Nucleic Acids Research, 2021, Vol. 49, Database issue ***D1–D9***
doi: 10.1093/nar/gkaa1216

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

Daniel J. Rigden[1,*] and Xosé M. Fernández[2]

[1]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [2]Institut 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 GreenPhylDB. 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 curation 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-

*To whom correspondence should be addressed. Tel: +44 151 795 4467; Email: nardatabase@gmail.com

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

velopment 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 lncRNAs from 23 species, including their gene expression, function annotation and sequence conservation between species. NONCODE also reports new efforts to document associations between mammalian lncRNAs 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. Rfam (13) reaches version 14 with new RNA families, some community-derived (including from masters students the paper reports) using a new Rfam 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:tRNA 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 Nucleome 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 archaebacteria. 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 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_lab.hzau.edu.cn/Animal-APAdb/	Alternative polyadenylation in animals
AcrDB	http://bcb.unl.edu/AcrDB/	Anti-CRISPR operons
AcrHub	http://pacrispr.erc.monash.edu/AcrHub	Anti-CRISPR proteins
ATACdb	http://www.licpathway.net/ATACdb	Human Assay-for-Transposase-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	https://bigfam.bioinformatics.nl	Biosynthetic gene cluster families
CancerImmunityQTL	http://www.cancerimmunityqtl-hust.com/	ImmunQTLs across multiple cancer types
Chewie-NS	https://chewbbaca.online/	Gene-by-gene schemas for microbial strain identification
CMNPD	https://www.cmnpd.org/	Comprehensive Marine Natural Product Database
CNCDatabase	https://cncdatabase.med.cornell.edu/	Cancer drivers at non-coding regions
cncRNAdb	http://www.rna-society.org/cncrnadb/	Coding and non-coding RNA
ConjuPepDB	http://conjupepdb.ttk.hu/	Drug-peptide conjugates
CovInDB	http://cadd.zju.edu.cn/cidb/	Covalent Inhibitor DataBase
crisprSQL	http://www.crisprsql.com/	CRISPR/Cas9 Off-Target Cleavage Assays
CRISPR-view	http://crispviewdatabase.weililab.org/	Functional genetic screens
CSEA-DB	https://bioinfo.uth.edu/CSEADB/	Cell type specificity of genetic traits
CSVS	http://csvs.babelomics.org/	Spanish genomes and exomes
Cyanorak	http://www.sb-roscoff.fr/cyanorak	Comparative genomics of cyanobacteria
Datanator	https://datanator.info/	Molecular data for modeling biochemical networks
dbCAN-PUL	http://bcb.unl.edu/dbCAN_PUL/home	Experimentally characterized CAZyme gene clusters
dbGuide	https://sgrnascorer.cancer.gov/dbguide	Manually curated and functionally validated guide RNAs
DescribePROT	http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/	Residue level prediction of structure and function across proteomes
DKK	https://darkkinome.org/	Dark Kinome Knowledgebase
DIGGER	https://exbio.wzw.tum.de/digger	Alternative splicing and protein-protein interactions
DNAmoreDB	http://www.genesilico.pl/DNAmoreDB	DNAzymes, i.e. DNA molecules with catalytic activity
DrugSpaceX	https://drugspacex.simm.ac.cn/	100 million compounds for virtual screening
DualSeqDB	http://www.tartaglialab.com/dualseq/	Dual RNA-seq host-pathogen sequencing
FireProtDB	https://loschmidt.chemi.muni.cz/fireprotdb	Protein stability data
gcType	http://gctype.wdcm.org/	WDCM 10K sequencing projects
GIMICA	https://idrblab.org/gimica/	Host Genetic and Immune Factors Shaping Human Microbiota
GlycoPOST	https://glycopost.glycosmos.org/	Raw Mass Spectrometry glycomics data
GRNdb	http://www.grndb.com/	TF-target relationships inferred from single cell and bulk RNA-seq datasets
GSRS	https://gsrs.ncats.nih.gov/app/substances	Global Substance Registration System
HeRA	https://hanlab.uth.edu/HeRA/	Human enhancer RNA Atlas
HERB	http://herb.ac.cn/	High-throughput Experiment- and Reference-guided dataBase of TCM
Housekeeping Transcript Atlas	http://www.housekeeping.unicamp.br/	Human and mouse housekeeping genes
HumanMetagenomeDB	https://webapp.ufz.de/hmgdb/	Curated and standardized metadata for human metagenomes
iCSDB	http://www.kobic.re.kr/icsdb	integrated CRISPR Screens DataBase
IDDB	http://mdl.shsmu.edu.cn/IDDB	Infertility Disease DataBase
iModulonDB	https://imodulondb.org/	'iModulons', groups of independently-modulated genes
IndiGenomes	http://clingen.igib.res.in/indigen/	Genetic variation in 1000 Indian individuals
INTEDE	https://idrblab.org/intede/	Interactome of Drug-Metabolizing Enzymes
KinaseMD	https://bioinfo.uth.edu/kmd/	Kinase Mutations and Drug responses
LnCeCell	http://bio-bigdata.hrbmu.edu.cn/LnCeCell/	lncRNA-associated ceRNA networks
LncExpDB	https://bigd.big.ac.cn/lncexpdb	Human lncRNA expression database
LncSEA	http://bio.liclab.net/LncSEA/	Reference human lncRNA sets
m6A-Atlas	http://www.xjtlu.edu.cn/biologicalsciences/atlas	The N6-methyladenosine (m6A) epitranscriptome including base-resolution data
markerDB	http://www.markerdb.ca/	Biomarkers: chemical, protein, chromosomal and genetic
MASI	http://www.aiddlab.com/MASI/	Microbiota – Active Substance Interactions database
MeDAS	https://das.chenlulab.com	Alternative splicing during development in 20 species
MemMoRF	http://memmorf.hegelab.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://oncovar.org/	Driver mutations, genes and pathways in cancer
Open Targets Genetics	https://genetics.opentargets.org/	Drug targets prioritised from genetic data
PCAT	http://pedtranscriptome.org	Pediatric cancer transcriptome explorer
Peryton	https://dianalab.e-ce.uth.gr/peryton	Microbe–disease associations
PheLiGe	https://phelige.com/	Genotype–phenotype associations
PhycoCosm	http://phycocosm.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_lab.hzau.edu.cn/Plant_imputeDB	Reference panels and imputation methods for plant genomes

