Published online 27 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D29–D37***
doi: 10.1093/nar/gkaa1077

The European Bioinformatics Institute: empowering cooperation in response to a global health crisis

Gaia Cantelli [1,*], Guy Cochrane [1], Cath Brooksbank[1], Ellen McDonagh [1,2], Paul Flicek [1], Johanna McEntyre[1], Ewan Birney[1] and Rolf Apweiler[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Open Targets, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 24, 2020; Revised October 20, 2020; Editorial Decision October 21, 2020; Accepted October 22, 2020

ABSTRACT

The European Bioinformatics Institute (EMBL-EBI; https://www.ebi.ac.uk/) provides freely available data and bioinformatics services to the scientific community, alongside its research activity and training provision. The 2020 COVID-19 pandemic has brought to the forefront a need for the scientific community to work even more cooperatively to effectively tackle a global health crisis. EMBL-EBI has been able to build on its position to contribute to the fight against COVID-19 in a number of ways. Firstly, EMBL-EBI has used its infrastructure, expertise and network of international collaborations to help build the European COVID-19 Data Platform (https://www.covid19dataportal.org/), which brings together COVID-19 biomolecular data and connects it to researchers, clinicians and public health professionals. By September 2020, the COVID-19 Data Platform has integrated in excess of 170 000 COVID-19 biomolecular data and literature records, collected through a number of EMBL-EBI resources. Secondly, EMBL-EBI has strived to continue its support of the life science communities through the crisis, with updated Training provision and improved service provision throughout its resources. The COVID-19 pandemic has highlighted the importance of EMBL-EBI's core principles, including international cooperation, resource sharing and central data brokering, and has further empowered scientific cooperation.

INTRODUCTION

The European Bioinformatics Institute (EMBL-EBI) is part of the European Molecular Biology Laboratory, an intergovernmental research organisation focusing on advancing the study and understanding of molecular biology across Europe. EMBL-EBI provides freely available data and bioinformatics services to the scientific community, contributing to the advancement of biology through investigator-driven research and supporting scientists at all levels with advanced bioinformatics training. We host a suite of open data resources and tools, which cover every data type in molecular biology, including, among others, nucleotide sequence data, protein sequences, chemical biology and the scientific literature (Figure 1). Resources hosted by EMBL-EBI include deposition databases (which archive experimental data) and added-value databases (which add value to archived data by providing annotation, curation, reanalysis and integration), as well as open-source software tools. EMBL-EBI resources are central to the life sciences community. On an average day in 2019, EMBL-EBI resources received over 62 million requests. Throughout 2019, EMBL-EBI websites were visited by almost 24 million IP addresses (the number of IP addresses is an indication of the number of users, but not an exact count), and EMBL-EBI trained over 30 000 people through its training programmes. Moreover, in 2019 EMBL-EBI reached over 300 petabytes of raw data storage space.

EMBL-EBI is committed to keeping the data it manages open, freely accessible and FAIR (findable, accessible, interoperable and reusable (1)). Building on its 25 years of international collaboration in bioinformatics and data brokering, EMBL-EBI is working with the international scientific community to build new scientific partnerships and coordinate global research efforts. This involves, for example, active participation in the development of the Global Alliance for Genomics and Health (https://www.ga4gh.org/), leading in the creation of standardised methods for accessing large-scale genomic data, and the use of the Data Use Ontology (2) by the European Genome-phenome Archive (EGA) to ensure that data use conditions are machine readable and standardised. Moreover, EMBL-EBI is a key player in the ARGENT project, an international effort that aims to create a global, live database of *Mycobacterium tuberculosis* genomes to monitor tuberculosis outbreaks around the world. Through these efforts, EMBL-EBI is developing new

[*]To whom correspondence should be addressed. Tel: +44 1223 49 4588; Email: gaiacantelli@ebi.ac.uk

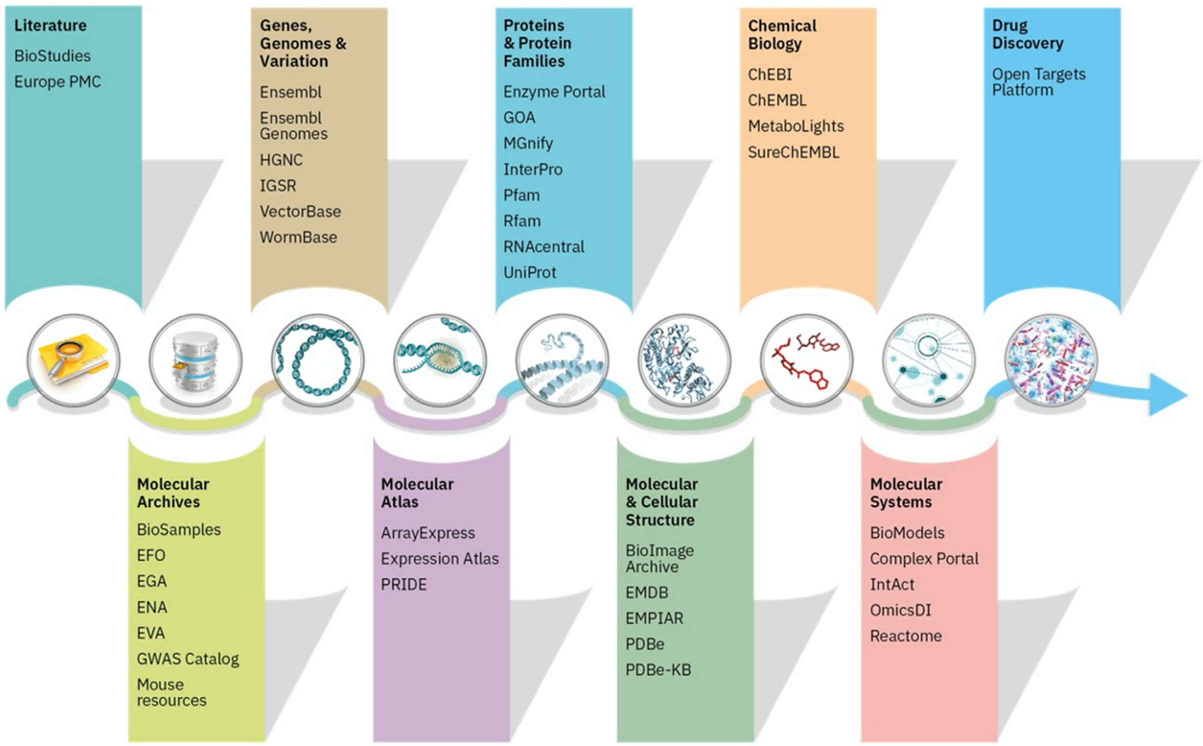


Figure 1. EMBL-EBI Resources. Summary of all EMBL-EBI Data Resources as of September 2020.

