

What everybody should know about the rat genome and its online resources

Simon N Twigger, Kim D Pruitt, Xosé M Fernández-Suárez, Donna Karolchik, Kim C Worley, Donna R Maglott, Garth Brown, George Weinstock, Richard A Gibbs, Jim Kent, Ewan Birney & Howard J Jacob

It has been four years since the original publication of the draft sequence of the rat genome. Five groups are now working together to assemble, annotate and release an updated version of the rat genome. As the prevailing model for physiology, complex disease and pharmacological studies, there is an acute need for the rat's genomic resources to keep pace with the rat's prominence in the laboratory. In this commentary, we describe the current status of the rat genome sequence and the plans for its impending 'upgrade'. We then cover the key online resources providing access to the rat genome, including the new SNP views at Ensembl, the RefSeq and Genes databases at the US National Center for Biotechnology Information, Genome Browser at the University of California Santa Cruz and the disease portals for cardiovascular disease and obesity at the Rat Genome Database.

Rat genome sequencing and assembly

The draft genome sequence of the Norway rat (*Rattus norvegicus*, Brown Norway (BN) strain BN/NHsdMcwi) was generated and assembled in a public project supported by the US National Human Genome Research Institute (NHGRI) and National Heart, Lung and Blood Institute and led by the Baylor College of Medicine Human Genome Sequencing Group. This assembly (RGSC v3.4) used a variety of sequencing and mapping resources based on a BAC plus whole-genome shotgun (WGS) strategy. After its initial publication¹ in 2004, one

subsequent update was released in which portions of the draft were replaced with finished sequence from BACs². The sequence quality of the current assembly is good, although some difficult-to-assemble regions remain to be addressed.

More sequence data have been produced since the initial assembly, enhancing the sequence-based resources for the rat community as well as expanding the input data available for use in calculating the genome annotation. For example, ongoing sequencing projects by both individual labs and larger efforts have resulted in more than 4,500 new mRNA submissions to GenBank in the past three years (see **Table 1** for some general rat genome and sequence statistics compared to mouse and human). Other data include more than 6,000 cDNA sequences produced by the Mammalian Gene Collection and 3 million WGS reads for SNP discovery released by the Baylor College of Medicine Human Genome Sequencing Group. The scope of this extra sequence data extends to one-fold coverage of the Sprague-Dawley (SD) genome completed by Applied Biosystems and a complete WGS assembly released by Celera. More details of these projects have been published elsewhere².

Genome upgrade in 2008. The extra sequence data now available provide an oppor-

tunity to upgrade the rat assembly, and work is underway to incorporate these new data into a genome upgrade planned for early 2008. The upgrade will combine the unique sequence data from the WGS-only assembly with that from the combined WGS-plus-BAC assembly, resulting in a more complete representation of the genome. Also in development are plans to improve the coverage of the Y chromosome. The Y chromosome was originally sequenced only to about twofold coverage as a consequence of the unusually large size of the Brown Norway Y chromosome. This will be addressed by including additional Y chromosome sequence from a rat strain with a more moderately sized Y chromosome, combining WGS sequence and low-coverage sequence from male BAC clones.

SNP identification and haplotype map resources. Another crucial genomic resource for any model system is SNPs. A great deal of recent progress has occurred in this area that will soon translate into a rapid increase in the number of SNPs available for the rat. Until recently, there were fewer than 40,000 rat reference SNPs in the dbSNP database; however, a number of groups are working to greatly increase this number. The European STAR consortium, together with Japanese colleagues, has identified 2.9 million SNPs from genomic DNA of six strains (S Sprague-Dawley, SS/Jr,

Simon N. Twigger and Howard J. Jacob are at the Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, USA. Kim D. Pruitt, Donna R. Maglott and Garth Brown are at the National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. Xosé M. Fernández-Suárez and Ewan Birney are at the European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK. Donna Karolchik and Jim Kent are at the University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. Kim C. Worley, George Weinstock and Richard A. Gibbs are at the Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. e-mail: simont@hmgc.mcw.edu

Table 1 Current rat genome statistics compared to mouse and human^a

	Rat (RGSC v3.4)	Mouse (NCBI m37)	Human (NCBI 36)
Genome size	2.51 Gbp	3.42 Gbp	3.25 Gbp
Assembly status	Draft	Finished	Finished
Protein coding genes with RefSeqs	24,948	29,171	24,291
Homologenes	17,868	19,504	19,436
Complete mRNAs	80,510	110,870	97,822
ESTs	869,431	4,923,946	8,251,867
UniGene clusters	81,726	106,980	157,342

^aBased on queries performed at NCBI on 25 March 2008.

