The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection

Daniel J. Rigden¹,* and Xosé M. Fernández²

¹Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and ²Institut Curie, 25 rue d’Ulm, 75005 Paris, France

**ABSTRACT**

The 2021 *Nucleic Acids Research* database Issue contains 189 papers spanning a wide range of biological fields and investigation. It includes 89 papers reporting on new databases and 90 covering recent changes to resources previously published in the Issue. A further ten are updates on databases most recently published elsewhere. Seven new databases focus on COVID-19 and SARS-CoV-2 and many others offer resources for studying the virus. Major returning nucleic acid databases include NONCODE, Rfam and RNAcentral. Protein family and domain databases include COG, Pfam, SMART and Panther. Protein structures are covered by RCSB PDB and dispersed proteins by PED and MobiDB. In metabolism and signalling, STRING, KEGG and WikiPathways are featured, along with returning KLIFS and new DKK and KinaseMD, all focused on kinases. IMG/M and IMG/VR update in the microbial and viral genome resources section, while human and model organism genomics resources include Flybase, Ensembl and UCSC Genome Browser. Cancer studies are covered by updates from canSAR and PINA, as well as newcomers CNCDatabase and Oncovar for cancer drivers. Plant comparative genomics is catered for by updates from Gramene and GreenPhyDB. The entire Database Issue is freely available online on the *Nucleic Acids Research* website (https://academic.oup.com/nar). The NAR online Molecular Biology Database Collection has been substantially updated, revisiting nearly 1000 entries, adding 90 new resources and eliminating 86 obsolete databases, bringing the current total to 1641 databases. It is available at https://www.oxfordjournals.org/nar/database/c/.

**NEW AND UPDATED DATABASES**

The 28th annual *Nucleic Acids Research* Database Issue contains 189 papers spanning, as usual, a wide range of biology. Unsurprisingly, COVID-19 casts a long shadow over the Issue. Seven new databases specifically address the pandemic and the SARS-CoV-2 virus responsible (Table 1) but new and returning databases in all areas have rushed to support research into the viral pandemic: the reader will find reference to it throughout the Issue, sometimes in quite unexpected places. The Issue contains a further 82 papers (Table 2) on new databases as well as 90 update papers on databases previously published in NAR. To complete the Issue, resources previously published elsewhere update in a further 10 papers (Table 3).

As is customary, the Issue starts with reports from the major database providers at the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the National Genomics Data Center (NGDC) in China (1–3). Thereafter, the usual categorisation applies: (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signaling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases. Many resources are not easily pigeon-holed so browsing of the whole Issue is strongly encouraged.

The COVID-19 papers span a number of sections clearly indicating the multidisciplinary nature of the huge scientific response to the pandemic. Navigating the deluge of COVID-19 papers is a significant challenge in its own right and one addressed by the NCBI’s LitCovid database (4) which features manual curations supported by sophisticated machine-learning assistance. SARS-CoV-2 nucleic acid sequence data and associated curated metadata can be conveniently obtained from the ViruSurf database (5) which also covers other human pathogenic viruses. SARS-CoV-2 comparative genomics is covered by the GESS database (6) where temporal and geographical patterns of SNVs can be analysed. SARS-CoV-2 protein structures – alone and in complex with antibodies, receptors, and small molecules – are collected at the CoV3D database (7) and made available with a variety of bespoke analyses of sequential and conformational diversity. Obviously, drug and vaccine de-
development are the primary drivers of SARS-CoV-2 protein structure determination. Supporting these efforts is the COVID19 Drugs Repository (8) which curates and integrates experimental findings from the literature with information from well-established small molecule databases to support drug repurposing. In a similar vein, DockCoV2 (9) contains in silico docking results for already approved drugs against putative drug targets in the SARS-CoV-2 proteome. Finally, recognising the therapeutic importance of an understanding of the host response, PAGER-CoV (10) covers pathways and gene lists relating to the viral infection and its consequences.

In the ‘Nucleic acid databases’ section, two significant returning databases focus on ncRNAs. The major news from the NONCODE database (11) is the inclusion of plant IncRNAs from 23 species, including their gene expression, function annotation and sequence conservation between species. NONCODE also reports new efforts to document associations between mammalian IncRNAs and cancers. MNDR (12) reports progress on a number of fronts—including the addition of circRNA-disease associations and ncRNA subcellular localization—resulting in a quadrupling of database entries. Pfam (13) reaches version 14 with new RNA families, some community-derived (including from masters students the paper reports) using a new Pfam Cloud platform, and including coronavirus and flavivirus entries. Integration with RNAcentral, also updating here (14), is evident in a new sequence search and work to make miRNA families more consistent with fellow member miRBase (15). RNAcentral continues to grow impressively, now encompassing 44 member ncRNA databases. Resources added since the last publication expand content in different directions, adding new classes of ncRNA such as snoRNAs, new links to diseases and new organism coverage.