ways for researchers, clinicians and public health professionals to share data, gather new insights and build on each other's discoveries.

The 2020 COVID-19 pandemic has re-shaped the landscape of research and scientific cooperation. The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) causes the disease COVID-19: between January and September 2020, SARS-CoV-2 has been responsible for over 890 000 deaths. Moreover, the pandemic and the lockdown measures implemented by governments across the globe to contain it have caused unprecedented economic and societal disruption. The pandemic showed that the scientific community needed effective international collaboration, backed by trusted expertise and an open dialogue with the public. The position EMBL-EBI holds as an international data broker, a service provider and home to cutting-edge research has allowed the Institute to contribute to the fight against COVID-19 in a number of ways. Crucially, the contributions of the Institute hinge not solely on coronavirus research, but on empowering international cooperation across research fields, allowing researchers to access trustworthy information and to cooperate in the face of a global crisis.

DATA SERVICE RESPONSES TO THE PANDEMIC

In the early months of 2020, the future global impact of the newly emerged SARS-CoV-2 became apparent. From March 2020, EMBL-EBI, supported by the European Commission and other funders, set to work on building a foundation for the scientific response to COVID-19 centred on biomolecular data. On the 20th of April 2020 the **European COVID-19 Data Platform** (https://www.covid19dataportal.org/) was launched, as announced by the President of the European Commission (https://audiovisual.ec.europa.eu/en/video/I-189639). The platform enables researchers to upload, access and analyse COVID-19 related reference data and specialist datasets as part of the global research effort. Drawing on our infrastructure, membership of ELIXIR and involvement in the European Open Science Cloud, we developed resources to link the clinical, epidemiological, and public health worlds. The European COVID-19 Data Platform comprises three components: the COVID-19 Data Portal, the SARS-CoV-2 Data Hub and the Federated EGA (FEGA) (Figure 2).

The COVID-19 Data Portal

The COVID-19 Data Portal (https://www.covid19dataportal.org, Figure 3) brings together biomolecular data across the diversity of relevant data types to provide an integrated search (based on EMBL-EBI's central data search service (3)), navigation across data and data download. In addition, the Portal provides an entry point to data resources and services beyond the biomolecular domain, including clinical and epidemiology data.

In September 2020, the system presents data from 14 of EMBL-EBI's biomolecular data resources, including UniProt (4), European Nucleotide Archive (5), Electron Microscopy Data Bank (5), Protein Data Bank in Europe (6), Europe PMC (7), PRIDE (8), InterPro (9), ChEMBL (10), Complex Portal (11), Protein Data Bank in Europe-Knowledge Base (12), Ensembl (13), Reactome (14) IntAct

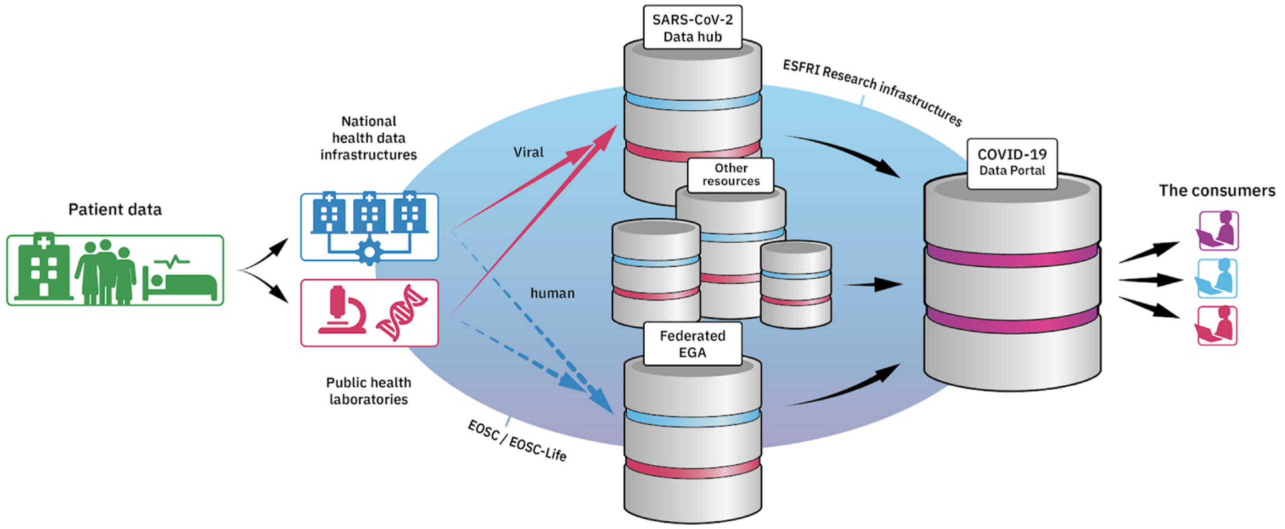


Figure 2. Data flow through the components of the European COVID-19 Data Platform. Patient data is collected by national health data infrastructures and public health laboratories, which can be deposited in specific SARS-CoV-2 Data Hubs (numbering 12 at the time of writing), Federated EGA (FEGA) and other data resources. This data feed into the COVID-19 Data Portal, which can be accessed by its user community of researchers, practitioners and public health specialists.

