

# The 2013 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection

Xosé M. Fernández-Suárez<sup>1,\*</sup> and Michael Y. Galperin<sup>2,\*</sup>

<sup>1</sup>Cambridge, CB24 6DZ, UK and <sup>2</sup>National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received November 14, 2012; Accepted November 15, 2012

## ABSTRACT

The 20th annual Database Issue of *Nucleic Acids Research* includes 176 articles, half of which describe new online molecular biology databases and the other half provide updates on the databases previously featured in *NAR* and other journals. This year's highlights include two databases of DNA repeat elements; several databases of transcriptional factors and transcriptional factor-binding sites; databases on various aspects of protein structure and protein–protein interactions; databases for metagenomic and rRNA sequence analysis; and four databases specifically dedicated to *Escherichia coli*. The increased emphasis on using the genome data to improve human health is reflected in the development of the databases of genomic structural variation (NCBI's dbVar and EBI's DGVa), the NIH Genetic Testing Registry and several other databases centered on the genetic basis of human disease, potential drugs, their targets and the mechanisms of protein–ligand binding. Two new databases present genomic and RNAseq data for monkeys, providing wealth of data on our closest relatives for comparative genomics purposes. The *NAR* online Molecular Biology Database Collection, available at <http://www.oxfordjournals.org/nar/database/a/>, has been updated and currently lists 1512 online databases. The full content of the Database Issue is freely available online on the *Nucleic Acids Research* website (<http://nar.oxfordjournals.org/>).

## NEW AND UPDATED DATABASES

This 1300-page virtual volume represents the 20th annual Database Issue of *Nucleic Acids Research* (*NAR*). It includes descriptions of 88 new online databases, 77 update articles on databases that have been previously featured in the *NAR* Database Issue (Table 1) and 11 articles with updates on database resources whose descriptions have been previously published in other journals (Table 2).

At this point it might be instructive to look back at the origin and evolution of the *NAR* Database Issue. Its history started from two supplementary issues that were published in *NAR* in April of 1991 and in May of 1992 and consisted of 18 and 19 articles, respectively (see <http://nar.oxfordjournals.org/content/19/supplement.toc> and <http://nar.oxfordjournals.org/content/20/supplement.toc>). These articles offered descriptions of several nucleotide sequence databases, such as GenBank, the EMBL Data Library, compilations of small RNA, tRNA, and 5S, 16S, and 23S rRNA sequences (including the Ribosomal Database Project), DNA sequences from *Escherichia coli* and a human genome database (GDB). Those first issues also included descriptions of several protein databases, such as SWISS-PROT, PIR, PROSITE, Restriction Enzyme Database (REBASE), Transcription Factors Database (TFD) and Histone database. There was also a medical genetics database, Haemophilia B, listing point mutations and indels in the coagulation factor IX (*F9*) gene that caused this blood clotting disorder, which has affected the royal families of several European countries.

The next issue, published on July 1, 1993, was the first one formally labelled as the Database Issue. It consisted of 24 articles, which added databases of RNA and protein structure and the ENZYME database. It was followed by

\*To whom correspondence should be addressed. Tel: +1 301 435 5910; Fax: +1 301 435 7793; Email: [nardatabase@gmail.com](mailto:nardatabase@gmail.com) or [galperin@ncbi.nlm.nih.gov](mailto:galperin@ncbi.nlm.nih.gov)

Correspondence may also be addressed to Xosé M. Fernández-Suárez. Email: [xose.m.fernandez@gmail.com](mailto:xose.m.fernandez@gmail.com)