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	https://rbp2go.dkfz.de	RNA-binding proteins across species
RJunBase	www.RJunBase.org	RNA splice junctions
RMDisease	http://www.xjtlu.edu.cn/biologicalsciences/rmd	Genetic variants that affect RNA modifications vs disease
RMVar	http://rmvar.renlab.org	Genetic variants that affect RNA modifications vs disease
SC2disease	http://easybioai.com/sc2disease/	Single cell transcriptomics data and disease
SilencerDB	http://health.tsinghua.edu.cn/silencerdb or http://bioinfo.au.tsinghua.edu.cn/silencerdb/	Human silencers, validated or predicted
STAB	http://stab.comp-sysbio.org/	Spatio-Temporal Cell Atlas of the Human Brain
TBDB	https://tbdb.io	Structurally annotated T-box riboswitch: tRNA pairs
TCRdb	http://bioinfo.life.hust.edu.cn/TCRdb	T-cell receptor (TCR) sequences
ThermoMutDB	http://biosig.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://trftarget.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://genelab.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
PLncDB	http://plncdb.tobaccodb.org/	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

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 improved treatments of viruses and new visualisations, a more dynamic and interactive pathway map viewer and global map display with more flexible and informative colouring. A number of papers cover metabolic pathways. WikiPathways (37) reports increased numbers of pathways, metabolites and—importantly, given its modus operandi—contributors. Distinct portals with their own URLs support communities focused on topics as diverse as COVID-19, rare diseases and lipid metabolism. The popular ModelSEED Biochemistry database (38), publishing here for the first time, reports a doubling of content since its first iteration. Quantitative pathway modelling requires easy access to relevant data. The new database Datanator (39) steps in here to collect highly diverse data from a wide range of relevant resources and present it in a normalised and integrated fashion. Kinases are particularly well covered this year with KLIFS, the database for kinase–inhibitor interactions, reporting an update (40) that extends coverage to atypical kinases and provides an API for easy access. It is joined by two new resources, DKK (41) focussing on the 'Dark Kinome' and collecting information on substrates, inhibitor and tissue expression; and KinaseMD (42) which provides 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