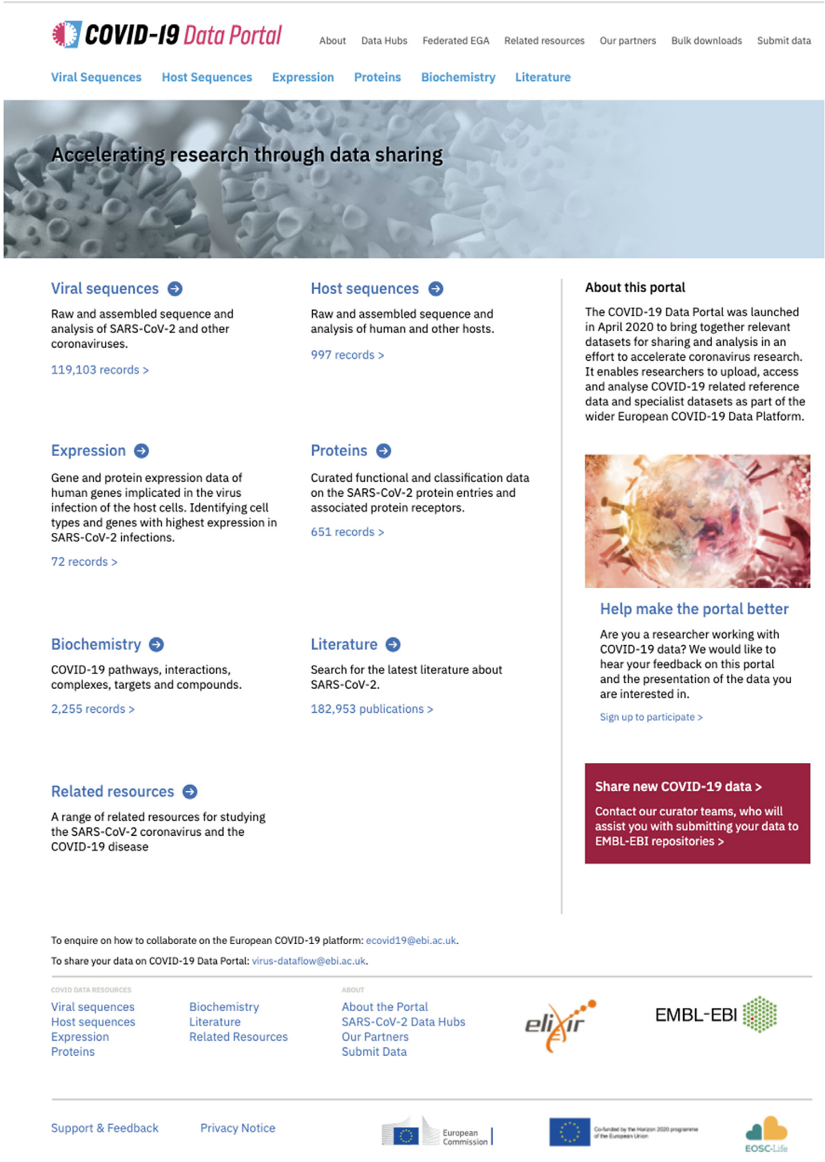
(15) and the European Genome-phenome Archive (16), featuring many of ELIXIR's Core Data Resources (17). Data in the system is growing rapidly and spans viral and host sequences, gene and protein expression, protein functions and structures, biochemical pathways and compounds, and the scientific literature (Box 1).

Data content available from the COVID-19 Data Portal (September 2020).

- **9,777** viral sequences
- **30,122** raw viral sequence data sets
- **2,332** variants
- **171** protein records
- **314** structures
- **8** compound screens
- **24** targets
- **376** pathways
- **27** expression studies covering implicated human genes
- **10** protein expression studies
- **110,379** scientific literature publications

To integrate and make this data accessible to the user community and support research into COVID-19, several of the contributing data resources developed specific COVID-19 browsers and tools. For example, Ensembl has developed a SARS-CoV-2 genome browser and related resources (https://covid-19.ensembl.org/index.html), which support genome browsing on the sequence of the virus, its gene functions, variation data and information on its protein sequence (13). Providing reliable sequence information for COVID-19 is essential both for our understanding of the virus and its pathogenicity, and for the development of any vaccines or treatments. A further essential area of study is understanding the genetic factors that dictate susceptibility to the disease from Genome Wide Association Studies (GWAS). The GWAS Catalog aggregates data from large cohort studies, making curated SNP-trait associations and summary statistics from both genome-wide and targeted studies (18). In the past year, the GWAS Catalog has released a new data deposition service, recognising the need to make available data from preprints prior to peer-reviewed publication. At the time of writing, there were 3997 studies from pre-prints with summary statistics, 4 of which were COVID-19 related. There were 8 published studies and 25 available published associations between SNPs and COVID-19 susceptibility.

EMBL-EBI resources have also been able to support drug discovery research to the COVID-19 portal. ChEMBL, EMBL-EBI's manually curated database of bioactive molecules with drug-like properties, has been curating the bioactivity data of potential anti-SARS-CoV-2 drugs. Drug discovery researchers have been screening candidate compounds to identify those with potential anti-SARS-CoV-2 activity, to identify already approved medications that can be repurposed for COVID-19. To date, several large-scale drug screening studies have been described and made available to the scientific community, either via preprints or as peer-reviewed publications. ChEMBL launched a special release in May 2020 (ChEMBL_27) focusing on studies that assess the potential anti-SARS-CoV-2 activity of compounds, with a further focus on studies that use cell-based assays in their screens. To date, ChEMBL has curated eight datasets on anti-SARS-CoV-2 compounds. Across all studies, 142 compounds have been shown to have some anti-SARS-CoV-2 activity (however, some of these also showed a high degree of cytotoxicity alongside their antiviral activity). Interestingly, only a few compounds have so far been shown to be active against SARS-CoV-2 in multiple studies: 14 compounds have been shown to be active in more than one study, and only 5 have shown activity across three or more studies. The ChEMBL_27 release also enables users to view the COVID-19-related results alongside previously available information for each compound, and to



COVID-19 Data Portal

About | Data Hubs | Federated EGA | Related resources | Our partners | Bulk downloads | Submit data

Viral Sequences | Host Sequences | Expression | Proteins | Biochemistry | Literature

Accelerating research through data sharing

Viral sequences

Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.

