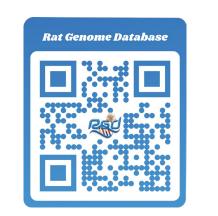# Rat ReMap Initiative at the Rat Genome Database (RGD): A Resource to Integrate Transcriptomic and Epigenomic Data for Laboratory Rats

Aron M. Geurts, Monika Tutaj, Wendy Demos, Jeffrey De Pons, Logan Lamers, Andrew S. Greene, Sridhar Rao, Allen W. Cowley, Jr., Mingyu Liang, Pengyuan Liu, and **Anne E. Kwitek**

Medical College of Wisconsin, Department of Physiology and the Rat Genome Database, Milwaukee WI 53226

https://rgd.mcw.edu

INTRODUCTION:

- Human genome wide association studies (GWAS) have identified over 1000 loci associated with blood pressure or hypertension. However, many of these loci are intergenic and the mechanisms leading to blood pressure variation are largely unknown, thus requiring animal models and translational studies

- Translating non-coding mechanisms between human and animal model systems such as the laboratory rat are complicated by a general lack of multi-omic integration both within and across species

- For the rat, few studies are mapped to the current genome reference (mRatBN7.2)

- The goal of the Rat ReMap Initiative is to establish an integrated resource within the Rat Genome Database (RGD) for publicly available rat genomic, transcriptomic and epigenomic data.

- The Rat ReMap Initiative involves:
  - Comprehensive curation of the public datasets
  - Reanalysis of targeted multi-omic data sets to map them to the most current rat reference genome
  - Data integration and visualization of data within the rat genome and through comparative genomics to human

METHODS:

1) Established a curation schema and an interface tool to standardize GEO metadata for transcriptome and epigenome data and store in the RGD Oracle database

2) Obtained publicly available data sets from the Short Read Archive (SRA) or Ensembl including rat liver transcriptome (PRJEB6938), ATAC-seq (PRJNA684678), and ChIP-seq for CTCF (PRJEB2329), transcription factors (PRJEB1571, PRJEB1243), and histone marks (PRJNA454317, PRJNA540150). Data were reanalyzed to map each to the mRatBN7.2 rat genome
   a) RNA-seq data was reanalyzed at Ensembl by their standard RNA-Seq Pipeline and made available for download (https://ftp.ensembl.org/pub/rapid-release/species/Rattus_norvegicus/GCA_015227675.2/ensembl/rnaseq/)
   b) The ENCODE ATAC-seq standard pipeline was used for quality control and statistical signal processing of ATAC-seq data, producing alignments and measures of enrichment (https://www.encodeproject.org/atac-seq/)
   c) ChIP-seq data were trimmed and processed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and were then aligned using Bowtie 2 (PMID: 22388286). Peaks were called from the aligned BAM files using MASC2 (PMID: 22936215)

3) Implemented a new genome browser - JBrowse2 - to visualize genes, transcripts, transcriptomic and epigenome data, and phenotype data curated by RGD

4) Developed a comparative genomics visualization tool – VCMap – of evolutionarily conserved synteny between rat and human integrating genes with variant and epigenome density plots

FINANCIAL DISCLOSURE AND ACKNOWLEDGEMENTS:

RESULTS:

Metadata curation of GEO records to tag records with standardized nomenclature and Ontology terms

Curation Tool Landing Page. This page displays projects identified by Unique GEO Accession IDs. Projects are sorted by *Pending* (to be curated), *Loaded* (curation completed), *Not For Curation* (does not meet curation criteria), and *Future Curation* (missing Information, not priority).

Project Level Curation. Some metadata can be pre-populated by the loading process. Metadata terms for the GEO Series (highlighted in yellow) are edited or added based on information found on the GEO Accession Display website, in the associated publication(s), or if necessary, confirmation through direct contact with the author(s). If a submitter provides more than 1 citation, additional PMIDs can be attached to the data.

GEO Accession Display – Landing page for a given Geo Series. Links to the individual sample details are provided (Boxed in blue.). (https://www.ncbi.nlm.nih.gov/geo/query/).

