RGD: A comparative genomics platform

Mary Shimoyama, Jennifer R. Smith, Tom Hayman, Stan Laulederkind, Tim Lowry, Rajni Nigam, Victoria Petri, Shur-Jen Wang, Melinda Dwinell, Howard Jacob and RGD Team*

Rat Genome Database, Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA
**Correspondence to:* Tel: +1 414 456 7505; Fax: +1 414 456 6516; E-mail: shimoyama@mcw.edu

Date received (in revised form): 10th August 2010

Abstract

The Rat Genome Database (RGD) (http://rgd.mcw.edu) provides a comprehensive platform for comparative genomics and genetics research. RGD houses gene, QTL and polymorphic marker data for rat, mouse and human and provides easy access to data through sophisticated searches, disease portals, interactive pathway diagrams and rat and human genome browsers.

Keywords: *genomics, database, disease*

Introduction

The Rat Genome Database (RGD) (http://rgd.mcw.edu) is recognised as the premier resource for genetic, genomic and phenotype data for the laboratory rat, *Rattus norvegicus*. Since 1999, RGD has provided a comprehensive catalogue of genes, quantitative trait loci (QTL) and strains, along with software tools to retrieve and display data of interest to investigators using this organism. The disease focus of these researchers often results in the use of multiple model organisms, in addition to clinical studies, in their efforts to elucidate the mechanisms and underlying genetic factors involved in human disease. To meet the needs of such users, RGD focuses its manual curation efforts on the functional, phenotype and pathway data related to specific disease areas and has integrated human and mouse data to create a comprehensive platform for comparative genomics and genetics. Several of these components are highlighted here.

Disease portals

The wealth of data at RGD includes genes and QTLs for rat, human and mouse, as well as polymorphic markers for rat and human (Table 1). Information on inbred, outbred, mutant, congenic, consomic and other types of rat strains is also provided. A team of scientific curators validates the identity of genomic elements, provides official nomenclature and annotates these elements with functional data from published literature.[1] With more than 1.3 million published rat research papers, prioritising data for curation is a vital task, and a project approach has proved effective. Such projects revolve around gene families, molecular pathways, ultra-conserved gene sets and diseases.

The disease portals (http://rgd.mcw.edu/wg/portals/) create a structure for prioritising rat data curation and integrating rat, human and mouse information, and provide a platform for researchers easily to access multiple data types related to a particular disease area (Table 2). RGD currently has portals for cardiovascular and neurological diseases, cancer, diabetes obesity/metabolic syndrome. For each portal, a list of genes is generated from literature and database searches, and these are prioritised by the weight of evidence suggesting disease association. All rat papers for each gene are curated and ontology-based annotations made for function, biological process, cellular component, pathway, phenotype and disease. In addition to the Gene Ontology, RGD uses the MeSH terms for disease, and the Mammalian Phenotype Ontology and the

Table 1. Major data elements in RGD. RGD houses data for rat, human and mouse genes and quantitative trait loci (QTL), rat and human polymorphic markers and rat strains. This table gives the number of each data element for each species contained in the RGD database

	Genes	QTL	Polymorphic markers	Strains
Rat	39,818	1766	13,142	2155
Human	28,855	1911	293,551	
Mouse	40,106	998		

Pathway Ontology for additional manual annotations. While rat literature provides the bulk of manual annotations, appropriate human papers providing experimental evidence of disease association are also used for disease annotations. Gene Ontology annotations for rat are manually curated, while those for mouse and human are imported from the Gene Ontology Consortium website (http://www.geneontology.org/).[2] Currently, there are over 10,000 rat, human and mouse genes with disease annotations. Targeted signalling pathways are curated and interactive diagrams created as part of the portal development, with more than 3,800 rat, human and mouse genes with pathway annotations. Further information, including map data, links to corresponding nucleotide and protein sequences and reports at outside databases — including Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim), Human Protein Reference Database (http://www.hprd.org), Uniprot (http://www.uniprot.org), Mouse Genome Informatics (MGI) (www.informatics.jax.org) and EntrezGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) — are provided on the RGD rat, human and mouse gene reports. Both human and rat QTL are acquired through literature curation and rat strains used as disease models are highlighted and phenotypes described. Users can search major disease categories or subsets and return lists of appropriate genes and QTLs for rat, mouse and human, with direct links to individual report pages in RGD (Figure 1). The portals provide a genome-wide view of genes and QTLs related to particular diseases, pathways or phenotypes with functionality for mouse, human and rat synteny views. Each portal also has a component highlighting the major strain models and their associated phenotypes, along with links to strain reports and associated references.