GK/Ox, WKY/Mdc, F344 and SHRSP/Mdc). Twenty thousand of these SNPs have been genotyped in more than 300 inbred and hybrid strains³. Eight strains (PVG, F344, SS, LEW, BB, FHH, DA and SHR) have been sequenced by the SNP discovery effort in the United States. The EURATools project intends to take these SNPs and genotype them across at least 150 inbred strains, enabling the development of a haplotype map. The previous rat SNP map was based on cDNA from four strains (SHRSP, Brown Norway, WKY and Sprague-Dawley) and had 12,395 interstrain polymorphic sites⁴. The databases and tools available for working with these SNP resources are described below.

Despite still not being assembled to 'finished' quality across the whole genome, the more complete and accurate genome sequence produced by the planned upgrade, combined with more SNP markers, will clearly be of great use to many researchers. Access to these resources is predominantly through the various genome databases that analyze, annotate and curate the assembled genome sequence. The following sections outline some of the resources available at these sites.

Rat annotation pipelines at Ensembl

Ensembl provides a comprehensive system combining open access to the underlying databases with genome annotation visualization using a user-friendly web browser⁵. As with the other genomes found in Ensembl, the rat genome is annotated using a well-established pipeline⁶ that begins with the latest assembly (now RGSC v3.4) and ultimately delivers the Ensembl gene models for the rat. This pipeline has been streamlined and optimized in recent years, taking advantage of new data types such as more accurate cDNA^{6,7} sequences. In the current release (v47), this has resulted in the annotation of more than 35,000 transcripts and almost 23,000 protein-coding genes. Of growing interest in recent years, 2,704 RNA genes (including microRNA, small nucleolar RNA, small nuclear RNA and ribosomal RNA) are also annotated through this system⁵. A further contribution of this pipeline is EST genes,

predicted splice variants based solely on EST evidence⁸. Ensembl also presents resequencing data from the Sprague-Dawley strain generated by Celera, enabling this to be compared to the Brown Norway reference genome.

SNPs and genome variation. Building on dbSNP, Ensembl includes approximately 2.9 million SNPs generated by the STAR project. New visualization tools such as TranscriptSNPView make it possible to compare this variation data among the different rat strains for which there is genotype data available⁹. Sequences from these strains can be visualized with SequenceAlignView, highlighting SNPs that exist between a particular strain and the reference assembly (**Supplementary Fig. 1** online). At present, data from the Sprague-Dawley, SS/Jr, GK/Ox, WKY/Mdc, F344 and SHRSP/Mdc strains are available, along with the reference strain BN/NHsdMwci.

Comparative genomics. Ensembl aligns whole genomes, identifying constrained elements (putative functional elements) and showing them in ContigView. These multiple sequence alignments are produced using PECAN, a progressive alignment algorithm. Based on these data, AlignSpliceView can display rat alongside the syntenic mouse or human regions, showing homologous genes in their genomic context, as well as conserved regions between the species. Homology relationships at the gene level are calculated using a phylogenetic approach and can be visualized in the GeneTreeView (**Supplementary Fig. 2** online). The one-to-one orthologies and other orthology categories (such as one-to-many and many-to-many) are available in the GeneView pages (**Supplementary Fig. 3** online).

