Exploring human disease using the Rat Genome Database

Mary Shimoyama1,*, Stanley J. F. Laulederkind1, Jeff De Pons1, Rajni Nigam1, Jennifer R. Smith1, Marek Tutaj1, Victoria Petri1, G. Thomas Hayman1, Shur-Jen Wang1, Omid Ghiasvand1, Jyothi Thota1 and Melinda R. Dwinell2

ABSTRACT

*Rattus norvegicus*, the laboratory rat, has been a crucial model for studies of the environmental and genetic factors associated with human diseases for over 150 years. It is the primary model organism for toxicology and pharmacology studies, and has features that make it the model of choice in many complex-disease studies. Since 1999, the Rat Genome Database (RGD; http://rgd.mcw.edu) has been the premier resource for genomic, genetic, phenotype and strain data for the laboratory rat. The primary role of RGD is to curate rat data and validate orthologous relationships with human and mouse genes, and make these data available for incorporation into other major databases such as NCBI, Ensembl and UniProt. RGD also provides official nomenclature for rat genes, quantitative trait loci, strains and genetic markers, as well as unique identifiers. The RGD team adds enormous value to these basic data elements through functional and disease annotations, the analysis and visual presentation of pathways, and the integration of phenotype measurement data for strains used as disease models. Because much of the rat research community focuses on understanding human diseases, RGD provides a number of datasets and software tools that allow users to easily explore and make disease-related connections among these datasets. RGD also provides comprehensive human and mouse data for comparative purposes, illustrating the value of the rat in translational research. This article introduces RGD and its suite of tools and datasets to researchers – within and beyond the rat community – who are particularly interested in leveraging rat-based insights to understand human diseases.

KEY WORDS: Rat Genome Database, Genomics, Disease, Data analysis, Online resource

Introduction

Since 1850, *Rattus norvegicus* (the laboratory rat) has been the model organism of choice for many investigations into the physiological mechanisms of complex diseases and the genetic and environmental factors that affect disease onset, progression and severity (Lindsey, 1979; Aitman at al., 2008). The more than 1.5 million publications of research using rat models reflect its use in laboratories around the world. Since the completion of the rat genome sequence in 2004 (Gibbs et al., 2004), more than 40 inbred rat strains commonly used as disease models have been sequenced and genomic variations among these strains identified, providing valuable tools for linking genotypes to phenotypes (Hermsen et al., 2015). Continued advancements in genetic-modification technologies have led to the generation of more refined models, further contributing to the increasing popularity of the rat as a genetic model of disease (Flister et al., 2015); the resulting targeted models provide important resources for researchers. Because of the recognized value of existing and emerging rat datasets, the Rat Genome Database (RGD; http://rgd.mcw.edu) was created in 1999 and has evolved into the leading resource for rat genomic, genetic, phenotype and strain data. The main responsibility of RGD is to retrieve rat data and confirm orthologous relationships with human and mouse genes. This collected and validated information is imported for use at several large data resources, such as Ensembl, NCBI and UniProtKB. RGD also provides official nomenclature for rat genes, quantitative trait loci (QTLs), strains and markers, as well as unique identifiers. The RGD team is supplementing these annotations via software pipelines which, on a monthly basis, add over 50,000 new annotations to RGD, thereby increasing the total number of annotations in RGD to over 1.5 million.

Disease-related data acquisition

RGD provides complete gene, QTL and strain catalogues with comprehensive functional annotations for rat-, human- and mouse-derived data. Because many RGD users focus their studies on particular diseases, data are manually curated according to specific disease areas, providing an efficient means for curators to prioritize literature and integrate associated functional information. Genes, QTLs, strains and pathways related to the prioritized disease area are identified, the related literature is reviewed, and data are added to the database in the form of annotations to the appropriate ontologies (Table 1). Each annotation associates a data object such as a gene, QTL or strain with an ontology term and the reference that provides evidence for the association. RGD curators manually annotate disease and pathway information across species, gene ontologies for rat, and phenotype data for rat and human. These manually curated annotations are supplemented via software pipelines which, on a daily basis, provide new annotations to the database.
weekly basis, automatically import data from outside sources and associate those data with RGD genes as follows: Gene Ontology (GO) annotations for mouse and human genes are imported from the Gene Ontology Annotation (GOA) database (Huntley et al., 2015); Mammalian Phenotype Ontology (MP) annotations for mouse genes are imported from the Mouse Genome Database (MGD) (Bult et al., 2016), which is part of Mouse Genome Informatics (MGI), an international database resource for mouse research; human