**Table 1.** New online databases featured in the 2013 NAR Database issue

Database name	URL	Brief description
APPRIS	<a href="http://appris.bioinfo.cnio.es/">http://appris.bioinfo.cnio.es/</a>	A system for annotating alternative splice isoforms
BioLiP	<a href="http://zhanglab.cmb.med.umich.edu/BioLiP/">http://zhanglab.cmb.med.umich.edu/BioLiP/</a>	Biologically relevant ligand-protein interactions
BSRD	<a href="http://kwanlab.bio.cuhk.edu.hk/BSRD">http://kwanlab.bio.cuhk.edu.hk/BSRD</a>	A repository of bacterial small regulatory RNA
CellLineNavigator	<a href="http://www.medicalgenomics.org/celllinenavigator">http://www.medicalgenomics.org/celllinenavigator</a>	Cell line expression profiles by microarray analysis
ChIPBase	<a href="http://deepbase.sysu.edu.cn/chipbase/">http://deepbase.sysu.edu.cn/chipbase/</a>	Transcriptional regulation of lncRNA and microRNA genes from ChIP-Seq data
ChiTaRS	<a href="http://chimerasrch.bioinfo.cnio.es/">http://chimerasrch.bioinfo.cnio.es/</a>	Chimeric RNAs of two or more different transcripts
CIL-CCDB	<a href="http://www.cellimagelibrary.org/">http://www.cellimagelibrary.org/</a>	Images, videos and animations of various cell types from diverse organisms
CircaDB	<a href="http://bioinf.itmat.upenn.edu/circa/">http://bioinf.itmat.upenn.edu/circa/</a>	Circadian gene expression profiles in human and mouse
CloneDB	<a href="http://www.ncbi.nlm.nih.gov/clone/">http://www.ncbi.nlm.nih.gov/clone/</a>	Clones and libraries: sequence data, map positions and distributor information
ClusterMine360	<a href="http://www.sigma54.ca/microbialclusters/">http://www.sigma54.ca/microbialclusters/</a>	Microbial PKS/NRPS Biosynthesis
Cyanolyase	<a href="http://cyanolyase.genouest.org/">http://cyanolyase.genouest.org/</a>	Sequences and motifs of the phycobilin lyase protein family
D2P2	<a href="http://d2p2.pro/">http://d2p2.pro/</a>	Database of Disordered Protein Predictions
dbVar	<a href="http://www.ncbi.nlm.nih.gov/dbvar">http://www.ncbi.nlm.nih.gov/dbvar</a>	Structural variation in chromosomes: inversions, translocations, insertions and deletions
DGVa	<a href="http://www.ebi.ac.uk/dgva/">http://www.ebi.ac.uk/dgva/</a>	domain-centric Gene Ontology
dcGO	<a href="http://supfam.org/SUPERFAMILY/dcGO">http://supfam.org/SUPERFAMILY/dcGO</a>	Human DNA repeat families
Dfam	<a href="http://dfam.janelia.org">http://dfam.janelia.org</a>	Disease and Gene Annotations database
DGA	<a href="http://dga.nubic.northwestern.edu/">http://dga.nubic.northwestern.edu/</a>	microRNA targets on long noncoding RNAs
DIANA-LncBase	<a href="http://www.microrna.gr/LncBase">http://www.microrna.gr/LncBase</a>	Database Of Biosynthesis clusters CURated and INtegrated
DoBISCUIT	<a href="http://www.bio.nite.go.jp/pks/">http://www.bio.nite.go.jp/pks/</a>	Various kinds of information about enzymes: small-molecule chemistry, biochemical pathways and drug compounds
EBI Enzyme Portal	<a href="http://www.ebi.ac.uk/enzymeportal">http://www.ebi.ac.uk/enzymeportal</a>	<i>Escherichia coli</i> Metabolome Database
ECMDB	<a href="http://www.ecmdb.ca/">http://www.ecmdb.ca/</a>	Engineered endonucleases: zinc finger nucleases and transcription activator-like effector nucleases
EENdb	<a href="http://eendb.zfgenetics.org/">http://eendb.zfgenetics.org/</a>	Energy profiles of protein structures
eProS	<a href="http://bioservices.hs-mittweida.de/EproS/">http://bioservices.hs-mittweida.de/EproS/</a>	Human transcription factor-binding data from ChIP-seq
Factorbook	<a href="http://www.factorbook.org/">http://www.factorbook.org/</a>	G-quadruplex Ligands Database
G4LDB	<a href="http://www.g4ldb.org/">http://www.g4ldb.org/</a>	Genomics of Drug Sensitivity in Cancer: Sensitivity for anti-cancer drugs in various cell lines
GDSC	<a href="http://www.cancerRxgene.org/">http://www.cancerRxgene.org/</a>	Genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences
GeneTack	<a href="http://topaz.gatech.edu/GeneTack/db.html">http://topaz.gatech.edu/GeneTack/db.html</a>	Domain structure predictions and 3D models for proteins from model genomes