Ensembl as a data integration platform. Although Ensembl contains a great deal of information, there is also a wealth of data available outside the Ensembl system. To address this issue, Ensembl makes extensive use of the Distributed Annotation System (DAS) to integrate external resources into the Ensembl browser. This enables any DAS-compatible sources to be displayed alongside the primary Ensembl annotation. Microarray data can be

integrated in this fashion, and GeneView allows users to visualize and access ArrayExpress data associated with a particular gene. Phenotype and expression data are of paramount importance in EURATools. Ensembl's contribution to this consortium involves integrating additional cDNA data obtained from the resequencing effort and presenting expression quantitative trait locus (QTL) data, using such tools as eQTL Explorer¹⁰, which relies on the DAS protocol.

Data mining with BioMart. BioMart¹¹ provides a data-mining tool to interact with the Ensembl database, enabling users to go beyond the genome browsers to retrieve information from Ensembl. For rat, this contains the current Ensembl genome data sets, a SNP data set and EURAMart, which is a compendium of gene expression data populated with the Gene Expression Atlas from the Genomic Institute of the Novartis Research Foundation¹². BioMart has been deployed by several other databases (for example, Rat Genome Database, or RGD¹³) and can be used to integrate data from these different installations. EURAMart can act as a bridge between Ensembl and RGD data, allowing the integration of expression data (from EURAMart) with phenotype and disease information (from RGD).

Rat resources at the NCBI

The US National Center for Biotechnology Information (NCBI) maintains several resources that support the rat research community, including the integrated suite of literature, sequence and BLAST databases; tools to query, retrieve and display biological information contained in those databases; reference sequence (RefSeq)¹⁴ and Gene¹⁵ records; variation data; two WGS assemblies annotated by NCBI; and radiation-hybrid and genetic maps. The Rat Genome Resources web page provides an up-to-date portal to access these and other rat-specific data (**Supplementary Table 1** online). The following highlights a subset of resources of interest to rat researchers. More information on NCBI resources is included in **Supplementary Table 1** and in the online NCBI Handbook and NCBI Help book.

Gene. Gene is the central resource for rat gene-specific information at NCBI and includes protein-coding and non-protein-coding genes, pseudogenes and mapped phenotypes. The database reports assigned gene ontology terms, cytogenetic locations, names and symbols, pathways, protein interactions, publications (including the GeneRIF annotated bibliography), sequences (GenBank and RefSeq) and links to numerous NCBI and other resources. Data are maintained through a combination of computation, collaboration

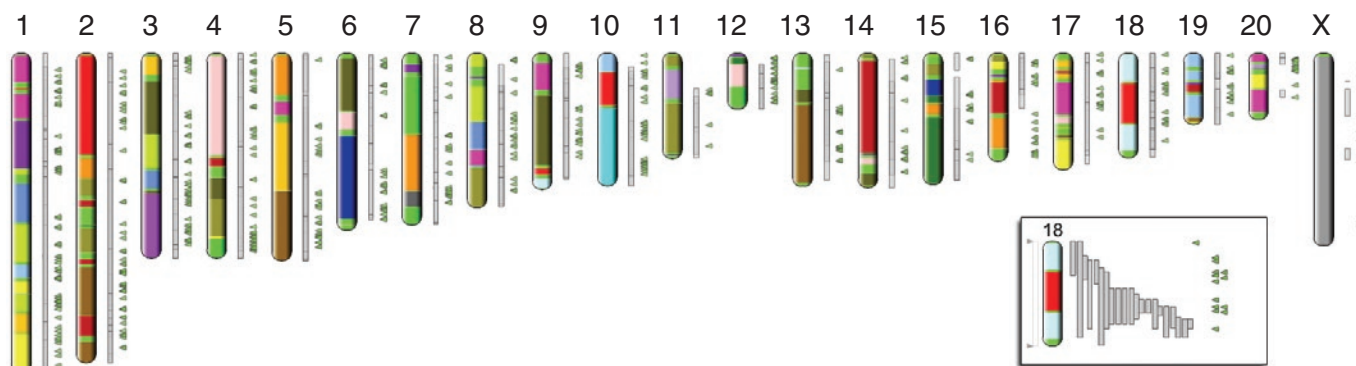


Figure 1 GViewer visualization of rat genes and QTLs associated with cardiovascular disease. The genes (green triangles) and QTLs (gray bars) are shown aligned with the rat genome, which has been colored according to the corresponding syntenic regions in the human genome. An expanded view for each chromosome is available that shows each individual gene and QTL location (inset). Each individual gene and QTL is hyperlinked back to the original database report page. Five hundred eighty-six rat genes, 388 rat QTLs and 127 rat strains have been associated with various types of cardiovascular disease. These are all available through the RGD Cardiovascular Disease Portal, along with 588 human genes, 559 mouse genes and 49 human QTLs.