GEO Accession Display – Sample Level Details. Data found on this page are used for project-level (highlighted in yellow) and individual sample-level curation (highlighted in red).

Project level metadata are populated to samples. Individual sample details can be refined. Highlighted in red is the specific experimental condition related to GSM2550549 as displayed on the GEO Accession display. This differs from what was entered on the project level page in the curation tool (the curation tool did not import multiple UBERON terms).
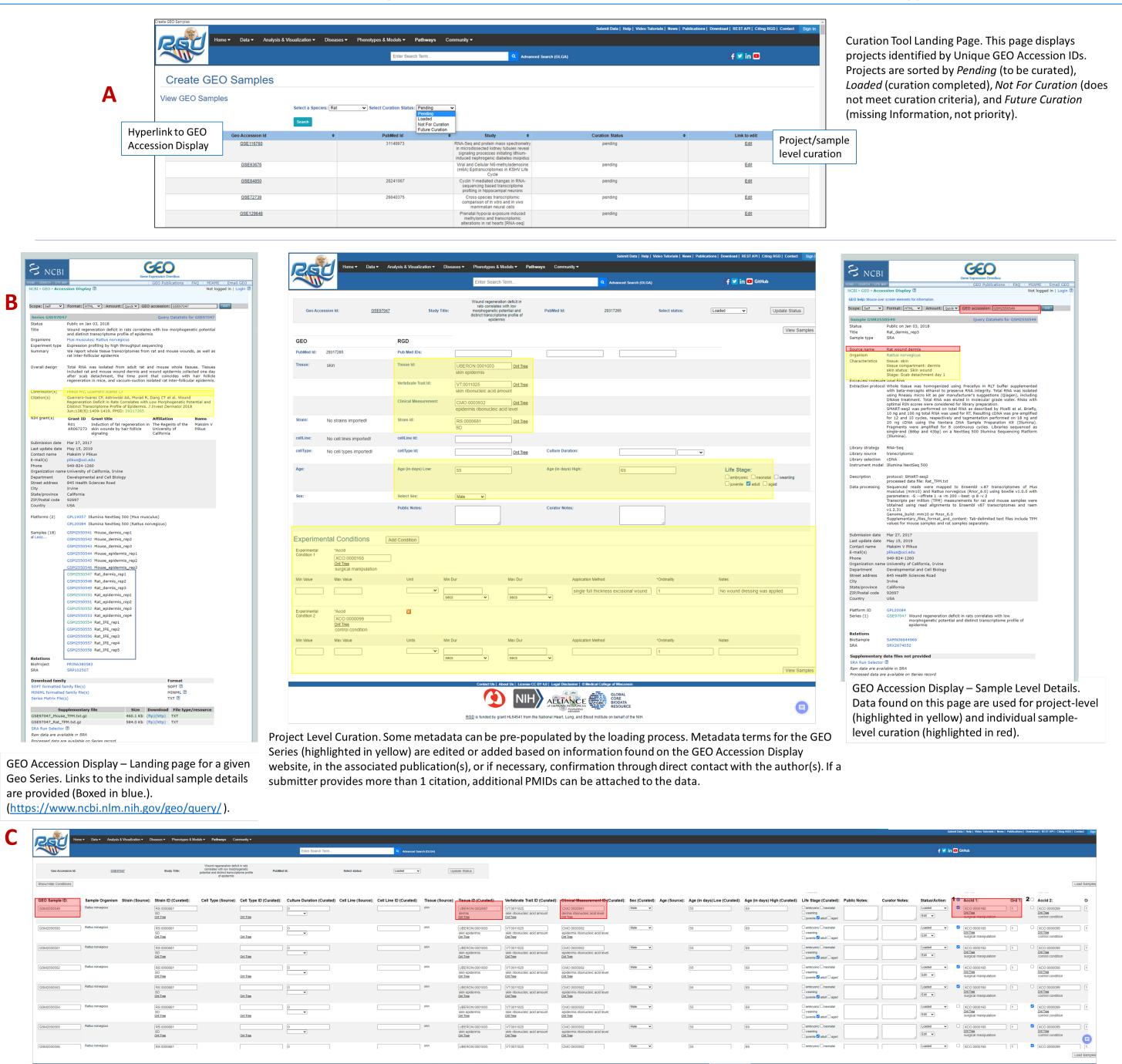
Fig 1. Metadata Curation of GEO Records. A. Curation tool landing page with preliminary information about the study including