Table 1. Numbers of functional annotations for rat, human and mouse genes

| Annotation category | Ontology used | Ontology reference | Species-specific | Inferred from other species | Total | Species-specific | Inferred from other species | Total |
|---------------------|--------------|--------------------|-----------------|-----------------------------|-------|-----------------|-----------------------------|-------|
| Disease             | RDO          | Hayman et al., 2016| 10,955          | 64,062                      | 75,017| 121,058         | 15,611                      | 136,669|
| Phenotype           | HPO          | Groza et al., 2015 | 1550            | 0                           | 1550  | 119,538         | 0                           | 119,538|
| Pathway             | MP           | Smith and Eppig, 2015| 12,292         | 17,987                      | 30,279| 28,274          | 1024                        | 29,298|
| Drug/chemical interactions | PW   | Hastings et al., 2016| 204,026        | 549,506                     | 753,532| 350,327        | 417,853                    | 768,180|
| Molecular function  | GO-MF        | Gene Ontology Consortium, 2015| 57,643         | 79,818                      | 137,461| 138,603        | 2750                        | 141,353|
| Biological process  | GO-BP        | Gene Ontology Consortium, 2015| 104,687        | 101,489                     | 206,176| 129,515        | 12,370                      | 141,885|
| Cellular component  | GO-CC        | Gene Ontology Consortium, 2015| 61,316         | 60,761                      | 122,077| 128,715        | 4389                        | 133,104|

RDO, modified MEDIC ontology; HPO, Human Phenotype Ontology – human only; MP, Mammalian Phenotype Ontology – rat and mouse; PW, Pathway Ontology; CHEBI, Chemical Entities of Biological Interest; GO, Gene Ontology; MF, Molecular Function; BP, Biological Process; CC, Cellular Component.

Fig. 1. The Cardiovascular Disease Portal home page. Selecting ‘Arrhythmias, Cardiac’ in the first disease category dropdown menu (1) results in a summary view of rat, human and mouse gene, QTL and rat strain objects annotated to the selected term (2). Below that is a Genome Viewer (GVViewer) display, showing the genomic positions of objects (genes, QTLs and strains) annotated to the term (3). These are presented in lists beneath the GVViewer, with links to report pages dedicated to individual genes (4). Accessed 15 April, 2016.
phenotype and disease annotations are imported through multiple pipelines: Online Mendelian Inheritance in Man (OMIM) (NCBI Resource Coordinators, 2016), ClinVar (Landrum et al., 2016) and the Genetic Association Database (GAD) (Becker et al., 2004), which has been retired, although data remain available; drug/chemical–gene interaction data are imported from the Comparative Toxicogenomics Database (CTD) (Davis et al., 2015); and data on molecular pathways are imported from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), the Small Molecule Pathway Database (SMPDB) (Jewison et al., 2014) and the Pathway Interaction Database (Schaefer et al., 2009) which, like GAD, is a retired legacy resource (Schaefer et al., 2009). As shown in Table 1, these data-acquisition efforts have resulted in enormous sets of annotations for diseases and disease-related phenotypes, as well as for functional categories such as pathways, biological processes and molecular functions, which can assist researchers in understanding how a gene or set of genes might be involved in the disease process. To provide easy access to these important datasets, RGD has created Disease Portals — entry points to consolidated disease-related data for researchers — and multiple software tools for data retrieval and analysis.

**Disease portals and data-mining tools**

**RGD Disease Portals**

Based on its manual-curation initiatives, RGD has generated Disease Portals, which provide easy access to multiple genomic and genetic data types associated with specific disease areas (Hayman et al., 2016). Currently, there are ten Disease Portals, covering: cardiovascular disease (CVD), cancer, diabetes, immune and inflammatory diseases, obesity and metabolic syndrome, and neurological, renal, respiratory, sensory organ and age-related diseases. Each portal integrates data for genes, QTLs and strains associated with the disease(s) highlighted by that portal. Each portal contains pages dedicated to different types of datasets, and these can be easily accessed via tabs at the top of each portal page. The different categories include ‘Diseases’, ‘Phenotypes’, ‘Biological Processes’ and ‘Pathways’. The dedicated pages each include a simple two-box search with dropdown menus. The first box contains major categories and, after a category is chosen, the second box presents the user with subcategories to choose from. For example, within the ‘Disease’ component of the cardiovascular Disease Portal, different types of CVD can be selected from the first dropdown menu, and further subcategories of the disease can be selected from the second dropdown menu (Fig. 1). The results shown in the main window include the number of genes, QTLs and strains associated with the selected disease term (Fig. 1, ‘2’), a visualization of them across the genome with a function to show human or mouse synteny (Fig. 1, ‘3’), and a listing of returned data elements annotated to the selected term for all three organisms with links to the respective report pages (Fig. 1, bottom). Report pages for genes, QTLs and strains provide summary views of annotations for diseases, phenotypes, drug/chemical–gene interactions, pathways and GO, with expanded views that include links to literature and other references from which the annotations were made (Fig. 2). Using the Disease Portals, researchers can get a fuller and better-rounded picture of their disease of interest, across three species.