Table 2. Numbers of genes, quantitative trait loci (QTL) and strains in disease portals. Each disease portal has a large number of manually curated rat, mouse and human genes, rat and human QTL and rat strains which have been experimentally shown to be related to the disease category covered by the portal

		Rat	Human	Mouse
Cancer portal (breast, prostate, urogenital)	Genes	523	532	521
	QTL	69	352	
	Strains	40		
Diabetes portal	Genes	890	893	868
	QTL	702	645	
	Strains	245		
Cardiovascular disease portal	Genes	838	875	813
	QTL	473	114	
	Strains	276		
Neurological disease portal	Genes	862	841	805
	QTL	116	35	
	Strains	115		
Obesity/ metabolic syndrome portal	Genes	988	988	957
	QTL	717	840	
	Strains	249		

Phenotype and models portal

The phenotype and models portal (http://rgd.mcw.edu/wg/physiology) provides a more extensive resource for those using, or interested in using, rat in their studies (Figure 2). Both veteran rat researchers and clinical researchers looking for a model for a particular human disease will find what they need at this site. The strains and models section provides links to information and sources for cages, rat food, colony management software, animal identification systems and protocols for

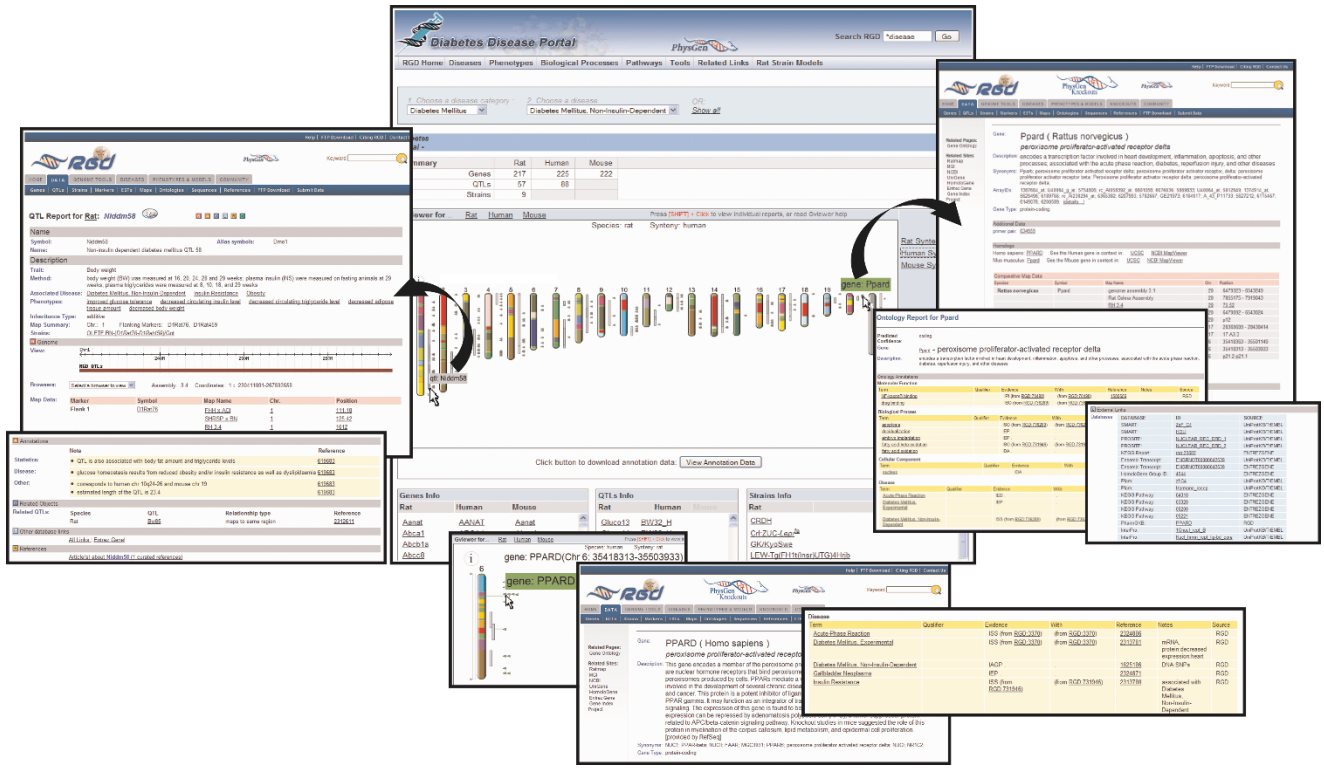

Figure 1. Each disease portal is designed to consolidate data for a particular category of disease — both across species and across data types. The inclusion of data for rat, mouse and human, and syteny views of the three genomes, facilitate cross-species comparisons. Links to individual report pages for genes, strains and QTL give users a more complete view of their data of interest.