accession ID, PMID. B. Study attributes automatically assigned in ETL pipeline with additional fields for manual validation and additional curation including age, sex, and standardized Ontology terms for strain, tissue, what was measured, and under what conditions. C. Sample attribute page for sample specific attribute editing. To date, 723 studies were loaded, reviewed and prioritized, and 100 records passed the initial filters and have been fully curated. 589 marked for future curation. An additional 1135 studies were recently loaded for review.
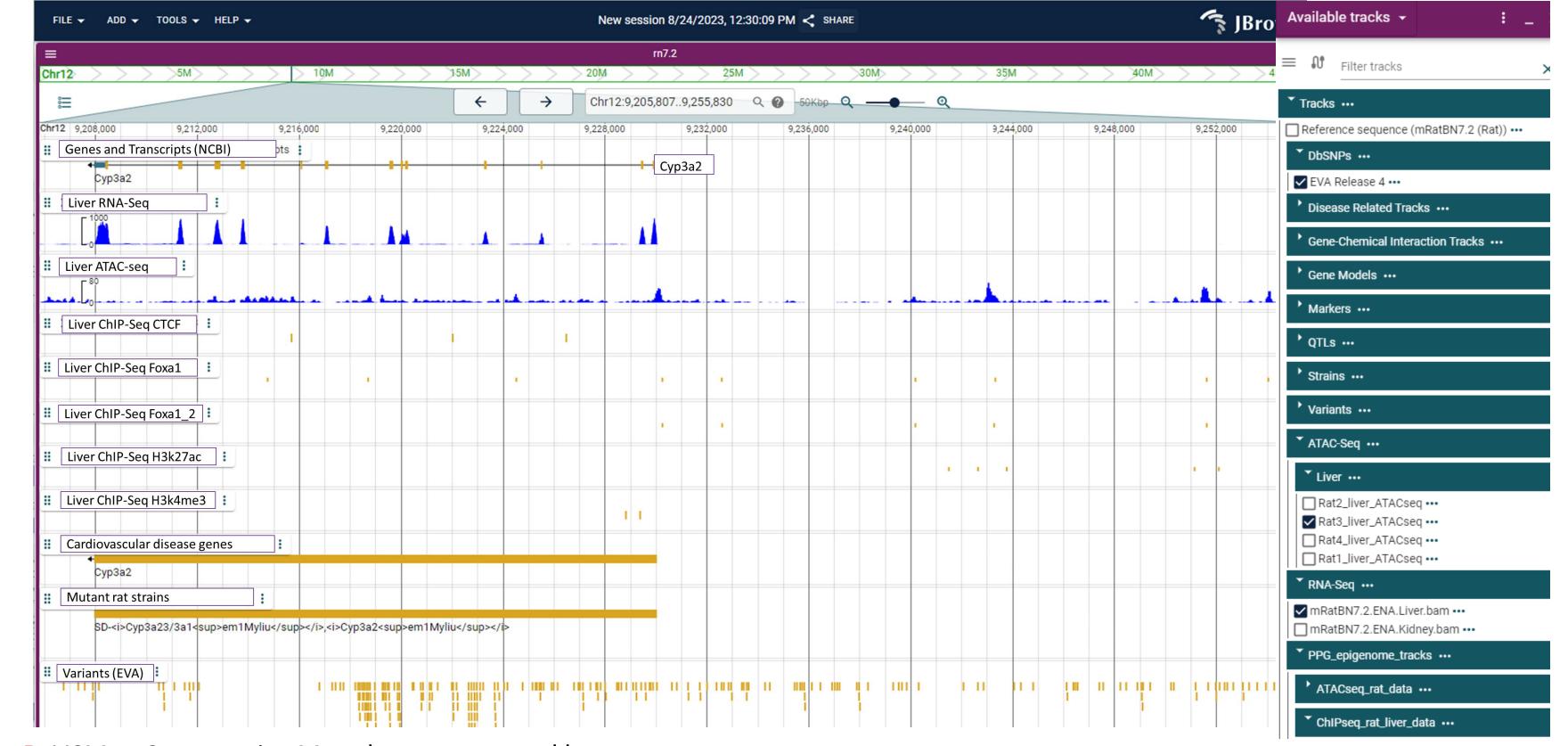
Mapping rat multi-omic data to mRatBN7.2 in Jbrowse 2 and comparative mapping between species with VCMap

