Automated Data Pipelines for Loading, Integrating, Annotating and Quality Control of Data at the Rat Genome Database

Marek Tutaj, Mary Shimoyama, Elizabeth A. Worthey, Jennifer Smith, Rajni Nigam, Victoria Petri, Stan Laulederkind, Timothy F. Lowry, Tom Hayman, Shur-Jen Wang, Jeff De Pons, Pushkala Jayaraman, Weisong Liu, Diane Munzenmaier, Melinda Dwinell, Simon Twigger, Howard Jacob
Rat Genome Database, Human and Molecular Genetics Center, Medical College of Wisconsin

Abstract

As the richness and diversity of biological data increase, model organism databases are confronted with the problem of quickly and efficiently populating their databases as well as providing timely updates to the information that they store. The Rat Genome Database (RGD, http://rgd.mcw.edu) provides comprehensive rat genetic, genomic and biological data through both manual and automated curation processes. A series of automated data pipelines have been implemented to acquire various data types from multiple sources, integrate them with existing data and provide comprehensive quality control data in order to maximize data coverage and reserve manual processes for targeted curation projects for data unavailable anywhere except the literature. Data acquired through these pipelines include 1) basic genomic elements such as genes and accompanying map, sequence and external database identifiers, protein information, genomic positions of exons and coding regions, 2) orthologs and ortholog relationships, 3) nomenclature alerts and reviews, 4) Gene Ontology annotations for human and mouse orthologs stored in RGD as well as appropriate annotations to rat genes, 5) ontology terms and relationships for GO, Mammalian Phenotype Ontology and Pathway Ontology. The pipelines at RGD are run with either incremental updates or delete-and-reload mechanisms and are run weekly to keep data up to date and synchronized with originating data sources. Pipeline mechanisms, quality control measures, and the methods to time and synchronize multiple pipelines will be presented along with the data types acquired and integrated and the process for resolving data errors and conflicts discovered during the QC processes.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Ontology Loading Pipeline	•Strain FTP Extract •QTL re-mapping and FTP Extract •Gene FTP Extract	RGD Terms Reindexing (for search engine)	GOC Annotations FTP Extract	Entrezgene Pipeline (rat, human, mouse) Ortholog Loading UniProtKB spiotr/trembl	Process GOC Annotations	
	GO Annotation Pipeline		Mouse and Human GO Annotation UniSTS Pipeline		Data Release	Data Release (cld.)

Fig 1. Synchronization of multiple pipelines at RGD assures data consistency while weekly runs providing timely updates.

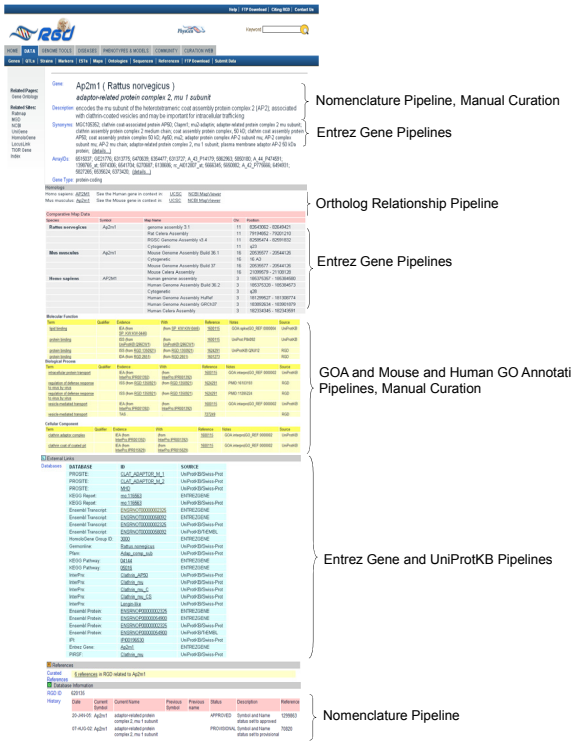

Fig 2. RGD pages are populated by multiple automated pipelines for various data types.