119,103 records >

Host sequences

Raw and assembled sequence and analysis of human and other hosts.

997 records >

Expression

Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.

72 records >

Proteins

Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.

651 records >

Biochemistry

COVID-19 pathways, interactions, complexes, targets and compounds.

2,255 records >

Literature

Search for the latest literature about SARS-CoV-2.

182,953 publications >

Related resources

A range of related resources for studying the SARS-CoV-2 coronavirus and the COVID-19 disease

About this portal

The COVID-19 Data Portal was launched in April 2020 to bring together relevant datasets for sharing and analysis in an effort to accelerate coronavirus research. It enables researchers to upload, access and analyse COVID-19 related reference data and specialist datasets as part of the wider European COVID-19 Data Platform.

Help make the portal better

Are you a researcher working with COVID-19 data? We would like to hear your feedback on this portal and the presentation of the data you are interested in.

Sign up to participate >

Share new COVID-19 data >

Contact our curator teams, who will assist you with submitting your data to EMBL-EBI repositories >

To enquire on how to collaborate on the European COVID-19 platform: ecovid19@ebi.ac.uk.

To share your data on COVID-19 Data Portal: virus-dataflow@ebi.ac.uk.

COVID DATA RESOURCES

Viral sequences
Host sequences
Expression
Proteins
Biochemistry
Literature
Related Resources

ABOUT

About the Portal
SARS-CoV-2 Data Hubs
Our Partners
Submit Data

Support & Feedback | Privacy Notice

elixir
EMBL-EBI
European Commission
Co-funded by the Horizon 2020 programme of the European Union
EOSC-Life

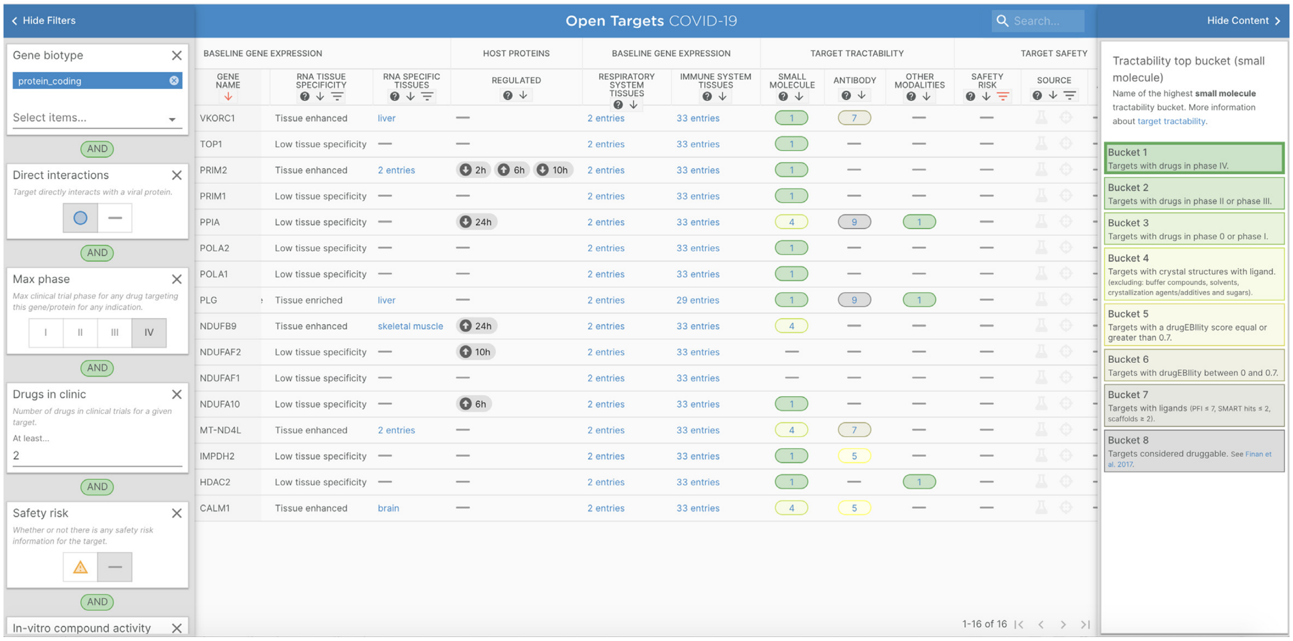
Figure 3. The COVID-19 Data Portal interface. The COVID-19 Data Portal provides a web entry point to services provided by the European COVID-19 Data Platform.

compare results for the same compound from different research groups.

EMBL-EBI is a partner in, and also hosts, Open Targets, a public-private partnership that uses human genetics and genomics data for systematic drug target identification and prioritisation (19). Open Targets has responded to the COVID-19 pandemic by developing a unique tool to aid filtering and prioritisation of human and viral (SARS-CoV-1 and SARS-CoV-2) proteins as potential drug targets for COVID-19 treatment (https://covid19.opentargets.org/). The tool integrates key datasets from publicly available resources via a simple user interface and is designed to filter potential drug targets by key properties of the data (Figure 4) to answer key questions such as 'which human proteins directly interact with a SARS-CoV-2 viral protein, are expressed in the respiratory system, and have compounds with in vitro activity against COVID-19 that modulate this target?'.

Datasets integrated in the tool include:

- Protein interaction and network data on whether the target interacts directly with a SARS-CoV-1 or -2 viral protein, or with another virus protein (source: IntAct).
- Expression, distribution, tissue and subcellular location data for a given human target (source: Human Protein Atlas and Expression Atlas).
- Whether target abundance is up- or down-regulated during SARS-CoV-2 infection, and at which time points (20).
- Drugs in clinical trials that modulate the target, including details on the number of drugs and maximum trial phase, and whether these trials are for COVID-19 treatment (source: ChEMBL).



