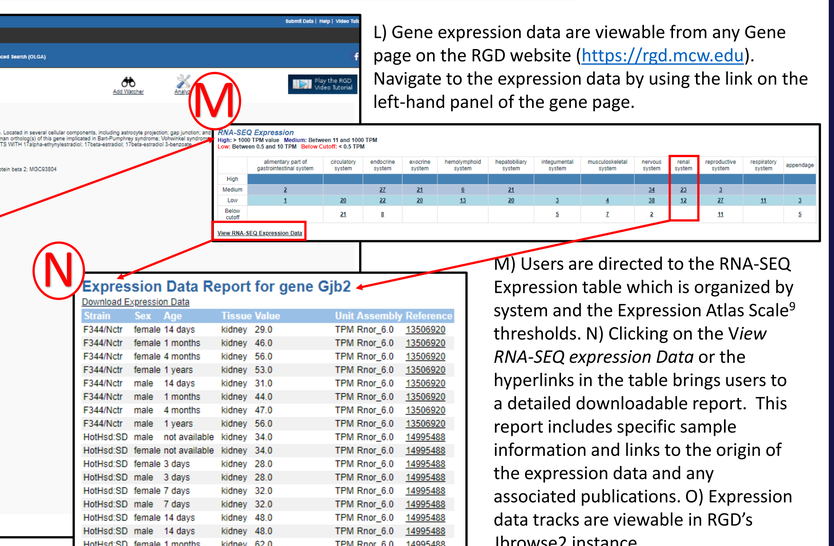




I) Data for Accession GSE220261[7] was used as a test to compare output by the RNAseq re-analysis pipeline. FPKM and count data were submitted to GEO and there is no associated publication. Other than knowing the data were aligned to "rn7", very few details were provided about the data. FPKM data were converted to TPM with R library GeoTcgaData[8]. Data were normalized by limiting the analysis to gene definitions in the Ensembl v109 gtf. J) The overall overlap of expressed genes between the submitted and RGD pipeline results was almost 92% for both the FPKM and TPM analysis. K) Using the expression thresholds as defined by the Expression Atlas Scale[9] the genes categorized as medium expression had the highest overlap rate (85%). The graph shows the count and classification for non-overlapping expression results within each threshold category.

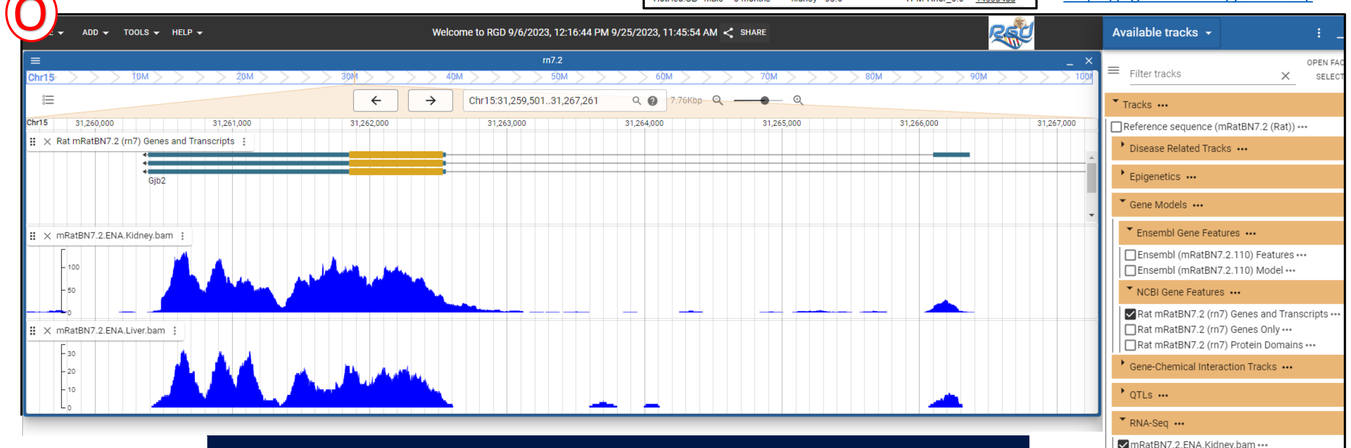
## Phase 3: Data Visualization




L) Gene expression data are viewable from any Gene page on the RGD website (https://rgd.mcw.edu). Navigate to the expression data by using the link on the left-hand panel of the gene page.




M) Users are directed to the RNA-SEQ Expression table which is organized by system and the Expression Atlas Scale[9] thresholds. N) Clicking on the *View RNA-SEQ expression Data* or the hyperlinks in the table brings users to a detailed downloadable report. This report includes specific sample information and links to the origin of the expression data and any associated publications. O) Expression data tracks are viewable in RGD's Jbrowse2 instance https://rgd.mcw.edu/jbrowse2/




## References

1. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras T: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2012, Jan:29(1): 15-21
2. Li B, Dewey C: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011 (12), 323
3. SRAtoolkit SRA Toolkit Development Team Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
4. Wingett SW and Andrews S. **FastQ Screen: A tool for multi-genome mapping and quality control [version 2; referees: 4 approved].** *F1000Research* 2018, 7:1338
5. Andrews S. **FastQC: a quality control tool for high throughput sequence data**. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc
6. Ewels P, Magnusson M, Lundin S, Käller M: **MultiQC: Summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics* (2016)
7. Li X: **GSE220261** Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220261
8. Hu E, **GeoTcgaData**. Available from https://github.com/YuLab-SMU/GeoTcgaData
9. EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. https://www.ebi.ac.uk/gxa/help/index.html

**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!

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