1. Gene QC at RGD

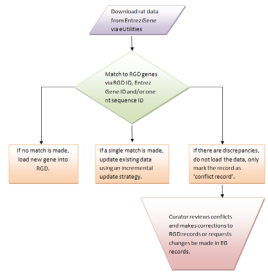

Fig 3. Entrez Gene Pipelines automatically query the NCBI databases for gene records that have been modified during last week and make the necessary updates to RGD genes.

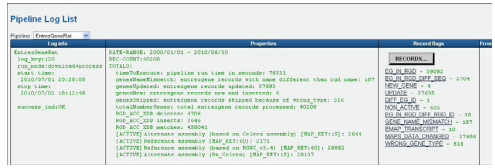

Fig 4. Pipeline web report pages show the results and summaries from the last runs. A set of flags is assigned to every record (gene) processed, so curator can quickly jump to group of records of interest.

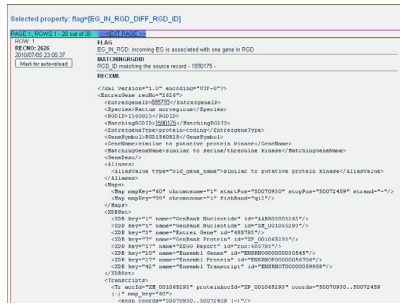

Fig 5. The curator can browse through particular class of conflicting records. The full hyperlinked XML representation of the incoming record is provided to allow for faster conflict resolution.

Types of conflicts:

- Entrez Gene/RGD ID pair do not agree between Entrez Gene and RGD.
- Entrez Gene ID is assigned to multiple RGD IDs.
- None of the nucleotide sequence IDs match between Entrez Gene and RGD.

Possible reasons for the conflicts:

- Entrez Gene has merged genes and RGD hasn't.
- RGD has merged genes and Entrez Gene hasn't.
- RGD has allele or splice variant records which haven't been designated as "allele" or "variant".
- Entrez Gene/Entrez Nucleotide has replaced existing nucleotide sequences with sequences having other identifiers.
- Entrez Gene has converted a gene from "protein-coding" to "pseudo" and changed or removed GenBank sequences.

2. Orthologs and Nomenclature QC at RGD

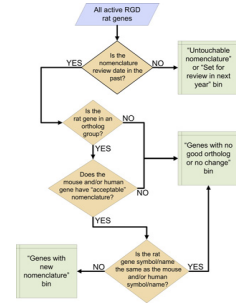

Fig 6. RGD Nomenclature Pipeline ensures nomenclature QC via proposing necessary nomenclature changes every time the tool is opened.

- Prevents problems with a delay between when the pipeline is run and when a curator can review results.
- "Binning" the output allows the curators to review gene nomenclature in any category, not just the ones for which the tool proposes changes.
- RGD Ortholog Relationship Pipeline keeps human-mouse-rat orthology information up to date with its delete-reload mechanism.

Fig 7. RGD Nomenclature Curation Software

- Provides rapid and efficient updates of rat nomenclature.
- A single queried gene or genes in bulk can be updated.
- Leverages the mouse and human Entrez Gene and ortholog pipelines to simplify the rat gene nomenclature review process.
- Nomenclature for >4000 genes is updated in less than 3 weeks.
- Compares rat nomenclature to mouse and human and proposes an updated symbol and name.
- Proposed new nomenclature can be edited.
- Nomenclature review date and reference is set.

3. Assuring Gene Identity in Multiple RGD Pipelines

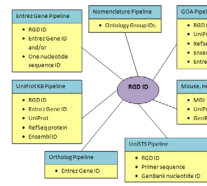

Fig 8. The pipelines at RGD use cross referencing identifiers, which expedites data exchange between different resources and ensures gene identity.