< Hide Filters

Open Targets COVID-19

Search...

Hide Content >

Gene biotype ✕
protein_coding ⊗
Select items...

AND

Direct interactions ✕
Target directly interacts with a viral protein.
○ —

AND

Max phase ✕
Max clinical trial phase for any drug targeting this gene/protein for any indication.
I II III IV

AND

Drugs in clinic ✕
Number of drugs in clinical trials for a given target.
At least...
2

AND

Safety risk ✕
Whether or not there is any safety risk information for the target.
⚠ —

AND

In-vitro compound activity ✕

BASELINE GENE EXPRESSION			HOST PROTEINS	BASELINE GENE EXPRESSION		TARGET TRACTABILITY			TARGET SAFETY	
GENE NAME	RNA TISSUE SPECIFICITY	RNA SPECIFIC TISSUES	REGULATED	RESPIRATORY SYSTEM TISSUES	IMMUNE SYSTEM TISSUES	SMALL MOLECULE	ANTIBODY	OTHER MODALITIES	SAFETY RISK	SOURCE
VKORC1	Tissue enhanced	liver	—	2 entries	33 entries	1	7	—	—	
TOP1	Low tissue specificity	—	—	2 entries	33 entries	1	—	—	—	
PRIM2	Tissue enhanced	2 entries	↓ 2h ↑ 6h ↓ 10h	2 entries	33 entries	1	—	—	—	
PRIM1	Low tissue specificity	—	—	2 entries	33 entries	1	—	—	—	
PPIA	Low tissue specificity	—	↓ 24h	2 entries	33 entries	4	9	1	—	
POLA2	Low tissue specificity	—	—	2 entries	33 entries	1	—	—	—	
POLA1	Low tissue specificity	—	—	2 entries	33 entries	1	—	—	—	
PLG	Tissue enriched	liver	—	2 entries	29 entries	1	9	1	—	
NDUFB9	Tissue enhanced	skeletal muscle	↑ 24h	2 entries	33 entries	4	—	—	—	
NDUFAF2	Low tissue specificity	—	↑ 10h	2 entries	33 entries	—	—	—	—	
NDUFAF1	Low tissue specificity	—	—	2 entries	33 entries	—	—	—	—	
NDUFA10	Low tissue specificity	—	↑ 6h	2 entries	33 entries	1	—	—	—	
MT-ND4L	Tissue enhanced	2 entries	—	2 entries	33 entries	4	7	—	—	
IMPDH2	Low tissue specificity	—	—	2 entries	33 entries	1	5	—	—	
HDAC2	Low tissue specificity	—	—	2 entries	33 entries	1	—	1	—	
CALM1	Tissue enhanced	brain	—	2 entries	33 entries	4	5	—	—	

1-16 of 16

Tractability top bucket (small molecule)
Name of the highest **small molecule** tractability bucket. More information about target tractability.

Bucket 1
Targets with drugs in phase IV.

Bucket 2
Targets with drugs in phase II or phase III.

Bucket 3
Targets with drugs in phase 0 or phase I.

Bucket 4
Targets with crystal structures with ligand. (excluding: buffer compounds, solvents, crystallization agents/additives and sugars).

Bucket 5
Targets with a drugEBIlity score equal or greater than 0.7.

Bucket 6
Targets with drugEBIlity between 0 and 0.7.

Bucket 7
Targets with ligands (PFI ≤ 7, SMART hits ≤ 2, scaffolds ≥ 2).

Bucket 8
Targets considered druggable. See Finan et al. 2017.

Figure 4. The Open Targets COVID-19 Target Prioritisation Tool. This example displays the results (listed in the centre window) of filters for targets that are: protein coding, have a direct interaction with a viral protein, have at least two drugs (for any indication) in a phase IV clinical trial, have a modulating compound in *in-vitro* assays, and have safety information (left content side bar). The right content sidebar displays further information for targets when entries in the list are selected. In this example, the small molecule target tractability button was selected for a particular target to display the level of predicted tractability information for the target.

- Compounds that modulate the target and have been tested *in vitro* against SARS-CoV-2 and whether they had activity against the virus or not (source: ChEMBL).
- Target tractability assessment, including suitability for modulation with a small molecule or antibody (source: Open Targets Platform).
- Target safety information including known toxicity and organs affected (source: Open Targets Platform).
- List of peer-reviewed articles in which human targets co-occur with COVID-19 (source: Europe PMC).

The COVID-19 Data Portal also provides links to data submission systems, which provide data to the Platform, a bulk data download function, and links to currently 56 external data resources and services, spanning areas such as computational, clinical and epidemiological content. Ongoing work on the Portal includes the development of dashboard functions, integration of SARS-CoV-2 Data Hubs, a variety of data visualisations and deeper connections into clinical data (such as via a ‘Cohort Browser’).

The COVID-19 Data Portal also provides connections with other related resources, which at the time of writing included 5 other European projects, 7 computing support resources and 30 other databases and atlases containing information relevant to COVID-19 research. Examples of related resources and projects include the Johns Hopkins University COVID-19 Case Tracker (https://www.covidtracker.com/), Coronavirus Phylomes from Phylome DB (http://beta.phylomedb.org/covid19) and the European Clinical Research Infrastructure Network (ECRIN) COVID-19 Taskforce (https://ecrin.org/covid-19-taskforce).