Elsewhere, a number of databases focus on the structure, at small or large scale, of nucleic acids. The new TBDB (16) focuses on T-box riboswitch:RNA pairs, where secondary structure modelling is required to predict functionality, while RASP (17) collects experimental data reporting on RNA secondary structure from a variety of sources including SARS-CoV-2. A major new arrival is the Nucleosome Data Bank (18), which brings together a repository of experimental structural genomic data, computational tools for the modelling of structures, and visualisation of structures in the context of sequential data. They also introduce, by analogy with PDB files, an NDB file format. In the same area, 3DIV (19) reports an update focusing largely on 3D cancer genomes and commenting on the consequences for chromatin structure of the DNA structural variations associated with some cancers. Finally, DNAmoreDB (20) collects data on catalytic DNA molecules or DNAzymes, already demonstrated to catalyse 20 different reactions with that number sure to grow as research on their practical applications continues.

In the section on protein sequence and structure databases, users of protein family and domain databases are particularly well-served. COG (21) returns to report well over 200 new protein families and a similar number of families updated to reflect recent experimental characterisation. The current COG includes almost 5000 COGs and annotates genomes of over 1200 microbes. Similarly popular resources Pfam, SMART and Panther also report updates. Pfam (22) adds over 350 new families, including entries for previously unmatched SARS-CoV-2 proteins, and improves the consistency of its content with the specialist RepeatsDB resource, also updating here (23). A major strategic decision reintroduces the automatically generated Pfam-B supplement of non-curated sequence clusters calculated using more efficient computational methods. SMART (24), in contrast, does not aim for complete proteome coverage and the most recent targeted focus has been mobile genetic elements in bacteria and archaeobacteria. The Panther update (25) reports an interesting new facility for assigning a function to a new sequence using tree grafting whereby Gene Ontology terms are attached to a protein according to the position at which it is best accommodated in the tree.

In the area of protein structure, the RCSB Protein Data Bank (26) reminds us that this venerable database turns 50 in 2021, having grown from just seven entries then (27), and distribution via magnetic tape, to ~170 000 now. Its increasingly comprehensive analytical and visualisation features are exemplified using SARS-CoV-2 structures in the update paper. Another database with a long history, ProThermDB, covering thermodynamic parameters relating to protein stability, returns revitalised and near-doubled in size after 15 years (28). Two new databases ThermoMutDB (29) and FireProtDB (30) offer competition in the same area. For proteins without regular stable structures, the Issue offers a trio of databases. The main innovations at the returning database MobiDB (31) cover functional annotations relating to disordered regions; regions undergoing a disorder to order transition on binding, predicted linear interaction motifs, post-translational modifications and regions implicated in phase separation. PED (32), also publishing an update, covers experimentally characterized structural ensembles of disordered regions and proteins. These are joined by the newcomer MemMoRF (33) which covers features within disordered regions that can interact with biological membranes. Finally the Gene Ontology (34), a cornerstone resource across all sections, contributes an update describing how collaborations with expert groups and databases

Table 1. Descriptions of new databases related to COVID-19 in the 2021 NAR database Issue

| Database name          | URL                                                   | Short description                                                                 |
|------------------------|-------------------------------------------------------|-----------------------------------------------------------------------------------|
| CoV3D                  | https://cov3d.ibbr.umd.edu                           | Experimental coronavirus protein structures                                         |
| COVID19 Drug Repository | http://covid19.md.biu.ac.il                          | Curated papers on COVID-19 drugs and drug repurposing                              |
| DockCoV2               | https://covirus.cc/drugs/                            | In silico drug docking against SARS-CoV2 targets                                   |
| GESS                   | https://wan-bioinfo.shinyapps.io/GESS/              | Global Evaluation of SARS-CoV2 Sequences                                           |
| LitCovid               | https://www.ncbi.nlm.nih.gov/research/coronavirus/  | Curated COVID-19 Literature                                                        |
| PAGER-COV              | http://discovery.informatics.uab.edu/PAGER-COV/     | Pathways and gene lists related to COVID-19                                          |
| ViruSurf               | http://gmql.eu/virusurf/                             | Portal to SARS-CoV2 sequences and variants                                          |
## Table 2. Descriptions of new databases in the 2021 *NAR* database Issue not specifically related to COVID-19