Genome3D	<a href="http://genome3d.eu/">http://genome3d.eu/</a>	Database of glycan 3D structures
Glycan Fragment DB	<a href="http://www.glycanstructure.org/fragment-db">http://www.glycanstructure.org/fragment-db</a>	Heritability data with trait-associated genomic loci
H2DB	<a href="http://tga.nig.ac.jp/h2db/">http://tga.nig.ac.jp/h2db/</a>	A knowledge database for the Hepatitis B Virus
HBVdb	<a href="http://hbvdb.ibcp.fr/HBVdb/">http://hbvdb.ibcp.fr/HBVdb/</a>	Gene expression profiles in haematopoiesis
HemaExplorer	<a href="http://servers.binf.ku.dk/shs/">http://servers.binf.ku.dk/shs/</a>	Human Exone Splicing Events
HEXEvent	<a href="http://hertellab.mmg.uci.edu/cgi-bin/HEXEvent/HEXEventWEB.cgi">http://hertellab.mmg.uci.edu/cgi-bin/HEXEvent/HEXEventWEB.cgi</a>	HOmo sapiens COmprehensive MOdel COllection of hand-curated transcription factor-binding site models
HOCOMOCO	<a href="http://autosome.ru/HOCOMOCO">http://autosome.ru/HOCOMOCO</a> , <a href="http://cbrckauust.edu.sa/hocomoco/">http://cbrckauust.edu.sa/hocomoco/</a>	Kinase-Inhibitor-Disease Family Map
KIDFamMap	<a href="http://gemdock.life.nctu.edu.tw/KIDFamMap/">http://gemdock.life.nctu.edu.tw/KIDFamMap/</a>	Library of Apicomplexan Metabolic Pathways
LAMP	<a href="http://www.llamp.net/">http://www.llamp.net/</a>	Human lncRNA gene sequences and structures
Lncipedia	<a href="http://www.lncipedia.org/">http://www.lncipedia.org/</a>	Long non-coding RNA-associated diseases
LncRNADisease	<a href="http://cmbi.bjmu.edu.cn/lncrnadisease">http://cmbi.bjmu.edu.cn/lncrnadisease</a>	Predicted genome of Last Universal Common Ancestor
LUCApedia	<a href="http://eeb.princeton.edu/lucapedia/">http://eeb.princeton.edu/lucapedia/</a>	Comparative genomic tools for metagenome analysis
meta.MicrobesOnline	<a href="http://meta.MicrobesOnline.org/">http://meta.MicrobesOnline.org/</a>	Metabolomics experiments and associated metadata
MetaboLights	<a href="http://www.ebi.ac.uk/metabolights">http://www.ebi.ac.uk/metabolights</a>	Metal-binding sites in macromolecular structures
MetalPDB	<a href="http://metalweb.cerm.unifi.it/">http://metalweb.cerm.unifi.it/</a>	Spatial organization of metabolic reactions in the mouse
METscout	<a href="http://metscout.mpg.de">http://metscout.mpg.de</a>	Genome biology of the monarch butterfly <i>Danaus plexippus</i>
MonarchBase	<a href="http://monarchbase.umassmed.edu/">http://monarchbase.umassmed.edu/</a>	Chemogenomic experiments in yeast: connection of drug response to biological pathways, phenotypes, and networks
NetwoRx	<a href="http://ophid.utoronto.ca/networx/">http://ophid.utoronto.ca/networx/</a>	Free online books on the NCBI website
NCBI Bookshelf	<a href="http://www.ncbi.nlm.nih.gov/books">http://www.ncbi.nlm.nih.gov/books</a>	Non-human Primate Reference Transcriptome Resource
NHPRTTR	<a href="http://nhprtr.org/">http://nhprtr.org/</a>	Genetic tests and laboratories that perform them
NIH Genetic Testing Registry	<a href="http://www.ncbi.nlm.nih.gov/gtr/">http://www.ncbi.nlm.nih.gov/gtr/</a>	Naturally occurring Plant-based Anticancer Compound Targets
NPACT	<a href="http://crdd.osdd.net/raghava/npact/">http://crdd.osdd.net/raghava/npact/</a>	Genome expression database of <i>Oikopleura dioica</i>
OikoBase	<a href="http://oikoarrays.biology.uiowa.edu/Oiko/">http://oikoarrays.biology.uiowa.edu/Oiko/</a>	Microbial orthology resource
OrtholugeDB	<a href="http://www.pathogenomics.sfu.ca/ortholugedb/">http://www.pathogenomics.sfu.ca/ortholugedb/</a>	Small secreted proteins from rice
OrySPSSP	<a href="http://genportal.org/SPD/index.do">http://genportal.org/SPD/index.do</a>	A database of Papillomaviridae family of viruses
Papillomavirus Episteme	<a href="http://PaVE.niaid.nih.gov/">http://PaVE.niaid.nih.gov/</a>	Plant Genome Duplication Database
PGDD	<a href="http://chibba.agtec.uga.edu/duplication/">http://chibba.agtec.uga.edu/duplication/</a>	Plant Intron Exon Comparison and Evolution
PIECE	<a href="http://probes.pw.usda.gov/piece/">http://probes.pw.usda.gov/piece/</a>	tRNAs of plants and algae
PlantRNA	<a href="http://plantrna.ibmp.cnrs.fr/">http://plantrna.ibmp.cnrs.fr/</a>	