The SARS-CoV-2 Data Hubs

The SARS-CoV-2 Data Hubs are the second component of the European COVID-19 Data Platform. They focus on providing tools and services for the validation, processing, analysis, interpretation, sharing and publication of viral sequence data. The SARS-CoV-2 Data Hubs are deployed to be used by public health agencies and scientists responsible for generating viral sequences at national or regional levels, with substantial user and tool support to help with their use. Specific tools on offer at the time of writing include a dedicated raw viral data uploader, a number of cloud-based processing workflows supporting data from different sequencing platforms and library preparation methods and a Notebook-based data exploration and visualisation system. Ongoing work includes the addition of further workflows, including phylogenetic analysis. A particular focus has been placed on ensuring that COVID-19 Data Hubs promote FAIR Data Principles, ensuring data remains Findable, Accessible, Interoperable and Reusable (1).

At the time of writing, there were 16 national SARS-CoV-2 Data Hubs, supporting 70% of data. Viral data included over 420 000 data records, including 97 000 raw viral sequence data sets from 38 countries and over 330 institutions.

The Federated European Genome-phenome Archive

The Federated European Genome-phenome Archive (FEGA) focuses on sensitive human biomolecular data relating to COVID-19 and is the third component of the European COVID-19 Data Platform. For these data, which are often required to remain within national borders, FEGA offers a unifying technology that allows national databases to be deployed and operated as nodes within a broader federation. This approach allows consistent interfaces (such as for submission and data access) across the federation while respecting requirements for national governance. Public metadata can be shared across the network so data discovery can operate from any point and lead to unambiguous endpoints for access requests and, ultimately, data access.

The COVID-19 pandemic has provided increased impetus for FEGA to become more widely operational to ensure rapid availability of COVID-19 host studies. This has led to an acceleration in the development of FEGA, including support for its underlying architecture, technical interfaces and software implementation to establish national nodes. To date, work has progressed on the governance and legal framework for FEGA, the technical interfaces, and demonstrator implementations in collaboration with the ELIXIR nodes (21). In addition, FEGA has been able to contribute to global standards for data sharing through the Global Alliance for Genomics and Health (GA4GH), a policy-framing and technical standards-setting organisation for responsible genomic data sharing. Its contributions have included standards for variation formats, secure data streaming, reference retrieval, data use ontology, encryption containers and phenotype exchange.

Building the COVID-19 Data Platform

EMBL-EBI was uniquely positioned to rapidly deploy a COVID-19 Platform due to a number of factors. First, EMBL-EBI has a 25 year history in operating bioinformatics infrastructure to provide biomolecular data and services. This infrastructure includes systems for the at-scale rapid capture, integration and publication of scientific data from the research community. Specifically, several EMBL-EBI data resources for relevant data types were already in place and, in many cases, already holding data of direct relevance to SARS-CoV-2 and COVID-19 research (e.g. EGA, PRIDE, Ensembl). Second, drawing on over 600 staff expert in bioinformatics, software engineering and data science, we were able to rapidly redirect human and computational capacity towards the effort. Third, our work over the last six years with European partners in pathogen data systems through the EU Horizon 2020 COMPARE project, particularly around the COMPARE Data Hubs (22), has allowed us to develop specific extensions to our existing infrastructure. Fourth, as a part of ELIXIR, with strong connections to national nodes, we have well-established working relations with those who operate infrastructure in national health and public health systems and the relevant European-level Research Infrastructures that sit alongside ELIXIR. Finally, our commitment to open science, through open data, FAIR principles (1) and data standards, has facilitated the integration of the diversity of data types of relevance to COVID-19 research.

While the Platform continues to be developed very actively with the incremental addition of new data, functions and services, we count a number of key milestones to date:

- Rapid deployment of the three components of the European COVID-19 Data Platform—the SARS-CoV-2 Data Hubs, the Federated EGA and the COVID-19 Data Portal;
- Integration of SARS-CoV-2 and COVID-19 data from 14 EMBL-EBI Data Resources and connection into a number of external resources and services;
- Mobilisation of raw viral sequence data at scale: with almost no sharing of raw viral sequence data in March 2020, by September the Platform has mobilised some 25 000 data sets, with 80% of these data routed through the SARS-CoV-2 Data Hubs; raw data are essential for systematic variation and assembly analysis with appropriate error handling;
- Establishment of a host of new tools, including a viral data submission tool (https://ebi-ait.github.io/sars-cov2-data-upload/), a selection of viral data processing workflows and a SARS-CoV-2 genome browser (https://covid-19.ensembl.org/index.htm);
- Deep collaboration with a broad range of partners at scientific and infrastructure level, and creation of a network of national coordinators and Platform nodes, such as in Sweden, Japan, Norway, Poland, and Slovenia (https://covid19dataportal.se, https://covid19dataportal.jp, https://covid19dataportal.no, https://covidhub.psnc.pl, https://covid19dataportal.si).
- As of October 2020, the COVID-19 Data Portal had received over 2.9 million web requests, from over 90 000 unique hosts. Since the COVID-19 Data Portal is an open resource, it is difficult to estimate the exact number of users who accessed it. The number of unique hosts, while not an exact count, is used as a useful indicator of the number of users.

While the COVID-19 crisis has driven the specific implementation of the European COVID-19 Data Platform, we have, by design, built concepts and technologies that will be sustained long into the future. Drawing on the foundation of existing EMBL-EBI and ELIXIR Data Resources, all new data available through the Platform are appropriately routed towards their respective underlying data resources for long-term preservation. Using such applications as EMBL-EBI Search (3), the future discoverability of the underlying data is assured. The new software elements that we have developed are reusable not only for rapid deployment in future infectious disease scenarios, but also in the service of broader scientific research, such as genetic disease applications of FEGA. Finally, our network of partners in the Platform, the data standards that we have worked upon and the conventions and practices that we have established will enrich our opportunities for data infrastructure development in the future.

SUPPORTING THE RESEARCH COMMUNITY THROUGH A GLOBAL CRISIS

Empowering cooperation in the specific field of coronavirus research is not the only way EMBL-EBI supports the research community. EMBL-EBI continues to support the life sciences research communities, both through its training programme and its regular service provision.