| Database name         | URL                                      | Short description                                                                 |
|-----------------------|------------------------------------------|------------------------------------------------------------------------------------|
| Aging Atlas           | https://bigd.big.ac.cn/aging             | Aging related omics data                                                           |
| Animal-APAdb          | http://gong.lahzau.edu.cn/Animal-APAdb/  | Alternative polyadenylation in animals                                              |
| AcrDB                 | http://bcb.unl.edu/AcrDB/               | Anti-CRISPR operons                                                                |
| AcrHub                | http://paciispr.erc.monash.edu/AcrHub   | Anti-CRISPR proteins                                                               |
| ATACdb                | http://www.licpathway.net/ATACdb        | Human Assay-for-Transposable-Accessible Chromatin data                              |
| AtMAD                 | http://www.megabionet.org/atmad         | Arabidopsis thaliana Multi-omics Association Database                               |
| BastionHub            | http://bastionhub.erc.monash.edu/       | Substrates of Gram-negative secretion systems                                        |
| Big-FAM               | http://bigfam.bioinformatics.nl          | Biosynthetic gene cluster families                                                  |
| CancerImmunityQTL     | http://www.cancerimmunityqtl-hust.com/   | ImmuneQTLs across multiple cancer types                                              |
| Chewie-NS             | https://chewbbaica.online/              | Gene-by-gene schemas for microbial strain identification                            |
| CNMNP                 | http://www.cnmnp.org/                   | Comprehensive Marine Natural Product Database                                        |
| CNCDatabase           | http://cnncdatabase.med.cornell.edu/     | Cancer drivers at non-coding regions                                                |
| cncRNAdb              | http://www.rna-society.org/cncraddb/     | Coding and non-coding RNA                                                           |
| ConJuPeDB             | http://conjupepdb.ttk.hu/               | Drug-peptide conjugates                                                             |
| CoViNDb               | http://caas.jxa.cn/coivdb/              | Covalenl Inhibitor DataBase                                                         |
| crispSQL              | http://www.crispsql.com/                | CRISPR/Cas9 Off-Target Cleavage Assays                                              |
| CRISPR-view           | http://crisprviewdatabase.weililab.org/  | Functional genetic functions                                                        |
| CSEA-DB               | http://bioinfo.uth.edu/CSEADB/          | Cell type specificity of genetic traits                                              |
| CVS                   | http://cvs.babelomics.org/              | Comparative genomics of cyanobacteria                                               |
| Cytanorak             | http://www.sb-roscoff.fr/cytanorak      | Molecular data for modeling biochemical networks                                     |
| Datanator             | http://datanator.info/                  | Experimentally-characterized CAZyme gene clusters                                    |
| dbCAN-PUL             | http://bcb.unl.edu/dbCAN_PUL/home        | Manually curated and functionally validated guide RNAS                              |
| dbGuide               | https://sgrnascorer.cancer.gov/dbguide  | Residue level prediction of structure and function across proteomes                 |
| DescribePROT          | http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/ | Dark Kinome Knowledgebase                                                        |
| DKK                   | https://darkkinome.org/                 | Alternative splicing and protein-protein interactions                                |
| DIGGER                | https://exbio.wzw.tum.de/digger         | DNAzymes, i.e. DNA molecules with catalytic activity                                 |
| DNAMoreDB             | http://www.genesiclo.pl/DNAMoreDB       | 100 million compounds for virtual screening                                          |
| DrugSpaceX            | http://www.tartaglia.com/dualseq/       | Dual RNA-seq host-pathogen sequencing                                                |
| DualSeqDB             | https://dualseq.chenyi.com/              | Protein stability data                                                               |
| FireProtDB            | http://loschmidt.chemi.muni.cz/fireprotdb | WDCM 10K sequencing projects                                                        |
| gcType                | http://gctype.wdcm.org/                 | Host Genetic and Immune Factors Shaping Human Microbiota                             |
| GIMICA                | https://idrlab.uth.com/gimica/          | Raw Mass Spectrometry glycomics data                                                |
| GlycoPOST             | https://glycoast.glycosmos.org/          | TF-target relationships inferred from single cell and bulk RNA-seq datasets         |
| GRNdb                 | http://www.grndb.com/                   | Global Substance Registration System                                                |
| GSRS                  | https://gsrs.ncats.nih.gov/app/substances | Human enhancer RNA Atlas                                                            |
| HeRA                  | https://hanlab.uth.edu/HeRA/            | High-throughput Experiment- and Reference-guided database of TCM                    |
| HERB                  | http://herb.ac/                         | Human and mouse housekeeping genes                                                  |
| Housekeeping Transcript | http://www.housekeeping.unicamp.br/     | Curated and standardized metadata for human metagenomes                              |
| HumanMetagenomeDB     | https://webapp.ufz.de/hmgdb/             | integrated CRISPR Screens DataBase                                                   |
| iCSDB                 | http://www.kobic.re.kict/icsdb          | Infertility Disease DataBase                                                         |
| iDDB                  | http://mdl.shsmu.edu.cn/iDDB/            | 'iModulons', groups of independently-modulated genes                               |
| iModulonDB            | https://imodulondb.org/                 | Genetic variation in 1000 Indian individuals                                         |
| IndiGenomes           | http://cлинген.ibig.res.in/indiigen/    | Interactome of Drug-Metabolizing Enzymes                                            |
| INTEDE                | https://idrlab.uth.com/intede/          | Kinase Mutations and Drug responses                                                  |
| KinaseMD              | https://bioinfo.uth.edu/kmd/            | Human IncRNA expression database                                                     |
| LncCell               | http://bio-bgdatalab.hbrmu.edu.cn/LncCell/ | The N6-methyladenosine (m6A) epitranscriptome                                       |
| LncExpDB              | https://bigd.big.ac.cn/lncexpdb         | The N6-methyladenosine (m6A) epitranscriptome                                       |
| LncSEA                | http://bio.liclab.net/LncSEA/           | Pharmacokinetics data from clinical trials and pre-clinical research                |
| m6A-Atlas             | http://www.xjtlu.edu.cn/biologicalsciences/atlases | Oceanographic omics datasets linked to environmental metadata                      |
| markerDB              | http://www.markerdb.ca/                 | Reference panels and imputation methods for plant genomes                           |
| MASI                  | http://www.aidlab.com/MASI/              | Microbiota – Active Substance Interactions database                                 |
| MeDAS                 | https://das.chenlab.com/                | Alternative splicing during development in 20 species                               |
| MemMoRF               | http://memnor.fhegelab.org/             | Membrane-Binding Molecular Recognition Features                                     |
| mMGE                  | http://mgedb.comp-sysbio.org/            | Human metagenomic extrachromosomal mobile genetic elements                           |
| MolluscDB             | http://mgb.biocloud.net/home             | Comparative genomics of molluscs                                                    |
| Nucleome Data Bank    | https://ndb.rice.edu/                   | 3D genome structures and simulations                                                |
| OncoVar               | https://oncовар.org/                   | Driver mutations, genes and pathways in cancer                                       |
| Open Targets Genetics | https://genetics.opentargets.org/       | Drug targets prioritised from genetic data                                          |
| PCAT                  | http://pcatranscriptome.org/            | Pediatric cancer transcriptome explorer                                              |
| Peryton               | https://dianalah-e-ce.uth.gr/peryton    | Microbe–disease associations                                                        |
| PhcLiGe               | https://phelge.com/                     | Genotype–phenotype associations                                                     |
| PhycoCosm             | http://phyccosm.jgi.doe.gov/            | Comparative genomics of algae                                                       |
| PK-DB                 | https://pk-db.com/                      | Pharmacokinetics data from clinical trials and pre-clinical research                |
| Planet Microbe        | https://www.planetmicrobe.org/          | Oceanographic omics datasets linked to environmental metadata                      |
| Plant-ImputeDB        | http://gong.lahzau.edu.cn/Plant_imputeDB | Reference panels and imputation methods for plant genomes                           |
are being used to enrich, expand and improve the accuracy of the ontology.