and ongoing curation by the Gene and RefSeq staff. Collaboration and curation enhance the content of Gene by (i) integrating information from, and establishing links to, databases such as RGD, RATMAP, Ensembl and UniProt, (ii) resolving identified data conflicts and ambiguities, and (iii) adding content such as sequences, names, phenotypes and publications. For example, regular updates synchronize rat gene nomenclature in the Gene database with that provided by RGD, add information based on new sequence submissions to GenBank, update gene ontology terms obtained by FTP from the Gene Ontology Consortium and, in collaboration with UniProtKB, update cross-links between Reference Sequence (RefSeq) proteins and the corresponding Swiss-Prot or TrEMBL proteins.

RefSeq. RefSeq data for rat include the genomic reference (RGSC v3.4) and Celera assemblies as well as gene-specific products. Accessions are assigned to chromosomes (**Supplementary Table 2** online), scaffolds and contigs. Gene-specific RefSeqs for RNAs and proteins include curated records based on submissions to GenBank and predicted records that are generated as a product of computing annotation for the genome assemblies.

Curation of RefSeq transcripts and proteins for rat is a continuous process and serves to (i) ensure accurate, full-length sequence for the complete set of transcripts and proteins, including loci that use selenocysteine or non-AUG codons, and (ii) provide additional RefSeq feature annotation such as mature peptides. In addition, the transcript-based curated RefSeq collection represents a high-quality complement to genome annotation because it can be used to identify genes that are not well represented in one or both genomic assemblies. For example, genes missing from the RGSC v3.4

reference assembly include the smooth muscle alpha-actin (*Acta2*, GeneID 81633), thymidine kinase 1 (*Tk1*, GeneID 24834) and gamma-glutamyltransferase 1 (*Ggt1*, GeneID 116568).

Map Viewer (see **Supplementary Fig. 4** online). NCBI provides a valuable service by annotating both available genome assemblies and displaying order of objects in two coordinate systems (physical distance (bp) and genetic distance (cM)). Genome annotation is computed based on alignments of the curated RefSeq collection described above, rat transcript data, and human, mouse and rat protein data. The results are distributed in the genomic RefSeq collection and in the RefSeq and Map Viewer FTP sites, and are available for browsing and querying by accession, text or sequence similarity (through BLAST) in the Map Viewer. Sequence-based data presented in the Map Viewer include the annotated genome at the gene and transcript level, plus an array of additional sequence details, including repeats, STS markers, CpG islands, alignments of rat genomic records and of human, mouse and rat transcript sequences, mapped phenotypes (QTLs) based on placement of flanking and peak markers and variation data from dbSNP. Alternative displays provide tabular reports and download support (Data as Table View), present alignments supporting the annotation (Evidence Viewer) or support using transcript alignments to generate alternative transcript models for further evaluation (Model Maker). Map Viewer also supports comparative displays of human, mouse and rat annotations, as well as review of order and orientation of assemblies based on placement of markers common to the sequence, genetic and radiation-hybrid maps.

BLAST. A rat-specific BLAST page facilitates access to several custom BLAST data-

bases. Options include the genome assemblies, RefSeq and GenBank RNAs and proteins, and trace reads from the Trace Archive. Query results for transcript and protein databases return links to the Gene, UniGene and/or GEO databases when an accession is known to that database. Query results for the genome assembly databases include links to view the results of Map Viewer in the context of the genome annotation.

dbSNP. dbSNP takes submissions of several classes of variation (for example, insertion/deletions, small tandem repeats, substitutions) and assigns them unique stable identifiers (ss numbers). Submissions are clustered periodically by alignment to the genome, and submissions of the same variant are assigned an rs number. The placement of these variants on the genome, and calculation of the effect of a variant on an encoded protein, is reported in Map Viewer and the dbSNP GeneView display.