(continued)

Table 1. Continued

Database name	URL	Brief description
PR2	<a href="http://ssu-rna.org/">http://ssu-rna.org/</a>	Protist Ribosomal reference database
prePPI	<a href="http://bhapp.c2b2.columbia.edu/PrePPI">http://bhapp.c2b2.columbia.edu/PrePPI</a>	Predicted and experimentally determined protein–protein interactions for yeast and human
PTMcode	<a href="http://ptmcode.embl.de/">http://ptmcode.embl.de/</a>	Functional associations between posttranslational modifications within proteins
Quorumpeps	<a href="http://quorumpeps.ugent.be/">http://quorumpeps.ugent.be/</a>	A database of quorum-sensing peptides
RhesusBase	<a href="http://www.rhesusbase.org/">http://www.rhesusbase.org/</a>	A Knowledgebase for the Monkey Research Community
RiceFRIEND	<a href="http://ricefrend.dna.affrc.go.jp/">http://ricefrend.dna.affrc.go.jp/</a>	Rice Functionally Related gene Expression Network Database
RNApathwaysDB	<a href="http://rnb.genesilico.pl/">http://rnb.genesilico.pl/</a>	A database of RNA processing pathways
SecReT4	<a href="http://db-mml.sjtu.edu.cn/SecReT4/">http://db-mml.sjtu.edu.cn/SecReT4/</a>	Type IV Secretion system Resource
SEVA	<a href="http://seva.cnbc.csic.es/SEVA/">http://seva.cnbc.csic.es/SEVA/</a>	Standard European Vector Architecture: a collection of plasmids to analyse complex prokaryotic phenotypes
SIFTS	<a href="http://www.ebi.ac.uk/pdbe/docs/sifts/">http://www.ebi.ac.uk/pdbe/docs/sifts/</a>	Structure Integration with Function, Taxonomy and Sequences
SINEBase	<a href="http://sines.eimb.ru">http://sines.eimb.ru</a>	A database of short interspersed elements (SINEs)
SomamiR	<a href="http://compbio.uthsc.edu/SomamiR/">http://compbio.uthsc.edu/SomamiR/</a>	Somatic mutations that impact microRNA targeting in cancer
Spermatogenesis Online	<a href="http://mcg.ustc.edu.cn/sdap1/spermgenes/">http://mcg.ustc.edu.cn/sdap1/spermgenes/</a>	Spermatogenesis-related genes
SpliceAid-F	<a href="http://mi.caspuir.it/SpliceAidF/">http://mi.caspuir.it/SpliceAidF/</a>	Human splicing factors and their RNA-binding sites
Spliceosome Database	<a href="http://spliceosomedb.ucsc.edu/">http://spliceosomedb.ucsc.edu/</a>	Spliceosome genes and proteins, splicing complexes
StreptomeDB	<a href="http://streptomedb.pharmaceutical-bioinformatics.de">http://streptomedb.pharmaceutical-bioinformatics.de</a>	Antibiotic, anti-tumour and immunosuppressant drugs produced by <i>Streptomyces</i> spp.
SwissBioisostere	<a href="http://www.swissbioisostere.ch/">http://www.swissbioisostere.ch/</a>	Molecular replacements for ligand design
SwissSidechain	<a href="http://www.swissidechain.ch/">http://www.swissidechain.ch/</a>	Non-natural amino acid sidechains for protein engineering
SynSysNet	<a href="http://bioinformatics.charite.de/synsysnet/">http://bioinformatics.charite.de/synsysnet/</a>	Synapse proteins, their structures and interactions
TCMID	<a href="http://www.megabionet.org/tcmid/">http://www.megabionet.org/tcmid/</a>	Traditional Chinese Medicine Integrated Database
TFClass	<a href="http://www.edgar-wingender.de/huTF_classification.html">http://www.edgar-wingender.de/huTF_classification.html</a>	Human transcription factors classified according to their DNA-binding domains
TissueNet	<a href="http://netbio.bgu.ac.il/tissuenet/">http://netbio.bgu.ac.il/tissuenet/</a>	Tissue distribution of protein–protein interactions
TOPPR	<a href="http://iomics.ugent.be/toppr/">http://iomics.ugent.be/toppr/</a>	The Online Protein Processing Resource
TSGene	<a href="http://bioinfo.mc.vanderbilt.edu/TSGene/">http://bioinfo.mc.vanderbilt.edu/TSGene/</a>	Tumor Suppressor Gene database
UCNEbase	<a href="http://cgc.vital-it.ch/UCNEbase/">http://cgc.vital-it.ch/UCNEbase/</a>	Ultraconserved non-coding elements and gene regulatory blocks
UUCD	<a href="http://uucd.biocuckoo.org/">http://uucd.biocuckoo.org/</a>	Ubiquitin and ubiquitin-like conjugation database
ValidNESs	<a href="http://validness.ym.edu.tw/">http://validness.ym.edu.tw/</a>	Validated nuclear export signals-containing proteins
Voronia4RNA	<a href="http://proteininformatics.charite.de/voronia4rna_tools/v4rna/index">http://proteininformatics.charite.de/voronia4rna_tools/v4rna/index</a>	Packing of RNA molecules and complexes
WDDD	<a href="http://so.qbic.riken.jp/wddd/">http://so.qbic.riken.jp/wddd/</a>	Worm Developmental Dynamics Database
WholeCellKB	<a href="http://wholecellkb.stanford.edu/">http://wholecellkb.stanford.edu/</a>	Pathway and genome database of <i>Mycoplasma genitalium</i> for whole-cell modelling
WormQTL	<a href="http://www.wormqtl.org">http://www.wormqtl.org</a>	Natural variation data in <i>Caenorhabditis</i> spp.
YM500	<a href="http://ngs.ym.edu.tw/ym500/">http://ngs.ym.edu.tw/ym500/</a>	smRNA-seq database for miRNA research
ZInC	<a href="http://research.nhgri.nih.gov/zinc">http://research.nhgri.nih.gov/zinc</a>	Zebrafish Insertions Collection