In the metabolic and signalling section, the hugely popular STRING database of functional associations (35) offers an update that describes improvements to the functional enrichment tests available when users upload lists of proteins. STRING also now allows users to visualise and score all functional associations, as previously, or to limit visualisation and scoring to only physical interactions. The equally influential KEGG returns (36) with a detailed analysis of the impact of mutations on kinase structures and their consequences for drug binding. Finally, two new databases feature gene clusters. BiG-FAM (43), from the developers of the well known antiSMASH database (44), contains families of Biosynthetic Gene Clusters and enables a better understanding (and exploitation) of the full range of microbial natural product synthetic capacity. dbCAN-PUL (45) also builds on and complements an earlier database, in this case dbCAN-seq (46), and offers an online repository of experimentally characterised Polysaccharide Utilization Loci.

The microbial genomics section begins with a pair of databases devoted to anti-CRISPR (Acr) proteins. AcrHub (47) brings together over 300 experimentally characterised Acr proteins, 70 000 predicted Acr proteins and a variety of modules for characterising potential Acrs in user-uploaded sequences. AcrDB (48) detects and presents operons containing Acr-coding genes and nearby Aca regulatory proteins and uses machine learning to score them further. For comparative genomics, sister resources from the JGI, IMG/M and IMG/VR report updates. IMG/M (49) contains billions of genes from genomes and metagenomes

| Table 2. Continued |
|-------------------|------------------|--------------------------------------------------|
| **Database name** | **URL**          | **Short description**                              |
| PROTAC-DB         | http://cadd.zju.edu.cn/protacdb/ | Proteolysis-targeting chimeras (PROTACs)          |
| RASP              | http://rasp.zhanglab.net/         | RNA secondary structure probing data              |
| RBP2GO            | http://bp2go.oki.dz.de            | RNA-binding proteins across species               |
| RJunBase          | www.RJunBase.org                 | RNA splice junctions                              |
| RMDisease         | http://www.xjtlu.edu.cn/biosciences/rmd | Genetic variants that affect RNA modifications vs disease |
| RMVar             | http://rmvar.renlab.org           | Genetic variants that affect RNA modifications vs disease |
| SC2dsisease       | http://easybioio.com/sc2dsisease/ | Single cell transcriptomics data and disease      |
| SilencerDB        | http://health.tsinghua.edu.cn/silencerdb | Human silencers, validated or predicted          |
| STAB              | http://stab.comp-sysbio.org       | Spatio-Temporal Cell Atlas of the Human Brain    |
| TBDDB             | https://tbd.db.io                 | Structurally annotated T-box riboswitch: tRNA pairs |
| TCRdb             | http://bioinfo.life.hust.edu.cn/TCRdb | T-cell receptor (TCR) sequences                 |
| ThermoMutDB       | http://biois.unimelb.edu.au/thermomutdb | Protein Mutation Thermodynamics Database     |
| TISCH             | http://tisch.comp-genomics.org    | Uniformly processed tumor (and microenvironment) scRNA-seq data |
| TransCirc         | https://www.biosino.org/transcirc/ | Protein coding potential of circRNAs           |
| tRFtarget         | http://trf.target.net             | Targets of tRNA-derived fragments                |
| tsRBase           | http://tsrbase.org                | tRNA-derived small RNA expression and function   |
| VARAdb            | http://www.licpathway.net/VARAdb/ | Variants and regulatory information             |