genotyping. It also provides an easy strain search tool with direct access to strain reports, as well as information on strain availability and sources. Detailed profiles of commonly used strains are provided in the strain medical records, including information on physical features, growth and development, reproduction and normal ranges for traits such as blood pressure, heart rate, glucose levels and others. Links to single nucleotide polymorphisms (SNPs), QTLs and microarray data make these records a valuable resource. The phenotypes section and PhenoMiner tool provide the necessary information to assist both novice and veteran researchers in choosing appropriate strains and developing adequate protocols for their studies. A snapshot of various trait values across multiple stressor conditions for the most commonly used strains is available, along with complete phenotype value data for hundreds of strains. Protocols for assessing major traits — and links to descriptions of methods, measurement equipment, induction agents and diagnostic test sets — make this section an excellent laboratory resource. Using the PhenoMiner tool, a researcher can compare phenotype values across multiple strains and multiple experimental conditions. Results can be downloaded into an Excel spreadsheet or viewed in a bar chart. Multiple phenotype values for a particular strain or strains can also be easily accessed. The phenotype and models portal provides the information necessary to choose the appropriate rat strain to function as a disease model and also allows users to compare the phenotype profile of these models with that for human disease.

Genome tools

RGD has adapted and developed a variety of online tools to assist researchers in analysing the data in RGD and that produced in their own