A. JBrowse2 view of mRatBN7.2

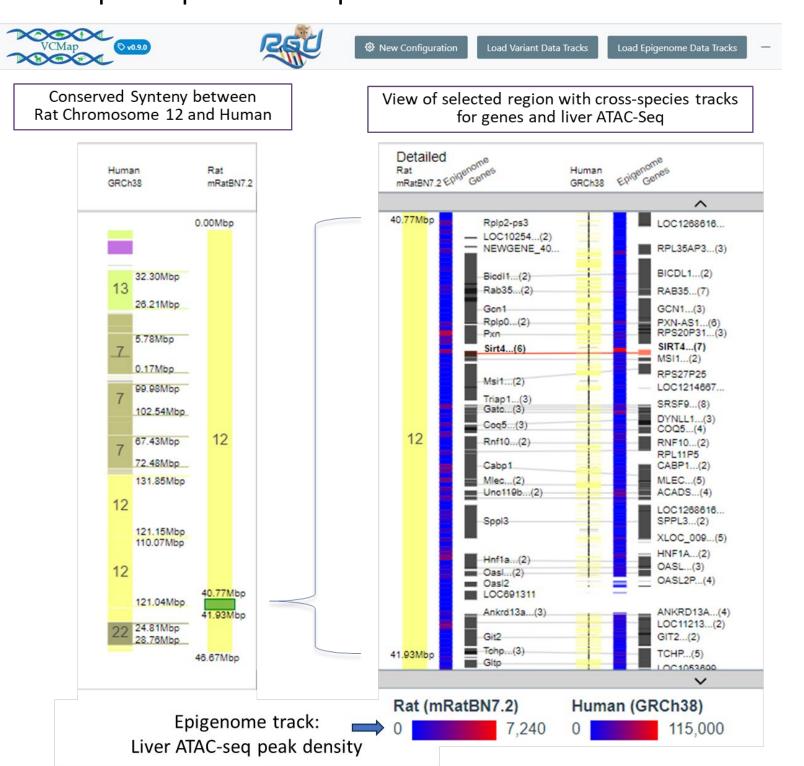**B.** VCMap Comparative Maps between rat and human

Fig 2. Remapping and data integration of multi-omic data. Data from RNA-Seq (2 tissues), ATAC-Seq (25 tissues), and ChIP-seq (liver) for transcription factors (Ctcf, Cebpa, Foxa1, Hnf4a, Hnf6) and histone modifiers (H3k27ac, H3k4me1, H3K27me3, H3K4me3) were remapped and loaded into genome browsers. A. JBrowse2 view of mRatBN7.2 with tracks for gene, transcriptome, epigenome, disease, strain, and variant data B. VCMap view of conserved synteny between rat and human with gene orthology connections and liver ATAC-seq density plots. The left panel shows rat chromosome 12 with the human conserved synteny to the left of it. A region of interest is highlighted in the green box. The right panel shows a detailed view of the selected region with the rat genome on the left and the human conserved synteny on the right. Lines connect orthologous gene pairs. Epigenome tracks are density plots of liver ATAC-seq data with red color indicating high number of peaks for rat and human in their respective tracks.

CONCLUSIONS:

- Automation assisted manual curation of GEO metadata with standardized nomenclature and Ontologies (e.g. for tissue, cell, measurement type, experimental conditions) allows for multi-omic data integration and visualization

- JBrowse 2 accepts multiple data formats being implemented at RGD with comprehensive genetic, transcriptome, and epigenome tracks

- VCMap is a novel comparative map tool showing conserved synteny. Species specific information is aligned to each species allowing comparisons of gene orthologs along with density tracks for variants and epigenomic data that will link out to more detailed information