Training

An essential part of the work at EMBL-EBI, alongside its data services and research activities, is its Training programme. The EMBL-EBI Training programme is now in its tenth year and aims to train scientists at all levels to get the most out of publicly available biological data (www.ebi.ac.uk/training). We responded rapidly to the COVID-19 pandemic; although we had to cancel some courses at short notice, we have created a robust mechanism for the delivery of highly interactive virtual courses and have received encouraging feedback from participants.

The training programme currently has five major components:

Live virtual: Aiming to provide the same impact as our face-to-face courses in a virtual environment, to reduce our carbon footprint, shield our participants from communicable disease and increase accessibility.

Train anytime: A comprehensive collection of online bioinformatics courses, pre-recorded webinars and course materials, allowing individuals to choose when, where and how they learn.

Trainer support: Enabling trainer communities to develop relevant and high-quality training courses, supported by EMBL-EBI experts and materials, and underpinned by global standards.

Face-to-face: Delivering courses to develop practical skills and knowledge, at EMBL-EBI and at host organisations worldwide (face-to-face training currently suspended)**.**

Visits and secondments: Visiting scientists are embedded in an EMBL-EBI group for up to 6 months, often leading to longer-term collaboration (visits and secondments are currently suspended).

This combination of activities allows us to deliver training that is both of high quality and scalable: in 2019, we participated in 305 training and outreach events, and we delivered three train-the-trainer events, training 26 bioinformatics instructors. Meanwhile, in 2019 our 'train online' collection was accessed by 599 549 unique IP addresses and we expanded our online training offerings by adding two new courses and 59 new webinars. The expansion of our virtual training content in 2019 allowed us to respond promptly and effectively to the challenges of 2020, when we were able to adapt our programme to the restrictions brought on by the COVID-19 pandemic.

Furthermore, in late 2019 and early 2020 we began testing the use of commercial cloud-based systems for their potential to provide on-demand compute capacity for computationally intensive off-site courses. We can now create virtual machines (VMs) and Docker containers in the commercial cloud, support trainers to test their software and then provide access to the final VM/container/set of containers for trainees, giving them a very similar user experience to the experience that they would have if they were being trained face-to-face at EMBL-EBI. We accelerated our first use of this system in response to the COVID-19 pandemic.

Empowering access through service provision

Crucially, EMBL-EBI empowers international cooperation through its regular service activity, enabling searchers from across the globe to collaborate daily by sharing their data through EMBL-EBI data resources.

All EMBL-EBI resources have been continuing their service development capabilities through the past year. Significant progress has been achieved in a number of core EMBL-EBI resources, including many referenced above (7,9,13) as well as Pfam (23), Rfam (24), and RNACentral (25).

We have been working to support the research communities that rely on our resources though both the COVID-19 pandemic and the ever-changing requirements of research. For example, the MetaboLights team (26) has been engaged with the metabolomic community in a comprehensive exercise to redevelop its submission systems to capture the growing amount and diversity of metabolomic information. By redeveloping the MetaboLights submission processes, we have enabled our user community to make more comprehensive and high-quality data available as easily as possible. At the time of writing, the database hosts around 700 publicly available studies with more in preparation and awaiting public release due to publication requirements. The MetaboLights team is committed to continually developing our resources and services to become a central hub for metabolomics related data and tools in collaboration with international collaborators and our diverse user community.

Data submission to other deposition databases has also changed through the recent COVID-19 pandemic. For instance, three of the months during the first half of 2020 have seen a record number of submitted data sets in a single month to the PRIDE database (8), the world-leading proteomics data repository and part of the ProteomeXchange Consortium (27), reaching 566 data sets (∼25 datasets per working day) submitted in June 2020 alone. The total number of proteomics data sets in PRIDE is currently around 18 000 and the number of proteomics datasets submitted to PRIDE continues to grow very significantly year by year. At the time of writing, 30 public data sets in PRIDE were COVID-19 related.

One of the strengths of EMBL-EBI as an institute is that it hosts services, training and research programmes within the same institute. Proteomics has been at the centre not only of expanded efforts by our services, but also of coronavirus-related research by our teams. Researchers from EMBL-EBI, in collaboration with partners from across the globe, have used proteomics to identify the mechanisms through which SARS-CoV-2 hijacks human cells during infection to promote its own replication. In turn these discoveries have allowed researchers to identify a number of existing drugs that may be effective at targeting the virus (28). Similarly, an international research group including scientists from EMBL-EBI identified hundreds of interactions between SARS-CoV-2 proteins and human proteins, which can in turn be used to identify potential COVID-19 treatments (29).

The activities of PRIDE are increasingly focused on integrating proteomics information into other added-value EMBL-EBI resources. In this context, more than 40 human and mouse quantitative proteomics datasets (both baseline and differential) are now available via Expression Atlas. Additionally, there is an ongoing effort to improve the annotation (and then facilitate the re-use) of public proteomics datasets by adopting and extending the file format SDRF (Sample and Data Relationship Format) used by the EMBL-EBI resources ArrayExpress and Expression Atlas (30).

As well as facilitating submission and access to its existing resources, EMBL-EBI is continually developing its resource portfolio to keep up with the ever-changing nature of molecular biology data. To this end, in 2020 EMBL-EBI has launched the eQTL Catalogue, a compendium of uniformly processed human gene expression data and splicing quantitative trait loci (31). The eQTL Catalogue focuses on expression of genetic variants that are associated with specific splicing events within genes, or with expression levels of nearby genes. eQTL datasets are rare, but they are hugely valuable for the process of drug discovery: data from the eQTL Catalogue has already been incorporated in the Open Targets platform for drug target identification and prioritisation.