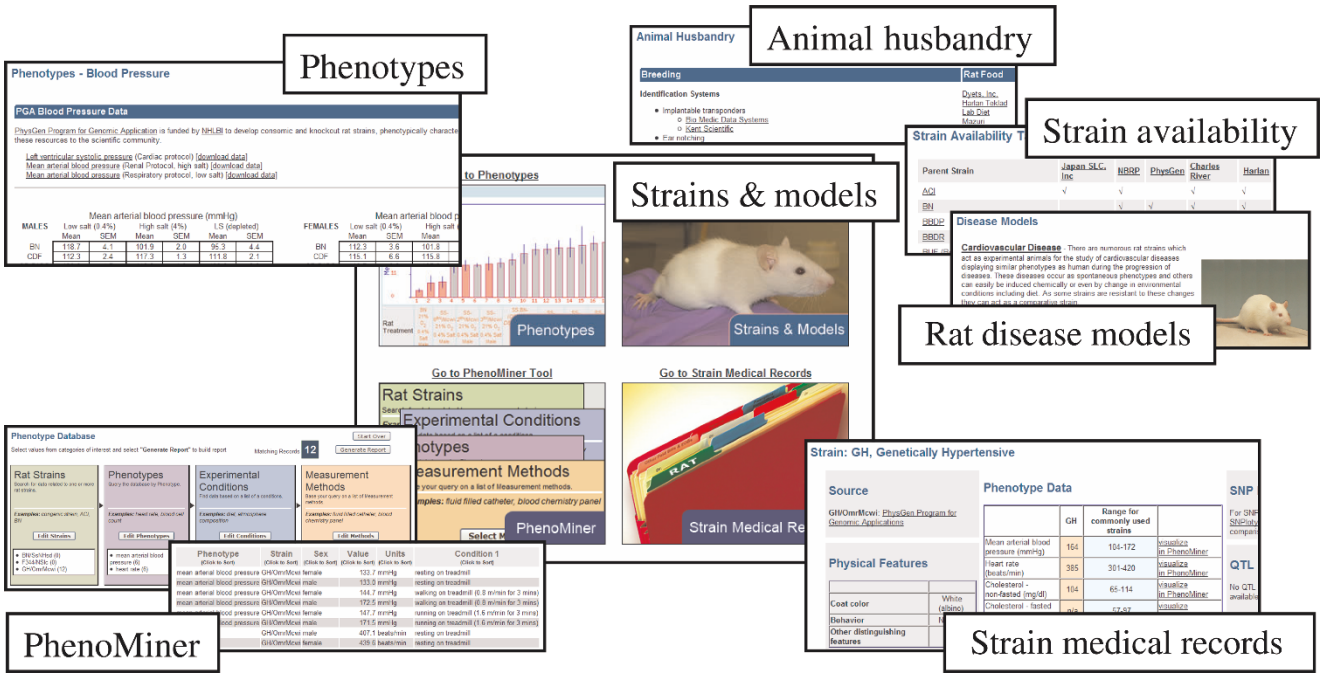

Figure 2. RGD's phenotypes and models portal is a valuable resource for anyone either currently working with, or considering working with, the rat as a disease model. The strains and models section gives tips and information about animal husbandry — such as identification systems, genotyping protocols and colony management software, strain availability and rat strains that are good models for human diseases such as cardiovascular disease and breast cancer. The strain medical records give detailed information about a number of important rat strains. The phenotypes section not only lists phenotypic values for many rat strains under a variety of conditions, but also gives detailed protocols so that researchers can do their own measurements. The PhenoMiner is a powerful and flexible tool which gives researchers the option to choose one or more rat strains, experimental conditions and phenotypes from a large number of options to obtain specific and detailed values for those phenotypes.

laboratories. These include the genome viewer, genome browsers for both rat and human, the rat SNPlotyper and RatMine (Figure 3).

GViewer provides users with a genome-wide view of rat genes and QTL retrieved from ontology-based searches. Single-term or Boolean searches across one or more ontologies allow for flexibility and ease of use for those unfamiliar with the ontologies. From the full genome view, users can zoom into a region to see additional details, and can add genes and QTL not retrieved with the initial search. The list of data objects in the display can easily be viewed, or exported for use in other applications. In addition, the GViewer provides links to the rat GBrowse (genome browser) for a closer view of genomic details.

Separate GBrowse applications for rat and human facilitate comparative analyses. The rat GBrowse includes a wealth of genomic data, including: gene models from RGD, the National Center for Biotechnology Information (NCBI) and Ensembl; genomic features such as QTLs, transfer RNAs, expressed sequence tags, and microRNAs; and SNPs from both NCBI's SNP database (dbSNP) and Ensembl. Additional information about each piece of data can be viewed as pop-up 'balloons' by mousing over the display. These balloons also include access to more complete information via links to records at RGD and other databases. Many of these features and data types are also available in RGD's human GBrowse, including human genes and QTLs. The Ensembl SNP track on the human GBrowse gives users the ability to browse through known polymorphisms in their genomic region of interest.