| Table 3. Updated descriptions of databases most recently published elsewhere |
|-------------------|------------------|--------------------------------------------------|
| **Database name** | **URL**          | **Short description**                              |
| Bgee              | https://bgee.org/ | Curated wild-type animal gene expression data     |
| CellMinerCDB      | https://discover.nci.nih.gov/cellminercdb/ | Cell line-based pharmacogenomics datasets         |
| GeneLab           | https://gene.lab.nasa.gov/ | Omics data relating to space biology and ionising radiation |
| HMPDACC           | https://portal.hmpdacc.org/ | Human Microbiome Project Data Coordination Center  |
| miRNASNP          | http://bioinfo.life.hust.edu.cn/miRNASNP/ | miRNA-related SNPs and mutations                  |
| MitImpact         | https://mitimpact.css-mendel.it/ | Precomputed pathogenicity predictions for human mitochondrial genome mutations |
| ModelSEED Biochemistry | https://modelseed.org/biochem | Biochemical reactions                              |
| PLoCDB            | http://phgntobacco.com/ | Plant long non-coding RNA                         |
| Project Score     | https://score.depmap.sanger.ac.uk/ | CRISPR-Cas9 screens to identify cancer dependencies |
| TREND-DB          | http://shiny.imbei.uni-mainz.de:3838/trend-db/ | Conditional alternative polyadenylation            |
and allows flexible mining and comparison between selected groups of sequences. IMG/VR (50) has near-tripled in size since its last update paper, its content dominated by 15,000 metagenomes that map across all continents and oceans. Also focused on viruses, VIPERdb (51), the database of viral capsids returns after a decade with new features including structure-based sequence alignments that allow identification of significantly conserved positions. Elsewhere metagenomics, and the human microbiome in particular, are a strong focus. Access to human metagenomes will be facilitated by the new curated metadata resource Human-MetagenomeDB (52) while another newcomer GIMICA (53) focuses on human genetic and immune factors that influence the human microbiota. The HMPDACC (54), published here for the first time, takes a multi-omics view of the human microbiome in health and disease, and encompasses 20 different data types. Finally, Planet Microbe (55) offers a home for oceanographic omics datasets, linking them to as much associated environmental context as possible.

In the next section, model organisms feature strongly. Flybase returns (56) to report many new features including Pathway Reports for major signalling pathways, better annotation of enzymes and flagging of proteins whose orthology to characterised human proteins renders them potentially relevant to disease. The Mouse Genome Database (57) also strongly emphasises the relevance to health and disease of comparisons with human proteins, and has a dedicated portal containing mouse research relevant to SARS-CoV-2. The ZFIN database (58) for zebrafish showcases newly designed pages driven by community feedback and describes the exchange of data with the Alliance of Genome Resources (59). Elsewhere two groups of organisms, arguably unfairly neglected hitherto, gain their own dedicated comparative genomics databases. MolluscDB (60) caters to members of the second largest animal phylum and includes not only 20 diverse and high-quality mollusc genomes, but also an array of functional genomic and even paleobiological data. Phycocosm (61) harbours genome (nuclear and plastid) and other omics data for the comparative study of algae. The UCSC Genome Browser (62) is a popular choice for interactive retrieval and display of genomes and associated tracks at its own site or embedded in other databases. Its response to the COVID-19 pandemic included rapidly making the SARS-CoV-2 sequence available with a variety of annotations, but also annotating human genome tracks with SNPs relevant to disease susceptibility and severity. Finally, the section also includes a trio of databases regarding ncRNA expression. The popular deepBase updates with a paper (63) describing a huge increase in expression datasets incorporated plus new and strong foci on ncRNA expression in cancer and exosomes. The new resource LncExpDB (64) intensively covers human lncRNA expression, including subcellular compartments and coexpression with potentially interacting mRNA molecules. Finally, LncSEA (65), also for human lncRNAs, supports research into lncRNA function by deriving reference sets of lncRNAs. Users can submit lncRNA lists for annotation and enrichment analyses.