Rat resources on the UCSC Genome Browser Website

The University of California, Santa Cruz (UCSC) Genome Bioinformatics Group provides a large collection of annotation data for the rat genome, along with a variety of web-based resources for displaying, querying, analyzing and downloading the assembly sequence and annotations. The complete toolset, downloadable data, documentation and related information are available through links on the Genome Browser website at <http://genome.ucsc.edu/> (**Supplementary Table 3** online).

Rat sequence and annotation data. The three most recent rat assemblies from the Baylor Human Genome Sequencing Center are featured on the main Genome Browser

website; an older release is archived at <http://genome-rn1.cse.ucsc.edu/cgi-bin/hgGateway>. New assemblies are added as they become available. **Supplementary Table 4** online provides a list of the assemblies supported at present on the UCSC website.

A broad set of annotations generated by UCSC and its collaborators is available for each rat assembly (**Supplementary Table 5** online). The annotation data, which is stored in a MySQL database shared by all the tools on the website, is roughly grouped into seven different categories based on shared characteristics: mapping and sequencing data, phenotype and disease association data, genes and gene predictions, mRNA and EST data, expression and regulation data, comparative genomics data, and variations and repeats.

Of particular note for the rat annotation, UCSC provides its own set of predicted protein-coding genes, UCSC Known Genes, based on protein data from UniProt and mRNA data from RefSeq and GenBank. This gene set will be further refined in the next rat assembly (rn5) to pull in further lines of evidence and include putative non-coding transcripts. Also of note is UCSC's extensive collection of rat comparative genomics data, which includes a measure of evolutionary conservation across several multiply-aligned vertebrates (Conservation), predictions of conserved elements within this group (Most Conserved) and pairwise alignments of several species to the rat genome showing both the alignment 'chains' and the best chain for every part of the rat genome ('nets')¹⁶.

Browsing, analysis and data-mining tools. The Genome Browser¹⁷ is a fast, interactive, web-based tool that displays a chromosome-oriented view of the annotation data aligned to the rat genome sequence as horizontal 'tracks' (**Supplementary Fig. 5** online). The rat genome can be queried using a wide range of search parameters, including chromosomal coordinate ranges, mRNA or EST accessions, gene names, clone accession numbers and keywords from mRNA GenBank descriptions. Navigation and configuration controls allow the user to adjust and customize the browsing window to focus on information of interest. Detailed description pages for the displayed data elements link to information from a wide range of external resources.

The Table Browser provides a graphical interface for downloading and manipulating the data underlying the rat Genome Browser annotations (**Supplementary Fig. 6** online). The user can view data tables, filter the output based on one or more criteria, intersect or correlate data from more than one table,

and restrict the output to specific coordinate ranges or lists of data elements. For more complex queries, Table Browser output may be exported to Penn State's Galaxy tool for further processing, or users can directly access the rat database through UCSC's public MySQL server.

UCSC's Custom Tracks functionality offers a convenient way for users to display and compare their own data with the built-in rat annotation. User-generated custom tracks may be viewed in the Genome Browser and manipulated using the full functionality of the Table Browser. They also provide an ideal medium for presenting a view of data submitted for publication or for sharing data with collaborators.

The Blat tool offers a fast method for quickly mapping rat sequence to the genome (**Supplementary Fig. 7** online). With DNA searches, Blat quickly finds matches of 95% and greater similarity of length 25 bases or more, finding perfect sequence matches of 33 bases and sometimes as few as 20 bases. On proteins, Blat finds sequences of 80% and greater similarity of length 20 amino acids or more.

The Gene Sorter provides a graphical interface for exploring expression, homology and other relationships among rat genes, as well as orthology with several model organisms such as human, mouse, zebrafish, fruitfly, worm and yeast (**Supplementary Fig. 8** online).

The Genome Graphs tool may be used to upload and view genome-wide datasets such as QTL mapping and linkage studies.

The Proteome Browser allows the user to examine many characteristics (such as exon structure, domains and structural features) of proteins associated with the gene (see **Supplementary Fig. 9** online).