**Ontology browser**

Another tool that allows easy access to multiple data types related to disease is the Ontology Browser, which can be accessed using the ‘Function’ button on the RGD homepage and with which a user can query across all the ontologies to find related data. Researchers can use a simple keyword search to retrieve the ontologies and respective terms associated with the search term (Fig. 3).
on any of the ontologies will display the terms retrieved plus an indication of the number of existing annotations to that term and/or its more specific child terms (subcategories). In the example in Fig. 3, a search is made for the term “arrhythmia.” Seven ontologies had terms that contained this search text (Fig. 3, ‘1’); from these, we selected Pathway Ontology to view the matching terms in that ontology (Fig. 3, ‘2’). The user can then retrieve the data objects annotated to a specific term by clicking on that term in the table or on the corresponding annotation number. In Fig. 3, the term “arrhythmogenic right ventricular cardiomyopathy pathway” was selected: the ontology and its associated annotations can be explored further by clicking on the branch icon or the link labelled ‘browse tree’ (Fig. 3, ‘3’); clicking on the term itself in the list (2) brings up the term ontology report page (4), which displays a Genome Viewer (GViewer) genome-wide view of objects annotated to the term. Below that is a list of rat, human and mouse genes annotated to the term, with links to the genome browser JBrowse to allow additional analysis. Accessed 15 April, 2016.

OLGA and the gene annotator
OLGA (Object List Generator and Analyzer) is a data-analysis tool that users can employ to assemble datasets for genes, QTLs or strains based on functional categories or genomic regions. Researchers can also upload gene lists and append or integrate these with gene lists created with the tool. Multiple lists of data objects can be created based on any of the categories. The lists can then be combined to create a non-redundant set, filtered to present just the common elements among the data sets, or subtracted, one list from another. Data mining for all three organisms (rat, mouse or human) is possible with OLGA. An example of dataset creation is presented in Fig. 4. The user initially chooses the data type (gene, QTL, strain) and then the method for creating the list [functional annotation (ontology), genomic position/region, QTL region, or symbol list upload]. In this example (Fig. 4, ‘1’), the user chooses to create a gene list based on the disease term ‘Arrhythmias, Cardiac’. The keyword search for each ontology offers an autocomplete feature and a list of potential matches so that the user can easily find the desired term. A list of genes annotated to ‘Arrhythmias, Cardiac’ or more specific categories of arrhythmias is assembled. At the time of access, the list contained 206 genes, but this number can vary as new annotations are added to the database. In step ‘2’, the user creates a second list of genes that interact with caffeine by browsing Chemical Entities of Biological Interest (CHEBI). This generated a list of 468 genes. The user then chooses how to integrate the two lists: through a union or intersection, or by subtracting the second list from the first (Fig. 4, ‘3’). In the example, an intersection of the two lists created a set of 18 genes that are associated with cardiac...
arrhythmias and also interact with caffeine. The user is then presented with the option to add another gene list or analyze the results (Fig. 5, '1'). ‘Analyze Result List’ presents the user with several options for displaying or analyzing the data aggregated to this point (Fig. 5, '2'). There is an option to download the gene set as a Microsoft Excel file, which includes the symbol, RGD ID, chromosome number and start/stop positions for each gene. The dataset can be displayed in the Genome Viewer, which shows the location of each of the genes beside the karyotype for that species, and also provides functions to add additional data objects or download the data. For rat genes, the gene list can also be uploaded into the Variant Visualizer tool in order to identify and visualize sequence variants within these genes across multiple strains. To obtain comprehensive data for each gene in the list and for further analysis of the dataset, the Gene Annotator (GA) tool is used (Fig. 5, '3'). Sending the gene set to the GA tool generates a full report for each gene, including human and mouse orthologue data. The report provides links to sequence data and corresponding gene reports at RGD, NCBI and Ensembl, as well as full functional annotations for GO, disease, phenotype, pathway and drug/chemical–gene interactions for all three organisms. Further investigation of the dataset can be conducted via the Annotation Distribution function, which shows the percentage of genes within the dataset associated with additional diseases, pathways, biological processes and other functional information. The Comparison Heat Map (Fig. 5, '4') function provides an easy method for identifying subsets of genes based on their annotations. The user chooses functional categories for the x and y axes from the multiple ontology types in the provided dropdown lists. In this example (Fig. 5, '5'), the user chose ‘Pathways’ and ‘GO biological processes’ for the axes. Browsing down the ontologies by selecting terms on the axes restricts the results to genes annotated to terms in these more specific branches. In this way, the user narrows the original dataset of 18 genes to a subset of six genes that, from OLGA analysis, are associated with cardiac arrhythmias and interact with caffeine, and via the GA tool are shown to be involved in a cellular metabolic process and participate in a cardiovascular system homeostasis pathway. Links to the comprehensive gene reports are provided for this subset. As demonstrated, OLGA gives users the ability to perform complex queries for disease-related data and easily funnel those results into analysis tools for additional investigation.