Table 2. Database updates new for the NAR Database issue

Database name	URL	Previous article	Brief description
2P2Idb	<a href="http://dimr.cnrs-mrs.fr">http://dimr.cnrs-mrs.fr</a>	2010	Structural data on protein–protein interactions and their inhibitors
Allen Brain Atlas	<a href="http://www.brain-map.org">http://www.brain-map.org</a>	2009	Gene expression and neuroanatomical data on human and mouse brain
BioGPS	<a href="http://biogps.org">http://biogps.org</a>	2009	Gene annotation portal and a resource on gene and protein function
DARNED	<a href="http://beamish.ucc.ie/">http://beamish.ucc.ie/</a>	2010	Database of RNA Editing
DoriC	<a href="http://tubic.tju.edu.cn/doric/">http://tubic.tju.edu.cn/doric/</a>	2007	Replication origin ( <i>oriC</i> ) regions in bacterial and archaeal genomes
FlyAtlas	<a href="http://flyatlas.org/">http://flyatlas.org/</a>	2007	<i>Drosophila</i> gene expression atlas
GenColors	<a href="http://sgb.fli-leibniz.de/">http://sgb.fli-leibniz.de/</a>	2005	Genome annotation and comparison database for small genomes
Genomicus	<a href="http://www.dyogen.ens.fr/genomicus">http://www.dyogen.ens.fr/genomicus</a>	2010	Syntenic relationships between eukaryote genomes
InnateDB	<a href="http://www.innatedb.com/">http://www.innatedb.com/</a>	2008	A database of mammalian innate immune response
MicroScope	<a href="http://www.genoscope.cns.fr/agc/microscope/">http://www.genoscope.cns.fr/agc/microscope/</a>	2009	Microbial genome annotation and analysis platform
NPIDB	<a href="http://npidb.belozersky.msu.ru/">http://npidb.belozersky.msu.ru/</a>	2007	Nucleic acids–protein interaction database

NAR Database Issues in September 1994, then in January 1996, and each January after that.

In the past 20 years, the Database Issue has gradually grown in size before stabilizing at the level of ~180 articles. However, despite the almost 10-fold increase in the

number of published articles, the key topics of the current issue remain largely the same as 20 years ago. This issue again features articles from GenBank and the European Nucleotide Archive (formerly the EMBL Data Library), which, together with the DNA Data Bank of Japan, form

the International Nucleotide Sequence Database collaboration, INSDC (1–4). Just as 20 years ago, there are updates from SWISS-PROT and PIR (now combined into UniProt) and PROSITE (5,6).

Continuing the tradition of featuring well-curated databases of RNA sequences, this issue includes an update on SILVA, a widely used comprehensive database of bacterial, archaeal and eukaryotic 16S/18S and 23S/28S rRNA sequences (7), and a description of Protist Ribosomal Reference database (PR2), a new database that catalogs small subunit rRNA sequences from unicellular eukaryotes (8). An update on the Ribosomal Database Project, a constant feature of the *NAR* Database Issue since 1991 (9), was last published in 2009 (10). Other RNA databases in this issue include an update on Rfam (11), the universally acclaimed database of RNA families, as well as several databases on long non-coding RNA, microRNA and their targets. An update of MODOMICS, a database on RNA modification, is now supplemented by RNAPATHWAYSDB, a database of RNA maturation and decay pathways developed by the same group (12,13).

As before, this issue presents several transcription factor (TF) databases. Two of them cover TFs themselves: TFClass offers a classification of human TFs, while NPIDB presents structural information on DNA–protein and RNA–protein complexes (14,15). Several other databases collect information on the TF-binding sites. These include Factorbook, a database of TF-binding data from the ENCODE project; HOCOMOCO, a collection of human TF-binding sites; CTCFBSDB, a database of CCCTC-binding factor (CTCF)-binding sites; RegulonDB, a database of transcriptional regulation in *E. coli*; and SwissRegulon, a database of regulatory sites in human, mouse and yeast genomes and in model bacteria (16–20).

The structural databases featured in this issue all show a trend towards a better integration and cross-referencing tools. This refers both to the updates of well-known databases, such as the RCSB Protein Data Bank (PDB), CATH and PDBTM, and to such databases as EBI's SIFTS, a joint effort of UniProt and PDB to provide a residue level mapping of their entries and supplement it with annotation from other public databases; Genome3D, a recent collaborative project aiming to provide structural annotation from CATH and SCOP to the genomic sequences; and dcGO, which develops domain-centric ontologies to link protein domains with functions, phenotypes and diseases (21–23).

Likewise, with *E. coli* remaining the workhorse of molecular biology, this issue includes update articles on the EcoGene (the first one since 2000), EcoCyc and RegulonDB databases, as well as a description of the newly developed *E. coli* Metabolome Database (20,24–26).