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 HumanMetagenomeDB (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. CSVS (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, RMDisease (70) and RMVar (71) each link genetic variants 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 PheLiGe (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 pangenomes. 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. GreenPhylDB (85) adopts the concept of the pangenome 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 Ventikos 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-NGDC Members and Partners (2020) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1022.
4. Chen,Q., Allot,A. and Lu,Z. (2020) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa952.
5. Canakoglu,A., Pinoli,P., Bernasconi,A., Alfonsi,T, Melidis,D.P. and Ceri,S. (2020) ViruSurf: 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) CoV3D: a database of high resolution coronavirus protein structures. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa731.
8. Tworowski,D., Gorohovski,A., Mukherjee,S., Carmi,G., Levy,E., Detroja,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., Cimino,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/gkaa1094.
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 ncRNA–disease repository with increased coverage and annotation. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa707.
13. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z., Rivas,E. *et al.* (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1047.
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) miRBase: 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., Hajitaheri,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. Kim,K., Jang,I., Kim,M., Choi,J., Kim,M.-S., Lee,B. and Jung,I. (2020) 3DIV update for 2021: a comprehensive resource of 3D

genome and 3D cancer genome. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1078.
20. Ponce-Salvatierra,A., Boccaletto,P. and Bujnicki,J.M. (2020) DNAmoreDB, a database of DNAzymes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa867.
21. Galperin,M.Y., Wolf,Y.I., Makarova,K.S., Alvarez,R.V., Landsman,D. and Koonin,E.V. (2020) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1018.
22. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2020) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa913.
23. Paladin,L., Bevilacqua,M., Errigo,S., Piovesan,D., Mičetić,I., Necci,M., Monzon,A.M., Fabre,M.L., Lopez,J.L., Nilsson,J.F. *et al.* (2020) RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1097.
24. Letunic,I., Khedkar,S. and Bork,P. (2020) SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa937.
25. Mi,H., Ebert,D., Muruganujan,A., Mills,C., Albou,L.-P., Mushayamaha,T. and Thomas,P.D. (2020) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1106.
26. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2020) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1038.
27. Protein Data Bank. (1971) *Nat. New Biol*, **233**, 223.
28. Nikam,R., Kulandaisamy,A., Harini,K., Sharma,D. and Gromiha,M.M. (2020) ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1035.
29. Xavier,J.S., Nguyen,T.-B., Karmarkar,M., Portelli,S., Rezende,P.M., Velloso,J.P.L., Ascher,D.B. and Pires,D.E.V. (2020) ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa925.
30. Stourac,J., Dubrava,J., Musil,M., Horackova,J., Damborsky,J., Mazurenko,S. and Bednar,D. (2020) FireProt^{DB}: database of manually curated protein stability data. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa981.
31. Piovesan,D, Necci,M., Escobedo,N., Monzon,A.M., Hatos,A., Mičetić,I., Quaglia,F., Paladin,L., Ramasamy,P., Dosztányi,Z. *et al.* (2020) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1058.
32. Lazar,T., Martínez-Pérez,E., Quaglia,F., Hatos,A., Chemes,L.B., Iserte,J.A., Méndez,N.A., Garrone,N.A., Saldaño,T.E., Marchetti,J. *et al.* (2020) PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1021.