Getting more information. Most of the programs and utilities on the UCSC website are documented through online user's guides. The website also provides a frequently asked questions page, links to several online tutorials, and subscription information and archives for three publicly accessible technical support mailing lists. See **Supplementary Table 6** online for a list of online resources.

Rat annotation resources at RGD

The Rat Genome Database (RGD, <http://rgd.mcw.edu>) is the model organism database for the laboratory rat¹². Its goal is to provide data and tools that build upon the rat genome and related genomic resources to maximize the utility of the rat as a model organism.

The core of RGD is information curated from the published literature combined with data imported from other authoritative

sources such as NCBI¹⁵, Ensembl⁵, UCSC¹⁸, Mouse Genome Informatics¹⁹ and Uniprot²⁰. RGD manually curates data for a variety of 'biological objects' and makes these available through the RGD web site and through downloadable files on the RGD FTP site. The website includes pages describing each rat gene and its function; an extensive catalog of rat strains used as experimental models, with their phenotypes and known diseases; and all rat QTLs that have been identified, mapping phenotypes to specific regions of the rat genome. A complete list of all RGD data objects and links to example reports are available in **Supplementary Table 7** online.

Bio-ontologies. The rat is very much a 'functional model': it is widely used to study genes involved in specific systems-level phenotypes or as a model of a particular disease. As a result, many researchers are looking for information to tie genes back to phenotypes or diseases of interest. In the present era, bio-ontologies²¹ provide the framework for describing this information in RGD and many other databases. As part of its ongoing curation, RGD provides the Gene Ontology²² annotation for the rat, describing the molecular function of a gene, the cellular component(s) in which it has been found and biological processes it is involved in. These are augmented with pathway, mammalian phenotype²³ and disease annotations. QTLs and strain records are also annotated with phenotype and disease ontology terms.

These annotations capture a wealth of functional knowledge within the database in a consistent fashion. They also tie the systems-level biology of disease, phenotype and pathway to the rat genome through the annotated QTLs and genes. Comparative genomics and orthology can then enable this information to be applied to the genomes of other species such as human and mouse.

RGD disease portals. Relating directly to the rat as a model of human disease, RGD has been creating specialized web sites (or 'portals') which provide an overview of disease-related genes, QTL and strains from rat, mouse and human in a single web page¹¹. The information on these sites is developed through targeted literature curation and ontology annotation and is focused on rat but also includes mouse data provided by Mouse Genome Informatics and human QTL data curated by RGD. The focus is on diseases that undergo significant research in the rat and that are of high clinical relevance. The portals include, at present, Cardiovascular (**Fig. 1**), Neurological and Obesity/Metabolic Syndrome. The Cancer/Neoplasms portal is in development, with release scheduled for

mid-2008. Links for these three portals are provided in **Supplementary Table 8** online; screenshots of the Cardiovascular portal are shown in **Supplementary Figure 10** online.

Bioinformatic tools for rat genomics. In addition to the curation and integration of rat genomic data, RGD provides a variety of tools to navigate, analyze and visualize this data. GViewer provides a broad, ontology-based search engine that can be used to find rat, mouse and human genes and other objects related to such things as pathways, diseases and phenotypes within the database. It provides a graphical view of the search results showing the distribution of these objects across the rat genome. RGD maintains a BioMart data warehouse in collaboration with the MCW Proteomics Center that can be used to create *ad hoc* datasets. BioMart can be linked to other BioMarts (such as EURAMart, described above) to enable complex queries between databases. SNPlopyer is a new tool supporting analysis and visualization of the emerging rat SNP data sets, particularly the most recent release from the STAR consortium. Links to each of these tools and others are provided in **Supplementary Table 9** online.

Rat strains. There are large numbers of different rat strains in use by the research community. Each has unique characteristics in terms of genotype and phenotype and, as such, embodies a unique model system. RGD curates strain data from the literature and at present has more than 1,400 records, including inbred, outbred, congenic and consomic strains, as well as heterogeneous stock and recombinant inbred lines. Many have been annotated according to their observed phenotypes or applicability to disease studies. By using this information, it is possible to iden-

tify strains that may provide a model system for subsequent research projects.