The section on human genomic variation, diseases and drugs is again the largest in the Issue. Two new databases report on genetic variability in national populations. CSVSN (66) reports on 2000 Spanish genomes and exomes and is notable for the crowd-sourcing from local projects behind its data, while IndiGenomes (67) covers 1000 Indian genomes from the country’s notably diverse population. The GVM (68) collects genetic variation information from 41 species and contributes an update reporting a doubling in size and an analysis of thousands of SARS-CoV-2 variants. The new VARAdb database (69) comprehensively annotates human variants with a welcome emphasis on non-coding changes. Other databases link SNPs to specific molecular phenomena. Thus, RM-Disease (70) and RMVar (71) each link variant to RNA modifications and consider their potential impacts on disease. miRNASNP plays a similar role for SNPs that affect miRNAs or their targets and this update (72) reports a dramatic expansion and a strong focus on disease-related variants. Linking genotypes to phenotypes is the focus of two new databases PhelLiGe (73) and Open Target Genetics (74). The latter’s sister resource, the Open Targets Platform (75) for linking drug targets to diseases, contributes an update describing new scoring of drug targets from an increased range of contributing datasets. These new sources include Open Targets Genetics and Project Score, which includes results from CRISPR knockout screens of cancer models and which also features in this Issue (76). Other cancer databases include two new resources specifically focusing on cancer drivers, CNCdatabase (77) for non-coding cancer drivers and Oncovar (78) which complements experimental data with bioinformatically predicted drivers. canSAR, the multi-faceted oncology database reports new data, interface and query options (79). The paper guides the reader through the database, from the notably clear target synopsis pages to further information that would support a researcher in validating a target in a particular cancer. Among resources relevant to drug candidates, PubChem (80) is the major returning database, reporting over 100 new data sources enabling, for example, better links to literature, patents, material properties and toxicological data. Finally, given the explosion of interest in PROTACs for the targeted degradation of disease-related proteins, it’s worth mentioning PROTAC-DB (81) which collects information on PROTACs’ structures, biological activities and drug-like properties.

Crops are, as usual, a major focus in the plant database section. SoyBase (82) returns with an update including new features for visualisation of gene expression and soya pan-genomes. The family is covered more broadly at LegumeIP (83) which now includes 17 legume genomes. A key focus is enabling translation of knowledge on agriculturally desirable traits from model genomes such as soybean to other species. A new database, Plant-ImputeDB (84), will also support crop breeding efforts by providing reference panels for 12 commercially important crops, thereby enabling better interpretation of genome variation. Two returning databases facilitate comparative plant genomics work. GreenPhyDB (85) adopts the concept of the pan-genome in order to facilitate exploration of evolutionary scenarios within and between species. Gramene contributes a comprehensive update (86) covering expanded content across genomes, pathways and gene expression. It concludes with an interesting view of the number of collaborations of the
enterprise with other databases, including the Alliance of Genome Resources (59). A multi-omics perspective of the model plant Arabidopsis is offered by the new resource AtMAD (87) which integrates genotypes, transcriptomes, methylomes and phenotypes for 620 accessions.

As ever, the final section contains a fascinating variety of databases that do not comfortably sit elsewhere. A single paper covers both ArrayExpress and BioStudies (88) and describes the migration of content from the former, that long outgrew its genesis (89) as a repository for microarray data, in favour of the more flexible and general possibilities of the latter. Glycans are covered by the returning GlyTouCan database (90) for glycan structures which brings improved submission and validation, and by the new GlycoPOST (91) which focuses on mass spectrometry of glycans and glycoproteins. Elsewhere NASA GeneLab (92) publishes in NAR for the first time and covers omics data obtained in space or simulated space conditions. Finally, MitoCarta, the popular focused resource for the mammalian mitochondrial proteome updates with curated sub-organelle localization and assignment of proteins to a set of 149 ‘MitoPathways’ (93).

NAR ONLINE MOLECULAR BIOLOGY DATABASE COLLECTION

The ongoing COVID-19 crisis has shown the resilience of the scientific community: resources were rapidly reallocated as a response to the pandemic and enabled sequencing thousands of strains, tracking infection rates as the virus spread across the globe, structural biology etc. Making the most of the data generated through various streams naturally involves databases and their efforts are somehow reflected in our database with the seven new COVID-specific databases mentioned above but this is but a small fraction of the international effort fueling the number of publications on SARS-CoV-2 and coronavirus-related research (94).

The new normal reduced our travel schedules providing additional time for a major revamp of our NAR online Molecular Database Collection (accessible at https://www.oxfordjournals.org/nar/database/c/). In addition to the customary removal of obsolete databases, hundreds of entries were updated, corrected or expanded bringing the total collection to 1641 databases. We thank the authors for their support for our ongoing effort monitoring the listed resources. Among the hundreds of entries updated, many were due to direct communication with xose.m.fernandez@gmail.com providing a plain text file as defined in https://www.oxfordjournals.org/nar/database/summary/1.

ACKNOWLEDGEMENTS

We thank Dr Martine Bernardes-Silva, especially, and the rest of the Oxford University Press team led by Joanna Venikos for their help in compiling this Issue.

FUNDING

Funding for open access charge: Oxford University Press. Conflict of interest statement. The authors’ opinions do not necessarily reflect the views of their respective institutions.