CONCLUSIONS

International cooperation, resource sharing and central data brokering are all principles upon which EMBl-EBI was founded. The COVID-19 pandemic has highlighted how important these practices are, and how crucial towards tackling and effectively addressing a global health crisis. International cooperation drove the creation and implementation of the European COVID-19 Data Platform, which in turn has connected scientists and practicing healthcare professionals not only throughout Europe, but all over the world. Cooperative resources have enabled new, groundbreaking research into the structure and mechanisms of action of the virus, which in turn are essential for the development of new effective treatments and vaccines.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of all who contributed to this paper. In particular, they would like to thank Anna Gaulton, Oana Stroe, Spencer Philips, Jessica Vamathevan, Helen Parkinson, Thomas Keane, Juan Antonio Vizcaino, Claire O'Donovan and Andrew Leach for their contributions, input and feedback. The authors also wish to acknowledge the Open Targets team, specifically David Ochoa, Daniel Suveges, Asier Gonzalez-Uriarte, Javier Ferrer, Aldo Segura-Cabrera, Andrew Hercules, David Hulcoop and Ian Dunham.

FUNDING

EMBL-EBI is indebted to its funders, including the EMBL member states, the European Commission, Wellcome, UK Research and Innovation, the US National Institutes of Health, our Industry Programme and many others; GWAS Catalog work was supported by National Human Genome Research Institute of the National Institutes of Health [5U41HG007823]. Funding for open access charge: European Molecular Biology Laboratory (www.embl.org).

Conflict of interest statement. None declared.

REFERENCES

1. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
2. Dyke,S.O., Philippakis,A.A., Rambla De Argila,J., Paltoo,D.N., Luetkemeier,E.S., Knoppers,B.M., Brookes,A.J., Spalding,J.D., Thompson,M., Roos,M. *et al.* (2016) Consent Codes: Upholding standard data use conditions. *PLoS Genet.*, **12**, e1005772.
3. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
4. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
5. Amid,C., Alako,B.T.F., BalavenkataramanKadhirvelu,V., Burdett,T., Burgin,J., Fan,J., Harrison,P.W., Holt,S., Hussein,A., Ivanov,E. *et al.* (2020) The European Nucleotide Archive in 2019. *Nucleic Acids Res.*, **48**, D70–D76.
6. Velankar,S., van Ginkel,G., Alhroub,Y., Battle,G.M., Berrisford,J.M., Conroy,M.J., Dana,J.M., Gore,S.P., Gutmanas,A., Haslam,P. *et al.* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.
7. Ferguson,C., Araújo,D., Faulk,L., Gou,Y., Hamelers,A., Huang,Z., Ide-Smith,M., Levchenko,M., Marinos,N., Nambiar,R *et al.* (2020) Europe PMC in 2020. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa994.
8. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
9. Blum,M., Chang,H-.Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2020) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa977.
10. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Felix,E., Magarinos,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
11. Meldal,B.H.M., Bye,A.J.H., Gajdos,L., Hammerova,Z., Horackova,A., Melicher,F., Perfetto,L., Pokorny,D., Lopez,M.R., Turkova,A. *et al.* (2019) Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.*, **47**, D550–D558.
12. consortium,P.D.-K. (2020) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.
13. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2020) Ensembl 2021. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa942.
14. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
15. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
16. Lappalainen,I., Almeida-King,J., Kumanduri,V., Senf,A., Spalding,J.D., Ur-Rehman,S., Saunders,G., Kandasamy,J., Caccamo,M., Leinonen,R. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.*, **47**, 692–695.
17. Drysdale,R., Cook,C.E., Petryszak,R., Baillie-Gerritsen,V., Barlow,M., Gasteiger,E., Gruhl,F., Haas,J., Lanfear,J., Lopez,R. *et al.* (2020) The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. *Bioinformatics*, **36**, 2636–2642.
18. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published

genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

19. Carvalho-Silva,D., Pierleoni,A., Pignatelli,M., Ong,C., Fumis,L., Karamanis,N., Carmona,M., Faulconbridge,A., Hercules,A., McAuley,E. *et al.* (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.*, **47**, D1056–D1065.
20. Bojkova,D., Klann,K., Koch,B., Widera,M., Krause,D., Ciesek,S., Cinatl,J. and Münch,C. (2020) Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*, **583**, 469–472.
21. Blomberg,N. and Lauer,K.B. (2020) Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. *Eur. J. Hum. Genet.*, **28**, 719–723.
22. Amid,C., Pakseresht,N., Silvester,N., Jayathilaka,S., Lund,O., Dynovski,L.D., Pataki,B.A., Visontai,D., Xavier,B.B., Alako,B.T.F. *et al.* (2019) The COMPARE Data Hubs. *Database (Oxford)*, **2019**, baz136.
23. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,GA., 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.
24. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z. *et al.* (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa1047.
25. RNAcentral Consortium (2020) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa921.
26. Haug,K., Cochrane,K., Nainala,V.C., Williams,M., Chang,J., Jayaseelan,K.V. and O'Donovan,C. (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.*, **48**, D440–D444.
27. Deutsch,E.W., Bandeira,N., Sharma,V., Perez-Riverol,Y., Carver,J.J., Kundu,D.J., Garcia-Seisdedos,D., Jarnuczak,A.F., Hewapathirana,S., Pullman,B.S. *et al.* (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.*, **48**, D1145–D1152.
28. Bouhaddou,M., Memon,D., Meyer,B., White,K.M., Rezelj,V.V., Correa Marrero,M., Polacco,B.J., Melnyk,J.E., Ulferts,S., Kaake,R.M. *et al.* (2020) The global phosphorylation landscape of SARS-CoV-2 infection. *Cell*, **182**, 685–712.
29. Gordon,D.E., Jang,G.M., Bouhaddou,M., Xu,J., Obernier,K., White,K.M., O'Meara,M.J., Rezelj,V.V., Guo,J.Z., Swaney,D.L. *et al.* (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, **583**, 459–468.
30. Perez-Riverol,Y. (2020) Towards a sample metadata standard in public proteomics repositories. *J. Proteome Res.*, **19**, 3906–3909.
31. Kerimov,N., Hayhurst,J.D., Manning,J.R., Walter,P., Kolberg,L., Peikova,K., Samoviča,M., Burdett,T., Jupp,S., Parkinson,H. *et al.* (2020) eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. bioRxiv doi: https://doi.org/10.1101/2020.01.29.924266, 29 January 2020, preprint: not peer reviewed.