Rat, mouse and human synteny can be displayed in both the rat and human GBrowse tools. Regions

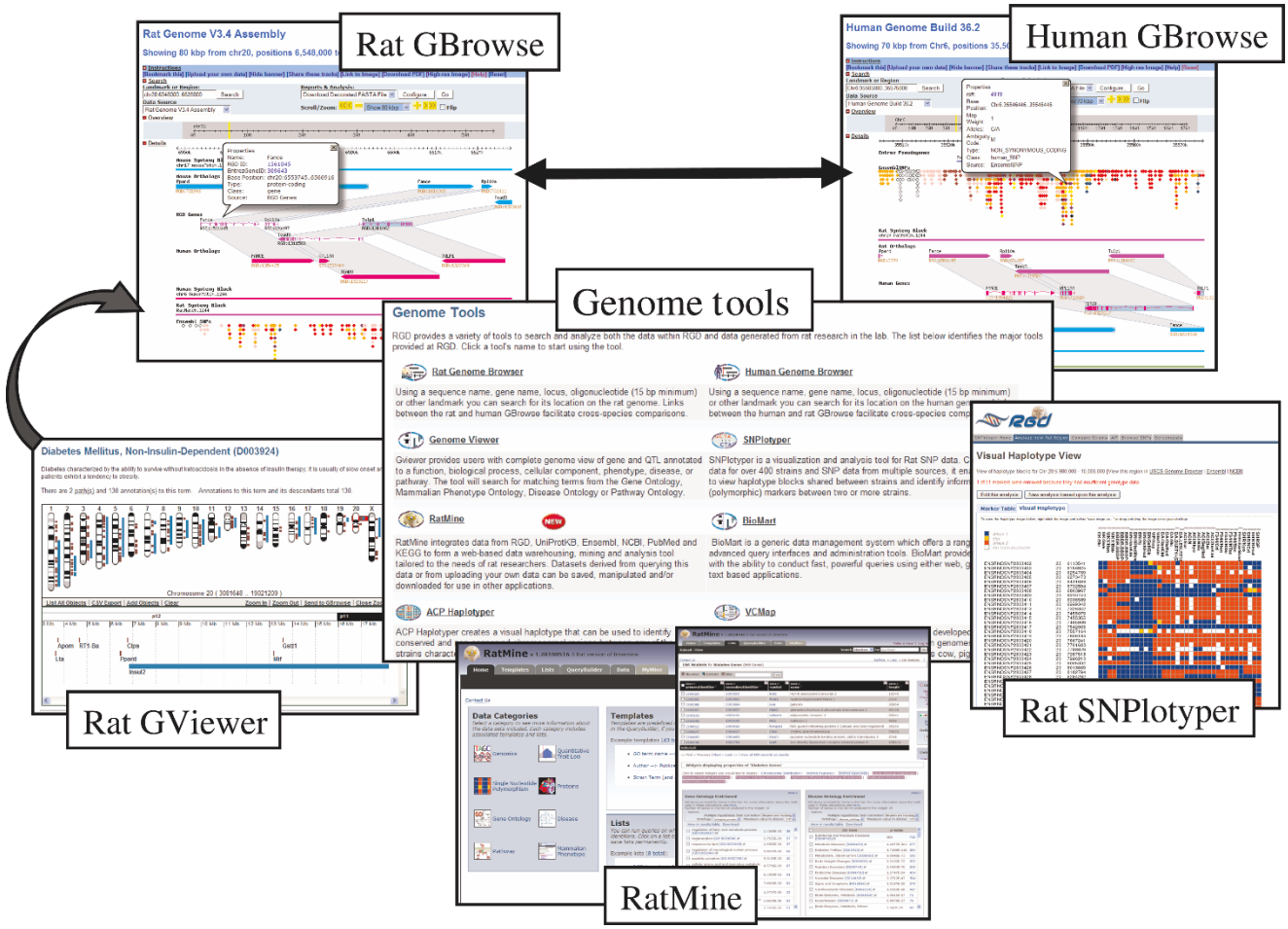

Figure 3. RGD's genome tools are designed to facilitate navigation and data analysis. From a genome-wide view in the GViewer, researchers can zoom in to a particular area, then view that area in the Rat GBrowse genome browser for a more complete view of the features located there. From there, navigate to the human GBrowse to determine how the features in the syntenic region of the human genome compare. Or a researcher might begin with a QTL in the human GBrowse and follow the link to the rat GBrowse to ascertain whether there are any known rat QTLs in the syntenic region. From there, researchers can move to the rat SNPlotyper to query which SNPs in that region are polymorphic between two or more rat strains. The resulting list of SNPs can be entered into the RatMine tool to view the genes associated with those SNPs and determine what functions, diseases, phenotypes or pathways may be overrepresented among those genes.

of synteny are shown as individual species-specific 'synteny blocks'. Orthologous genes are highlighted by shaded connections between genes in the three species. This makes evolutionary alterations such as differences in gene size and spacing, or areas of sequence inversion, easy to spot.

Navigating between the rat and human genome browsers is simplified by the inclusion of a link from the pop-up balloon for each human orthologue or human synteny block on the rat GBrowse to the same gene or synteny block in the human browser, and vice versa. Users interested in finding QTLs in corresponding regions in the two species, for instance, can use this feature to compare the human QTLs in the regions syntenic to a rat QTL of interest, or rat QTLs in the regions syntenic to a human QTL of interest, with ease. Synteny IDs such as 'ratMatch.1023' and 'humanMatch.1023' group the syntenic regions in all three species and are the same in both the rat and human browsers to aid this type of navigation.

To find polymorphic markers between two or more rat strains, RGD's SNPlotyper tool is ideal.

The tool includes SNPs from the STAR SNP consortium, dbSNP, funcSTAR, the CASCAD database and the Wellcome Trust Rat SNP resources. Users choose a region of the genome and strains from a list of 200 included in the tool. SNPlotyper returns both an exportable list of SNPs and their genotypes in each strain, and a haplotype diagram showing alternative alleles as different coloured blocks. This allows users to distinguish polymorphic SNPs and strain-specific haplotype blocks at a glance.