REFERENCES

1. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. et al. (2020) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., doi:10.1093/nar/gkaa892.
2. Cantelli,G., Cochrane,G., Brooksbank,C., McDonagh,E., Flicek,P., McEntyre,J., Birney,E. and Apweiler,R. (2020) The European Bioinformatics Institute: empowering cooperation in response to a global health crisis. Nucleic Acids Res., doi:10.1093/nar/gkaa1077.
3. CNCB-NGD Members and Partners (2020) Database Resources of the National Genomics Data Center, China National Center for Bioinformatics in 2021. Nucleic Acids Res., doi:10.1093/nar/gkaa1022.
4. Chen,Q., Allott,A. and Lu,Z. (2020) LitCovid: an open database of COVID-19 literature. Nucleic Acids Res., doi:10.1093/nar/gkaa952.
5. Canakoglu,A., Pinioli,P., Bernasconi,A., Alfonsi,T., Melidis,D.P. and Ceri,S. (2020) VirusSurf: an integrated database to investigate viral sequences. Nucleic Acids Res., doi:10.1093/nar/gkaa846.
6. Fang,S., Li,K., Shen,J., Liu,S., Liu,J., Yang,L., Hu,C.-D. and Wan,J. (2020) GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. Nucleic Acids Res., doi:10.1093/nar/gkaa808.
7. Gowthaman,R., Guest,J.D., Yin,R., Adolf-Bryfogle,J., Schief,W.R. and Pierce,B.G. (2020) Co3VD: a database of high resolution coronavirus protein structures. Nucleic Acids Res., doi:10.1093/nar/gkaa731.
8. Tworowski,D., Gorolohvski,A., Mukherjee,S., Carmi,G., Levy,E., Detoroja,R., Mukherjee,S.B. and Frenkel-Morgenstern,M. (2020) COVID19 Drug Repository: text-mining the literature in search of putative COVID19 therapeutics. Nucleic Acids Res., doi:10.1093/nar/gkaa969.
9. Chen,T.-F, Chang,Y.-C., Hsiao,Y., Lee,K.-H., Hsiao,Y.-C., Lin,Y.-H., Tu,Y.-C.E., Huang,H.-C., Chen,C.-Y. and Juan,H.-F. (2020) DockCoV2: a drug database against SARS-CoV-2. Nucleic Acids Res., doi:10.1093/nar/gkaa861.
10. Yue,Z., Zhang/E., Xu,C., Khurana,S., Batra,N., Dang,S.D.H., Cinimo,J.J. and Chen,J.Y. (2020) PAGER-CoV: a comprehensive collection of pathways, annotated gene-lists and gene signatures for coronavirus disease studies. Nucleic Acids Res., doi:10.1093/nar/gkaa1046.
11. Zhao,L., Wang,J., Li,Y., Song,T., Wu,Y., Fang,S., Bu,D., Li,H., Sun,L., Pei,D. et al. (2020) NONCODEv6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res., doi:10.1093/nar/gkaa1046.
12. Ning,L., Cui,T., Zheng,B., Wang,N., Luo,J., Yang,B., Du,M., Cheng,J., Dou,Y. and Wang,D. (2020) MNDR v3.0: mammal nRNA–disease repository with increased coverage and annotation. Nucleic Acids Res., doi:10.1093/nar/gkaa1047.
13. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Gautheret,D., Weinberg,Z., Rivas,E. and Griffiths-Jones,S. (2020) NONCODEv6: an updated database of non-coding RNA–disease annotation and interaction. Nucleic Acids Res., doi:10.1093/nar/gkaa707.
14. RNACentral Consortium. (2020) RNACentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res., doi:10.1093/nar/gkaa921.
15. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRNA from microRNA sequences to function. Nucleic Acids Res., 47, D155-D162.
16. Marchand,J.A., Smela,M.D.P., Jordan,T.H.H., Narasimhan,K. and Church,G.M. (2020) TBDB: a database of structurally annotated T-box riboswitch:tRNA pairs. Nucleic Acids Res., doi:10.1093/nar/gkaa721.
17. Li,P., Zhou,X., Xu,K. and Zhang,Q.C. (2020) RASP: an atlas of transcriptome-wide RNA secondary structure probing data. Nucleic Acids Res., doi:10.1093/nar/gkaa880.
18. Contessoto,V.G., Cheng,R.R., Hajitohari,A., Dodero-Rojas,E., Mello,M.F., Lieberman-Aiden,E., Wolynes,P.G., Di Pierro,M. and Onuchic,J.N. (2020) The Nucleome Data Bank: web-based resources to simulate and analyze the three-dimensional genome. Nucleic Acids Res., doi:10.1093/nar/gkaa818.
19. Lin,Y.-H., Tu,Y.-C.E., Huang,H.-C., Chen,C.-Y. and Juan,H.-F. (2020) DockCoV2: a drug database against SARS-CoV-2. Nucleic Acids Res., doi:10.1093/nar/gkaa1047.
20. T-box riboswitch:tRNA pairs.
55. Ponseno A.J., Bomhoff, M., Blumberg, K., Youens-Clark, K., Herz, N.M., Wood-Charlson, E.M., Delong, E.F. and Hurwitz, B.L. (2020) Planet microphone: a platform for marine microbiology to discover and analyze interconnected 'omics and environmental data. Nucleic Acids Res. doi:10.1093/nar/gkaa637.

56. Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B. et al. (2020) FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. doi:10.1093/nar/gkaa1026.

57. Blake, J.A., Baldarelli, R., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J. and the Mouse Genome Database Group (2020) Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. Nucleic Acids Res. doi:10.1093/nar/gkaa1083.