## HUMAN DISEASE GENOMICS—THE NEXT FRONTIER?

As discussed earlier (27), the original GDB did not survive the influx of the new data and multiple changes of ownership. Nevertheless, we now have a wide variety of

databases that cover different aspects of human genome and genomes of model organisms. This issue features annual updates from Ensembl and ENCODE projects and from the UCSC Genome Browser and the Japanese H-InvDB database (28–31). The model organism databases are represented by the updates to FlyBase, Mouse Genome database, Xenbase and ZFIN (32–35).

Two new databases, RhesusBase and NHPRTR, present extensive genome and RNAseq data for non-human primates, including great apes, old world monkeys, new world monkeys and prosimians (36,37). These data could go a long way towards establishing monkeys as model organisms for comparative genomics studies. One more database is dedicated to a more distant relative of human, the urochordate *Oikopleura dioica* (38).

A potentially important development is the construction of two new databases of repetitive DNA elements, Dfam and SINEBase (39,40). Along with the industry standard Repbase Update (41,42) and monthly RepBase Reports (<http://www.girinst.org/repbase/reports/>), these databases promise to contribute to a better understanding of eukaryotic repeat elements.

With the abundance of databases providing valuable tools for genome analysis, there is a clear trend towards bringing genomics 'from the bench to the bedside', i.e. using genomic data for a better understanding and, hopefully, better treatment of human disease. A number of projects, including ClinSeq (<http://www.genome.gov/20519355>), DDD (<http://www.ddduk.org/>) and UK10K (<http://www.uk10k.org/>) are working towards these goals, and several databases featured in this issue represent important steps in this direction. Last year's issue introduced the GWASdb database of human genetic variants identified by genome-wide association studies (43). GWAS Central, established in 2007 as HGVbaseG2P (44), has been revamped and now includes data from over 1000 studies. Now, a joint article from NCBI and EBI describes their databases of genomic structural variation, dbVar and DGVA (45). These databases cover diverse variation data including inversions, insertions and translocations that are >50 bp in length. NCBI is also developing ClinVar (<http://www.ncbi.nlm.nih.gov/clinvar/>), a database of relationships between human gene variation and the observed health status (46). The task of streamlining the genetic tests that provide such information is taken up by the recently created NIH Genetic Testing Registry, a database of genetic tests and laboratories that perform them, with detailed information about what exactly is measured in each test and its analytic and clinical validity (47).

The impact of the genomic data on developing targeted approaches for fighting disease is particularly evident in the case of cancer. This issue features updates from three great databases, the UCSC Cancer Genome Browser (48), the Atlas of Genetics and Cytogenetics in Oncology and Haematology (49) and the TP53 website [(50), the first update of the database on tumor factor p53 mutations since 1997]. In addition, there are two new databases dedicated to studying cancer at the level of specific cell lines. The CellLineNavigator database provides gene

expression profiles of different cancer cell lines in different pathological states (51), whereas the Genomics of Drug Sensitivity in Cancer (GDSC) collects the results of high-throughput studies examining the sensitivity for anti-cancer drugs in various cell lines (52).

### CURATION OF THE NAR DATABASE COLLECTION

During the past 20 years, all databases featured in the *NAR* Database Issues were added to the *NAR* online Molecular Biology Database Collection, available at <http://www.oxfordjournals.org/nar/database/a/>. With the annual attrition rate of <5%, this Collection has been steadily growing and, in 2012, exceeded 1400 database entries (53). It was clear that the list was due for a serious clean-up, and one of the authors (XMFS) devised and set in motion a semi-automated procedure to identify obsolete and non-responsive websites. Remarkably, >90% of the databases listed in the last year's release of the online Collection were found to be functional. Corresponding authors of close to a hundred non-responsive resources had been contacted and 44 websites (~3.2% of the total) have been approved for deletion. About 100 entries in the Collection have been updated by receiving corrected URLs, summaries highlighting recent developments, or some other changes in the deposited data.

Although deletion of 40 databases was well within the average drop-off rate and was hardly surprising, further analysis revealed that most of these resources were not lost. Instead, in the normal course of database evolution, they have been integrated into larger projects. For example, a couple of segmental duplications databases were merged into the Database of Genomic Variants (54), *NAR* Database Collection entry no. 655, while the NCBI's Cancer Chromosomes database has been merged into dbVar [described in detail in this issue, (45)]. Further, improved annotation of the human genome made redundant a number of resources that covered specific areas of the genome (e.g. the IXDB with its physical maps of human chromosome X).

In one instance, the ExDom database of exon-intron structures of genes in seven eukaryotic genomes (55) had to be removed from the Collection, as it has taken the commercial route and does not provide a free version anymore, although the author's company offered a discounted version for academic users. Unfortunately, the tightening budgets (56) might force other databases to follow the same path.

In total, the *NAR* online Molecular Biology Database Collection now includes 1512 databases sorted into 14 categories and 41 subcategories. The authors wishing to have their databases, published elsewhere, to be included in the Collection are welcome to contact XMFS directly.