Concluding remarks

As evidenced by the ongoing enhancements to the rat genome sequence and the comprehensive bioinformatics resources described above, the rat and its genome have a broad base of support. This is being strengthened by closer coordination between the genome assembly and annotation groups, beginning with the impending genome upgrade. Plans are in development for the creation of a consensus gene set for the rat. This will be based on comparisons of results from the Ensembl, UCSC and RefSeq pipelines, in collaboration with RGD, to manage nomenclature and functional annotations.

There are many exciting developments coming to fruition in rat genomics through the efforts of groups worldwide. An overview of these developments and a vision for the future of the rat are covered in more detail in the accompanying Perspective²⁴. However, it is indisputable that modern biology rests heavily on the support of bioinformatics and public databases, and in this and many other areas, the portents look very promising for the future of rat as a tractable genetic model for physiology and complex disease studies.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank T. Aitman, N. Hübner and A. Kwitek for helpful comments during the preparation of this manuscript. This work was supported in part by the Intramural Research Program of the US National Institutes of Health, National Library of Medicine. RGD is supported in part by US National Institutes of Health grants HL-64541 and HG-002273. EURATools and the STAR consortium are supported through

the Sixth Framework Programme of the European Union, action line LSH-2003-1.1.0-1. The UCSC Genome Browser project is funded by grants from the NHGRI, the Howard Hughes Medical Institute and the US National Cancer Institute. The Phase 2 genome project and SNP discovery at Baylor College of Medicine Human Genome Sequencing Center is funded by NHGRI HG-003273.

- Gibbs, R.A. *et al.* *Nature* **428**, 493–521 (2004).
- Worley, K.C., Weinstock, G.M. & Gibbs, R.A. *Physiol. Genomics* **32**, 273–282 (2007).
- STAR Consortium. *Nat. Genet.* **40**, 560–566 (2008).
- Zimdahl, H. *et al.* *Science* **303**, 807 (2004).
- Hubbard, T.J. *et al.* *Nucleic Acids Res.* **35**, D610–D617 (2007).
- Curwen, V. *et al.* *Genome Res.* **14**, 942–950 (2004).
- Fernandez-Suarez, X.M., Searle, S. & Birney, E. in *In Silico Genomics and Proteomics: Functional Annotation of Genomes and Proteins* (eds. Mulder, N. & Apweiler, R.) 109–113 (Nova Science, New York, 2006).
- Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. *Genome Res.* **14**, 976–987 (2004).
- Cunningham, F. *et al.* *Nat. Genet.* **38**, 853 (2006).
- Mueller, M. *et al.* *Bioinformatics* **22**, 509–511 (2006).
- Kasprzyk, A. *et al.* *Genome Res.* **14**, 160–169 (2004).
- Walker, J.R. *et al.* *Genome Res.* **14**, 742–749 (2004).
- Twigger, S.N., Shimoyama, M., Bromberg, S., Kwitek, A.E. & Jacob, H.J. *Nucleic Acids Res.* **35**, D658–D662 (2007).
- Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **35**, D61–D65 (2007).
- Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. *Nucleic Acids Res.* **35**, D26–D31 (2007).
- Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489 (2003).
- Kent, W.J. *et al.* *Genome Res.* **12**, 996–1006 (2002).
- Karolchik, D. *et al.* *Nucleic Acids Res.* **36**, D773–D779 (2008).
- Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. & Blake, J.A. *Nucleic Acids Res.* **36**, D724–D728 (2008).
- The Uniprot Consortium. *Nucleic Acids Res.* **35**, D193–D197 (2007).
- Bodenreider, O. & Stevens, R. *Brief. Bioinform.* **7**, 256–274 (2006).
- Ashburner, M. *et al.* *Nat. Genet.* **25**, 25–29 (2000).
- Smith, C.L., Goldsmith, C.A. & Eppig, J.T. *Genome Biol.* **6**, R7 (2005).
- Aitman, T. *et al.* *Nat. Genet.* **40**, 516–522 (2008).