Summary
RGD provides a unique platform for accessing comprehensive disease-associated data for rat, mouse and human. Its Disease Portals centralize multiple data types related to specific disease areas within a single website component. Researchers can access gene, QTL, strain, pathway, phenotype and biological-process data that is of interest to them via a single resource. OLGA and the Gene Annotator are innovative, user-friendly software tools for creating
and analyzing sets of genes, QTLs or strains related to multiple aspects of disease. RGD continues its commitment to providing the best in data and software tools for rat researchers and for researchers and clinicians beyond the rat research community.

This article is part of a special subject collection ‘Spotlight on Rat: Translational Impact’, guest edited by Tim Aitman and Aron Geurts. See related articles in this collection at http://dmm.biologists.org/collection/rat-disease-model.

Competing interests
The authors declare no competing or financial interests.

Funding
The Rat Genome Database (RGD) is funded by grant HL64541 from the National Heart, Lung, and Blood Institute on behalf of the National Institutes of Health (NIH).

References
Aitman, T. J., Critser, J. K., Cuppen, E., Dominiczak, A., Fernandez-Suarez, X. M., Flint, J., Gauguier, D., Geurts, A. M., Gould, M., Harris, P. C. et al. (2008). Progress and prospects in rat genetics: a community view. Nat. Genet. 40, 516-522.

Becker, K. G., Barnes, K. C., Bright, T. J. and Wang, S. A. (2004). The genetic association database. Nat. Genet. 36, 431-432.

Bult, C. J., Eppig, J. T., Blake, J. A., Kadin, J. A. and Richardson, J. E. (2016). Mouse genome database 2016. Nucleic Acids Res. 44, D840-D847.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Scicky, D., King, B. L., Wiegars, T. C. and Matlingly, C. J. (2015). The comparative toxicogenomics database’s 10th year anniversary: update 2015. Nucleic Acids Res. 43, D914-D920.

Flister, M. J., Prokop, J. W., Lazar, J., Shimoyama, M., Dwwlin, M. and Geurts, A. (2015). 2015 guidelines for establishing genetically modified rat models for cardiovascular research. J. Cardiovasc. Transl. Res. 8, 269-277.

Gene Ontology Consortium. (2015). Gene ontology consortium: going forward. Nucleic Acids Res. 43, D1049-D1056.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergen, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E. et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493-521.

Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemotiel, T., Schriml, L. M., Klibbe, W. A., Schofield, P. N., Beck, T. et al. (2015). The human phenotype ontology: semantic unification of common and rare disease. Am. J. Hum. Genet. 97, 111-124.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. and Steinbeck, C. (2016). ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res. 44, D1214-D1219.

Hayman, G. T., Laulederkind, S. J., Smith, J. R., Wang, S. J., Petri, V., Nigam, R., Tutaj, M., De Pons, J., Dwinell, M. R. and Shimoyama, M. (2016). The Disease
Portals, disease-gene annotation and the RGD disease ontology at the rat genome database. Database 2016, baw034.

Hermsen, R., de Ligt, J., Spee, W., Blokzijl, F., Schäfer, S., Adami, E., Boymans, S., Flink, S., van Boxtel, R., van der Weide, R. H. et al. (2015). Genomic landscape of rat strain and substrain variation. BMC Genomics 16, 357.

Huntley, R. P., Sawford, T., Mutow-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J. and O’Donovan, C. (2015). The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. 43, D1057-D1063.

Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arnzt, D. et al. (2014). SMPDB 2.0: big improvements to the small molecule pathway database. Nucleic Acids Res. 42, D478-D484.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D37-D462.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J. et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44, D862-D868.

Lindsey, J. R. (1979). Historical foundations in the laboratory rat. In The Laboratory Rat (ed. J. J. Baker, J. R. Lindsey and S. H. Weisbroth), pp. 1-36. New York, NY: Academic press.

NCBI Resource Coordinators. (2016). Database resources of the national center for biotechnology information. Nucleic Acids Res. 44, D7-D19.

Petri, V., Jayaraman, P., Tutaj, M., Hayman, G. T., Smith, J. R., De Pons, J., Laulederkind, S. J. F., Lowry, T. F., Nigam, R., Wang, S.-J. et al. (2014). The pathway ontology - updates and applications. J. Biomed. Semantics 5, 7.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K. H. (2009). PID: the pathway interaction database. Nucleic Acids Res. 37, D674-D679.

Smith, C. L. and Eppig, J. T. (2015). Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. J. Biomed. Semantics 6, 11.