### ACKNOWLEDGEMENTS

The authors thank Drs Javier Herrero and Michael Schuster for helpful comments and the Oxford

University Press team led by Jennifer Boyd and Andrew Malvern for their help in compiling this issue.

### FUNDING

Intramural Research Program of the U.S. National Institutes of Health at the National Library of Medicine [to M.Y.G.]. Funding for open access charge: Waived by Oxford University Press.

*Conflict of interest statement.* The authors' opinions do not necessarily reflect the views of their respective institutions.

### REFERENCES

- Benson,D., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
- Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
- Ogasawara,O., Mashima,J., Kodama,Y., Kaminuma,E., Nakamura,Y., Okubo,K. and Takagi,T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.*, **41**, D25–D29.
- Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41**, D21–D24.
- The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
- Sigrist,C.J.A., de Castro,E., Cerutti,L., Cucho,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
- Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
- Guillou,L., Bachar,D., Audic,S., Bass,D., Berney,C., Bittner,L., Boutte,C., Burgaud,G., de Vargas,C., Decelle,J. *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small subunit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**, D597–D604.
- Olsen,G.J., Larsen,N. and Woese,C.R. (1991) The ribosomal RNA database project. *Nucleic Acids Res.*, **19**, 2017–2021.
- Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
- Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
- Machnicka,M.A., Milanowska,K., Osman Oglou,O., Purta,E., Kurkowska,M., Olchowik,A., Januszewski,W., Kalinowski,S., Dunin-Horkawicz,S., Rother,K.M. *et al.* (2013) MODOMICS: a database of RNA modification pathways—2012 update. *Nucleic Acids Res.*, **41**, D262–D267.
- Milanowska,K., Mikołajczak,K., Lukasik,A., Skorupski,M., Balcer,Z., Mika,M., Rother,K.M. and Bujnicki,J.M. (2013) RNAPATHWAYSDB—a database of RNA maturation and decay pathways. *Nucleic Acids Res.*, **41**, D268–D272.
- Wingender,E., Schoeps,T. and Dönitz,J. (2013) TFClass: an expandable classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–D170.