33. Csizmadia,G., Erdős,G., Tordai,H., Padányi,R., Tosatto,S., Dosztányi,Z. and Hegedűs,T. (2020) The MemMoRF database for recognizing disordered protein regions interacting with cellular membranes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa954.
34. The Gene Ontology Consortium. (2020) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1113.
35. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T. and Bork,P. (2020) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1074.
36. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2020) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa970.
37. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., Miller,R.A., Digles,D., Lopes,E.N., Ehrhart,F. *et al.* (2020) WikiPathways: connecting communities. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1024.
38. Seaver,S.M.D., Liu,F., Zhang,Q., Jeffryes,J., Faria,J.P., Edirisinghe,J.N., Mundy,M., Chia,N., Noor,E., Beber,M.E. *et al.* (2020) The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa746.
39. Roth,Y.D., Lian,Z., Pochiraju,S., Shaikh,B. and Karr,J.R. (2020) Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1008.
40. Kanev,G.K., de Graaf,C., Westerman,B.A., de Esch,I.J.P. and Kooistra,A.J. (2020) KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa895.
41. Berginski,ME., Moret,N., Liu,C., Goldfarb,D., Sorger,P.K. and Gomez,S.M. (2020) The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa853.
42. Hu,R., Xu,H., Jia,P. and Zhao,Z. (2020) KinaseMD: kinase mutations and drug response database. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa945.
43. Kautsar,S.A., Blin,K., Shaw,S., Weber,T. and Medema,M.H. (2020) BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa812.
44. Blin,K., Pascal Andreu,V., de Los Santos,E.L.C., Del Carratore,F., Lee,S.Y., Medema,M.H. and Weber,T. (2019) The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **47**, D625–D630.
45. Ausland,C., Zheng,J., Yi,H., Yang,B., Li,T., Feng,X., Zheng,B. and Yin,Y. (2020) dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa742.
46. Huang,L., Zhang,H., Wu,P., Entwistle,S., Li,X., Yohe,T., Yi,H., Yang,Z. and Yin,Y. (2018) dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res.*, **46**, D516–D521.
47. Wang,J., Dai,W., Li,J., Li,Q., Xie,R., Zhang,Y., Stubenrauch,C. and Lithgow,T. (2020) AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa951.
48. Huang,L., Yang,B., Yi,H., Asif,A., Wang,J., Lithgow,T., Zhang,H., Minhas,F.U.A.A. and Yin,Y. (2020) AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa857.
49. Chen,I.-M.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S., Varghese,N. and Seshadri,R. (2020) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa939.
50. Roux,S., Páez-Espino,D., Chen,I.-M.A., Palaniappan,K., Ratner,A., Chu,K., Reddy,T.B.K., Nayfach,S., Schulz,F., Call,L. *et al.* (2020) IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa946.
51. Montiel-Garcia,D., Santoyo-Rivera,N., Ho,P., Carrillo-Tripp,M., Brooks,C.L. III, Johnson,J.E. and Reddy,V.S. (2020) VIPERdb v3.0: a structure-based data analytics platform for viral capsids. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1096.
52. Kasmanas,J.C., Bartholomäus,A., Corrêa,F.B., Tal,T., Jehmlich,N., Herberth,G., von Bergen,M., Stadler,P.F., de Leon Ferreira de Carvalho,A.C.P. and da Rocha,U.N. (2020) HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1031.
53. Tang,J., Wu,X., Mou,M., Wang,C., Wang,L., Li,F., Guo,M., Yin,J., Xie,W., Wang,X. *et al.* (2020) GIMICA: host genetic and immune factors shaping human microbiota. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa851.
54. Creasy,H.H., Felix,V., Aluvathingal,J., Crabtree,J., Ifeonu,O., Matsumura,J., McCracken,C., Nickel,L., Orvis,J., Schor,M. *et al.* (2020) HMPDACC: a Human Microbiome Project Multi-omic data resource. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa996.