58. Howe, D.G., Ramachandran, S., Bradford, Y.M., Fashena, D., Toso, S., Eagle, A., Frazer, K., Kalita, P., Mani, P., Martin, R. et al. (2020) The Zebrafish Information Network: major gene page and home page updates. Nucleic Acids Res. doi:10.1093/nar/gkaa1010.

59. The Alliance of Genome Resources Consortium. (2020) Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res., 48, D650–D658.

60. Liu, F., Li, Y., Hu, H., Zhang, L., Hu, J., Bao, Z. and Wang, S. (2020) MolluscaDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. Nucleic Acids Res. doi:10.1093/nar/gkaa918.

61. Grigoriev, I.V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Sergey, D., Nikitin, R., Mondo, S.J., Salamova, A. et al. (2020) PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. doi:10.1093/nar/gkaa898.

62. Gonzalez, J.N., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M. et al. (2020) The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. doi:10.1093/nar/gkaa1070.

63. Xie, F., Liu, S., Wang, J., Xuan, J., Zhang, X., Qu, L., Zheng, L. and Yang, J. (2020) DeepBase v3.0: expression atlas and interactive analysis of ncRNAs from thousands of deep-sequencing data. Nucleic Acids Res. doi:10.1093/nar/gkaa1039.

64. Li, Z., Liu, L., Jiang, S., Li, Q., Fung, C., Du, Q., Zou, D., Xiao, J., Zhang, Z. and Ma, L. (2020) LncExpDB: an expression database of human long non-coding RNAs. Nucleic Acids Res. doi:10.1093/nar/gkaa850.

65. Chen, J., Zhang, J., Gao, Y., Li, Y., Fung, C., Song, C., Ning, Z., Zhou, X., Zhao, J., Feng, M. et al. (2020) LncSEA: a platform for long non-coding RNA related sets and enrichment analysis. Nucleic Acids Res. doi:10.1093/nar/gkaa806.

66. Peña-Chilet, M., Boldt, G., Perez-Florido, J., Peña-Chilet, M., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B. et al. (2020) NucleusmA: a comprehensive database of the Spanish population genetic variability. Nucleic Acids Res. doi:10.1093/nar/gkaa894.

67. Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B. et al. (2020) FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. doi:10.1093/nar/gkaa1026.

68. Howe, D.G., Ramachandran, S., Bradford, Y.M., Fashena, D., Toso, S., Eagle, A., Frazer, K., Kalita, P., Mani, P., Martin, R. et al. (2020) The Zebrafish Information Network: major gene page and home page updates. Nucleic Acids Res. doi:10.1093/nar/gkaa1010.

69. The Alliance of Genome Resources Consortium. (2020) Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res., 48, D650–D658.

70. Liu, F., Li, Y., Hu, H., Zhang, L., Hu, J., Bao, Z. and Wang, S. (2020) MolluscaDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. Nucleic Acids Res. doi:10.1093/nar/gkaa918.

71. Grigoriev, I.V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Sergey, D., Nikitin, R., Mondo, S.J., Salamova, A. et al. (2020) PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. doi:10.1093/nar/gkaa898.

72. Gonzalez, J.N., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M. et al. (2020) The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. doi:10.1093/nar/gkaa1070.

73. Xie, F., Liu, S., Wang, J., Xuan, J., Zhang, X., Qu, L., Zheng, L. and Yang, J. (2020) DeepBase v3.0: expression atlas and interactive analysis of ncRNAs from thousands of deep-sequencing data. Nucleic Acids Res. doi:10.1093/nar/gkaa1039.

74. Li, Z., Liu, L., Jiang, S., Li, Q., Fung, C., Du, Q., Zou, D., Xiao, J., Zhang, Z. and Ma, L. (2020) LncExpDB: an expression database of human long non-coding RNAs. Nucleic Acids Res. doi:10.1093/nar/gkaa850.

75. Chen, J., Zhang, J., Gao, Y., Li, Y., Fung, C., Song, C., Ning, Z., Zhou, X., Zhao, J., Feng, M. et al. (2020) LncSEA: a platform for long non-coding RNA related sets and enrichment analysis. Nucleic Acids Res. doi:10.1093/nar/gkaa806.

76. Peña-Chilet, M., Boldt, G., Perez-Florido, J., Peña-Chilet, M., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B. et al. (2020) NucleusmA: a comprehensive database of the Spanish population genetic variability. Nucleic Acids Res. doi:10.1093/nar/gkaa894.

77. Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B. et al. (2020) FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res. doi:10.1093/nar/gkaa1026.
international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa947.

91. Watanabe, Y., Aoki-Kinoshita, K. F., Ishihama, Y. and Okuda, S. (2020) GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1012.

92. Berrios, D. C., Galazka, J., Grigorev, K., Gebre, S. and Costes, S. V. (2020) NASA GeneLab: interfaces for the exploration of space omics data. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa887.

93. Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T. J., Goodman, R. P., Grabarck, Z., Haas, M. E., Hung, W. H. W. *et al.* (2020) MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1011.

94. Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W. *et al.* (2020) CORD-19: the Covid-19 Open Research Dataset. arXiv doi: https://arxiv.org/abs/2004.10706v4, 22 April 2020, preprint: not peer reviewed.