When looking for information on how these SNPs relate to genes or QTLs, researchers can use RatMine, a tool based on the InterMine technology developed by the FlyMine and ModENCODE teams at the Cambridge Systems Biology Centre.[3] RatMine integrates data for genes, QTLs, proteins, SNPs and strains, as well as ontology annotations for these data types. Researchers can build their own queries to extract data of interest, or utilise prebuilt template queries for answers to commonly used questions. Once a query is run, it returns a list of genomic elements that can be saved and used for analyses within RatMine or exported for use in other applications. Lists of genes or proteins, whether derived from a RatMine query or uploaded by the researcher, are viewed on the 'list analysis' page. This provides information about objects in the list, as well as a variety of 'widgets' which automatically perform enrichment analyses for ontology annotations such as disease and pathway, chromosomal locations and publications associated with that group of genes or proteins.

Summary

The RGD provides a unique comparative genomics platform for researchers interested in comprehensive rat, mouse and human data. Unique datasets include human QTL, signalling and regulatory pathways and detailed disease data catalogues. In addition, multiple genome tools allow the user to zoom across chromosomes, as well as zoom in to specific gene models, while providing syntenic views of data for easier comparisons. These make RGD an outstanding resource, not only for rat researchers, but also for those using mouse or conducting human studies.

References

1. Shimoyama, M., Hayman, G.T., Laulederkind, S.J., Nigam, R. *et al.* (2009), 'The rat genome database curators: Who, what, where, why', *PLoS Computat. Biol.* Vol. 11, p. e1000582.
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. *et al.* (2000), 'Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium', *Nat. Genet.* Vol. 25, pp. 25–29.
3. Lyne, R., Smith, R., Rutherford, K., Wakeling, M. *et al.* (2007), 'FlyMine: An integrated database for *Drosophila* and *Anopheles* genomics', *Genome Biol.* Vol. 8, p. R129.