55. Ponsero,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 microbe: 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., Toro,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., Yu,H., Zhang,L., Hu,J., Bao,Z. and Wang,S. (2020) MolluscDB: 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., Salamov,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., Feng,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., Feng,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., Roldán,G., Perez-Florido,J., Ortuño,F.M., Carmona,R., Aquino,V., Lopez-Lopez,D., Loucera,C., Fernandez-Rueda,J.L., Gallego,A. *et al.* (2020) CSVS, a crowdsourcing database of the Spanish population genetic variability. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa794.
67. Jain,A., Bhoyar,R.C., Pandhare,K., Mishra,A., Sharma,D., Imran,M., Senthivel,V., Divakar,M.K., Rophina,M., Jolly,B. *et al.* (2020) IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa923.
68. Li,C., Tian,D., Tang,B., Liu,X., Teng,X., Zhao,W., Zhang,Z. and Song,S. (2020) Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1005.
69. Pan,Q., Liu,Y.-J., Bai,X.-F., Han,X.-L., Jiang,Y., Ai,B., Shi,S.-S., Wang,F., Xu,M.-C., Wang,Y.-Z. *et al.* (2020) VARAdb: a comprehensive variation annotation database for human. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa922.
70. Chen,K., Song,B., Tang,Y., Wei,Z., Xu,Q., Su,J., de Magalhães,J.P., Rigden,D.J. and Meng,J. (2020) RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa790.
71. Luo,X., Li,H., Liang,J., Zhao,Q., Xie,Y., Ren,J. and Zuo,Z. (2020) RMVar: an updated database of functional variants involved in RNA modifications. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa811.
72. Liu,C.-J., Fu,X., Xia,M., Zhang,Q., Gu,Z. and Guo,A.-Y. (2020) miRNASNP-v3: a comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa783.
73. Shashkova,T.I., Pakhomov,E.D., Gorev,D.D., Karssen,L.C., Joshi,P.K. and Aulchenko,Y.S. (2020) PheLiGe: an interactive database of billions of human genotype–phenotype associations. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1086.
74. Ghoussaini,M., Mountjoy,E., Carmona,M., Peat,G., Schmidt,E.M., Hercules,A., Fumis,L., Miranda,A., Carvalho-Silva,D., Buniello,A. *et al.* (2020) Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* doi:10.1093/nar/gkaa840.
75. Ochoa,D., Hercules,A., Carmona,M., Suveges,D., Gonzalez-Uriarte,A., Malangone,C., Miranda,A., Fumis,L., Carvalho-Silva,D., Spitzer,M. *et al.* (2020) Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1027.
76. Dwane,L., Behan,F.M., Gonçalves,E., Lightfoot,H., Yang,W., van derMeer,D., Shepherd,R., Pignatelli,M., Iorio,F. and Garnett,M.J. (2020) Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa882.
77. Liu,E.M., Martinez-Fundichely,A., Bollapragada,R., Spiewack,M. and Khurana,E. (2020) CNCDatabase: a database of non-coding cancer drivers. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa915.
78. Wang,T., Ruan,S., Zhao,X., Shi,X., Teng,H., Zhong,J., You,M., Xia,K., Sun,Z. and Mao,F. (2020) OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1033.
79. Mitsopoulos,C., Di Micco,P., Fernandez,E.V., Dolciami,D., Holt,E., Mica,I.L., Coker,E.A., Tym,J.E., Campbell,J., Che,K.H. *et al.* (2020) canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1059.
80. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2020) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa971.
81. Weng,G., Shen,C., Cao,D., Gao,J., Dong,X., He,Q., Yang,B., Li,D., Wu,J. and Hou,T. (2020) PROTAC-DB: an online database of PROTACs. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa807.
82. Brown,A.V., Conners,S.I., Huang,W., Wilkey,A.P., Grant,D., Weeks,N.T., Cannon,S.B., Graham,M.A. and Nelson,R.T. (2020) A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1107.
83. Dai,X., Zhuang,Z., Boschiero,C., Dong,Y. and Zhao,P.X. (2020) LegumeIP V3: from models to crops—an integrative gene discovery platform for translational genomics in legumes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa976.
84. Gao,Y., Yang,Z., Yang,W., Yang,Y., Gong,J., Yang,Q.-Y. and Niu,X. (2020) Plant-ImputeDB: an integrated multiple plant reference panel database for genotype imputation. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa953.
85. Valentin,G., Abdel,T., Gaëtan,D., Jean-François,D., Matthieu,C. and Mathieu,R. (2020) GreenPhylDB v5: a comparative pangenomic database for plant genomes. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1068.
86. Tello-Ruiz,M.K., Naithani,S., Gupta,P., Olson,A., Wei,S., Preece,J., Jiao,Y., Wang,B., Chougule,K., Garg,P. *et al.* (2020) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa979.
87. Lan,Y., Sun,R., Ouyang,J., Ding,W., Kim,M.-J., Wu,J., Li,Y. and Shi,T. (2020) AtMAD: *Arabidopsis thaliana* multi-omics association database. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1042.
88. Sarkans,U., Füllgrabe,A., Ali,A., Athar,A., Behrangi,E., Diaz,N., Fexova,S., George,N., Iqbal,H., Kurri,S. *et al.* (2020) From ArrayExpress to BioStudies. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1062.
89. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
90. Fujita,A., Aoki,N.P., Shinmachi,D., Matsubara,M., Tsuchiya,S., Shiota,M., Ono,T., Yamada,I. and Aoki-Kinoshita,K.F. (2020) The

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., Grabarek,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.