15. Kirsanov,D., Zaneagina,O., Spirin,S., Karyagina,A. and Alexeevski,A. (2013) NPIDB: Nucleic acids – Protein Interaction DataBase. *Nucleic Acids Res.*, **41**, D517–D523.
16. Wang,J., Zhuang,J., Iyer,S., Lin,X., Greven,M., Kim,B., Moore,J., Dong,X., Virgil,D., Birney,E. *et al.* (2013) Factorbook.org: a wiki-based database for transcription factor binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.
17. Kulakovskiy,I.V., Medvedeva,Y.A., Schaefer,U., Kasianov,A.S., Vorontsov,I.E., Bajic,V.B. and Makeev,V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195–D202.
18. Ziebarth,J., Bhattacharya,A. and Cui,Y. (2013) CTCFBSDB 2.0: a database for CTCF binding sites and genome organization. *Nucleic Acids Res.*, **41**, D188–D194.
19. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, **41**, D214–D220.
20. Salgado,H., Peralta,M., Gama-Castro,S., Santos-Zavaleta,A., Muñoz-Rascado,L.J., Garcia-Sotelo,J.S., Weiss,V., Solano-Lira,H., Martinez-Flores,I., Medina-Rivera,A. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards, and more. *Nucleic Acids Res.*, **41**, D203–D213.
21. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfield,T.J., O'Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
22. Lewis,T.E., Sillitoe,I., Andreeva,A., Blundell,T.L., Buchan,D., Chothia,C., Cuff,A., Dana,J.M., Filippis,I., Gough,J. *et al.* (2013) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res.*, **41**, D499–D507.
23. Fang,H. and Gough,J. (2013) dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.*, **41**, D536–D544.
24. Zhou,J. and Rudd,K.E. (2013) EcoGene 3.0. *Nucleic Acids Res.*, **41**, D613–D624.
25. Keseler,I.M., Mackie,A., Peralta-Gil,M., Santos-Zavaleta,A., Gama-Castro,S., Bonavides-Martinez,C., Fulcher,C., Huerta,A.M., Kothari,A., Krummenacker,M. *et al.* (2013) EcoCyc: fusing model-organism databases with systems biology. *Nucleic Acids Res.*, **41**, D605–D612.
26. Guo,A.C., Jewison,T., Wilson,M., Liu,Y., Knox,C., Djoumbou,Y., Lo,P., Mandal,R., Krishnamurthy,R. and Wishart,D.S. (2013) ECMDB: the *E. coli* Metabolome Database. *Nucleic Acids Res.*, **41**, D625–D630.
27. Galperin,M.Y. and Cochrane,G.R. (2009) Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res.*, **37**, D1–D4.
28. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
29. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Wong,M.C., Kirkup,V.M., Maddren,M., Fang,R., Heitner,S. *et al.* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
30. Meyer,L.R., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G., Rhead,B., Raney,B.J. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
31. Takeda,J., Yamasaki,C., Murakami,K., Nagai,Y., Sera,M., Hara,Y., Obi,N., Habara,T., Imanishi,T. and Gojohori,T. (2013) H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Res.*, **41**, D915–D919.
32. Marygold,S.J., Leyland,P.C., Seal,R.L., Goodman,J.L., Thurmond,J., Strelets,V.B. and Wilson,R.J. (2013) FlyBase: improvements to the bibliography. *Nucleic Acids Res.*, **41**, D751–D757.
33. Bult,C.J., Eppig,J.T., Blake,J.A., Kadin,J.A., Richardson,J.E. and The Mouse Genome Database Group (2013) The Mouse Genome Database (MGD): phenotype, function and models of human disease. *Nucleic Acids Res.*, **41**, D885–D891.
34. James-Zorn,C., Ponferrada,V.G., Jarabek,C.J., Burns,K., Segerdell,E.J., Lee,J., Synder,K., Bhattacharyya,B., Karpinka,J.B., Fortriede,J. *et al.* (2013) Xenbase: expansion and updates of the *Xenopus* model organism database. *Nucleic Acids Res.*, **41**, D865–D870.
35. Howe,D.G., Bradford,Y.M., Conlin,T., Eagle,A.E., Fashena,D., Frazer,K., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2013) ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res.*, **41**, D854–D860.
36. Zhang,S.J., Liu,C.J., Shi,M., Kong,L., Chen,J.Y., Zhou,W.Z., Zhu,X., Yu,P., Wang,J., Yang,X. *et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res.*, **41**, D892–D905.
37. Pipes,L., Li,S., Bozinoski,M., Palermo,R., Peng,X., Blood,P., Kelly,S., Weiss,J., Thierry-Mieg,J., Thierry-Mieg,D. *et al.* (2013) The Nonhuman Primate Reference Transcriptome Resource (NHPRT) for comparative functional genomics. *Nucleic Acids Res.*, **41**, D906–D914.
38. Danks,G., Campsteijn,C., Parida,M., Butcher,S., Doddapaneni,H., Fu,B., Petrin,R., Metpally,R., Lenhard,B., Wincker,P. *et al.* (2013) OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res.*, **41**, D845–D853.
39. Wheeler,T.J., Clements,J., Eddy,S.R., Hubley,R., Jones,T.A., Jurka,J., Smit,A.F.A. and Finn,R.D. (2013) Dfam: a database of repetitive DNA based on profile Hidden Markov Models. *Nucleic Acids Res.*, **41**, D70–D82.
40. Vassetzky,N. and Kramerov,D. (2013) SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.*, **41**, D83–D89.
41. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
42. Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412.
43. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.
44. Thorisson,G.A., Lancaster,O., Free,R.C., Hastings,R.K., Sarmah,P., Dash,D., Brahmachari,S.K. and Brookes,A.J. (2009) HGVBbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.
45. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,D., Chen,C., Maguire,M., Corbett,M., Zhou,Z., Paschall,J. *et al.* (2013) dbVar and DGVA: Public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
46. NCBI Resource Coordinators. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
47. Rubinstein,W., Maglott,D., Lee,J., Kattman,B., Malheiro,A., Ovetsky,M., Hem,V., Gorelenkov,V., Song,G., Wallin,C. *et al.* (2013) The NIH Genetic Testing Registry: A new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.*, **41**, D925–D935.
48. Goldman,M., Craft,B., Swatloski,T., Ellrott,K., Cline,M., Diekhans,M., Ma,S., Wilks,C., Stuart,J., Haussler,D. *et al.* (2013) The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res.*, **41**, D949–D954.
49. Huret,J.L., Ahmad,M., Arsaban,M., Bernheim,A., Cigna,J., Desangles,F., Guignard,J.-C., Jacquemot-Perbal,M.-C., Labarussias,M., Leberre,V. *et al.* (2013) Atlas of Genetics and Cytogenetics in Oncology and Haematology in 2013. *Nucleic Acids Res.*, **41**, D920–D924.
50. Leroy,B., Fournier,J.L., Ishioka,C., Monti,P., Inga,A., Fronza,G. and Soussi,T. (2013) The TP53 web site: an integrative resource

- centre for the TP53 mutation database and TP53 mutant analysis. *Nucleic Acids Res.*, **41**, D962–D969.
51. Krupp, M., Itzel, T., Maass, T., Hildebrandt, A., Galle, P.R. and Teufel, A. (2013) CellLineNavigator: a workbench for cancer cell line analysis. *Nucleic Acids Res.*, **41**, D942–D948.
52. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindahl, N., Beare, D., Smith, J.A., Thompson, I.R. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.
53. Galperin, M.Y. and Fernandez-Suarez, X.M. (2012) The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **40**, D1–D8.
54. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. and Scherer, S.W. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.*, **115**, 205–214.
55. Bhasi, A., Philip, P., Manikandan, V. and Senapathy, P. (2009) ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, **37**, D703–D711.
56. Baker, M. (2012) Databases fight funding cuts. *Nature*, **489**, 19.