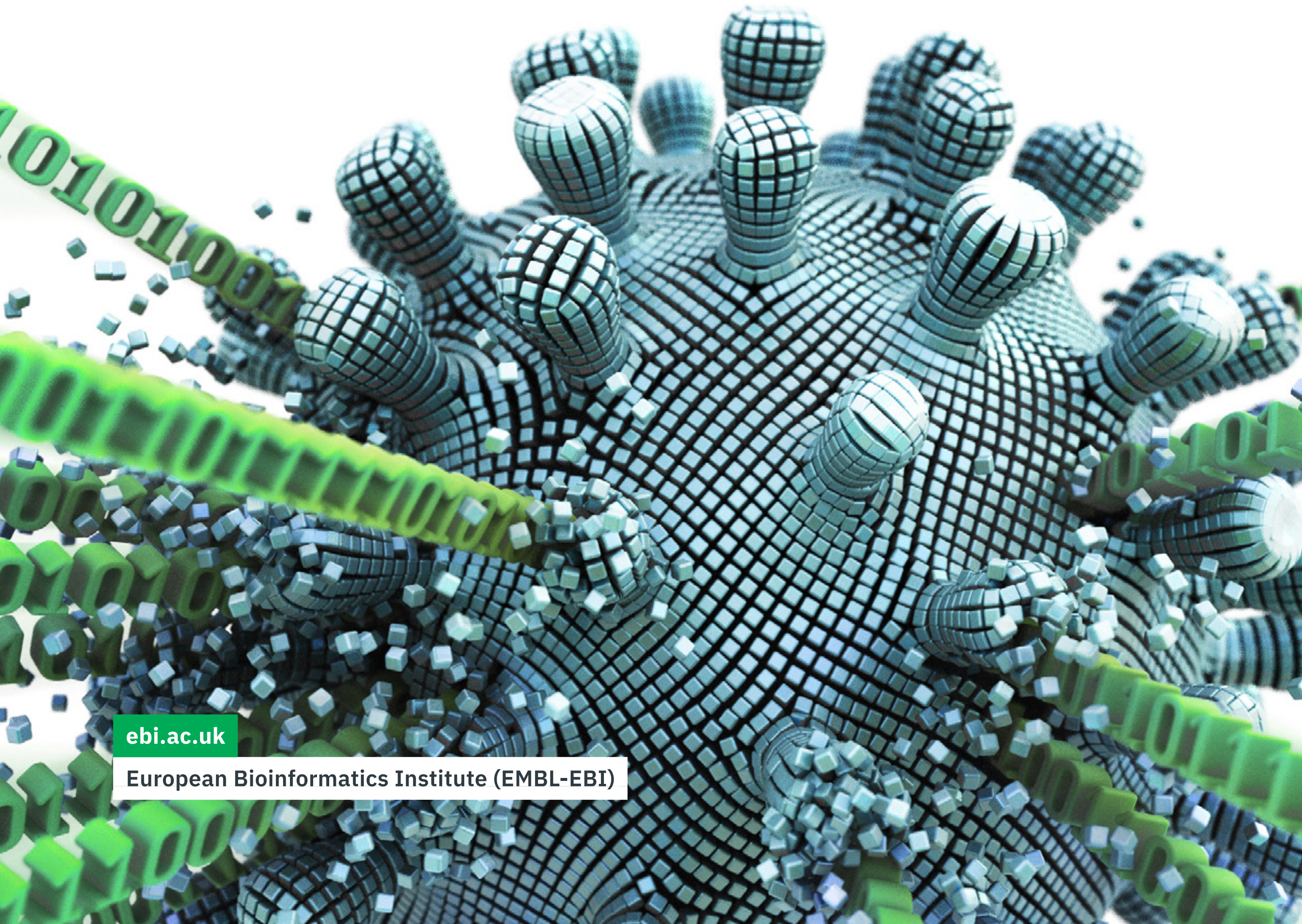
EMBL-EBI

Annual Report

2020



ebi.ac.uk

European Bioinformatics Institute (EMBL-EBI)

This publication was produced by the Communications team at EMBL's European Bioinformatics Institute (EMBL-EBI).

Cover illustration: Spencer Phillips

For more information about EMBL-EBI please contact:
comms@ebi.ac.uk

Table of contents

Who we are

EMBL's European Bioinformatics Institute (EMBL-EBI) is the world's leading source of public biological and biomolecular data. Our mission is to enable life science research and its translation to medicine, agriculture, industry and society by providing biological data, tools and knowledge.

We are part of the European Molecular Biology Laboratory (EMBL), an open science intergovernmental organisation and Europe's centre of excellence in life science research, services and training. EMBL is primarily funded by public research monies from over 20 member states.

Our vision

To benefit humankind by advancing scientific discovery and impact through bioinformatics.

Our missions

- To freely provide data and bioinformatics services to the scientific community in ways that promote scientific progress.
- To contribute to the advancement of biology through investigator-driven research in bioinformatics.
- To provide bioinformatics training to scientists at all levels.
- To disseminate cutting-edge technologies to industry and applications of science.
- To support, as an ELIXIR Node, the coordination of biomolecular data provision in Europe.

Our strategic priorities

- Increasing usage, utility and application of bioinformatics
- Extending collaboration and coordination
- Continuous improvement, maximising efficiency
- Building capacity and capability
- Supporting global expansion of biomolecular resources

Foreword

The emergence of the COVID-19 pandemic highlighted more clearly and poignantly than ever before the importance of open data and open science.

As scientists around the world uprooted their research to study the new coronavirus, EMBL-EBI's response to the pandemic focused on mobilising an impressive range of data through the European COVID-19 Data Platform, to power new insights and discoveries.

We doubled our data sharing, research and training efforts, acting as an enabler of open science. We also supported national initiatives with data coordination and setting up data infrastructures for COVID-19.

This rapid response was only possible because it could build on years of previous investment in research data infrastructure, international collaborations, and a growing culture of open data sharing.

We want to thank our funders and collaborators for their sustained support. We're grateful to the scientific community for championing open data, and for their intense drive to collaborate. Finally, we want to thank our staff who have shown incredible resilience and tenacity under pressure. Their work has kept the data flowing.

The hard-earned lessons from the COVID-19 pandemic are already crystallising: the need for more international collaborations to support knowledge exchange, better coordination across countries, a multidisciplinary approach to research, new technological solutions to accommodate new data types and answer new research questions, and the need for robust response mechanisms should another pandemic arise.

This is a unique opportunity to craft better global responses to future public health emergencies, from antimicrobial resistance to climate change. We must seize this chance and ensure open science, coordination and collaboration sit at the heart of what we build today to respond to tomorrow's challenges.

Rolf Apweiler, Joint Director

Ewan Birney, Joint Director

EMBL-EBI's response to COVID-19

Despite the COVID-19 pandemic, our data resources continued to operate and their usage grew, highlighting the essential nature of open data and open science during a global health crisis. Our collaborative and flexible approach enabled us to quickly adapt and contribute to the scientific response to the pandemic.

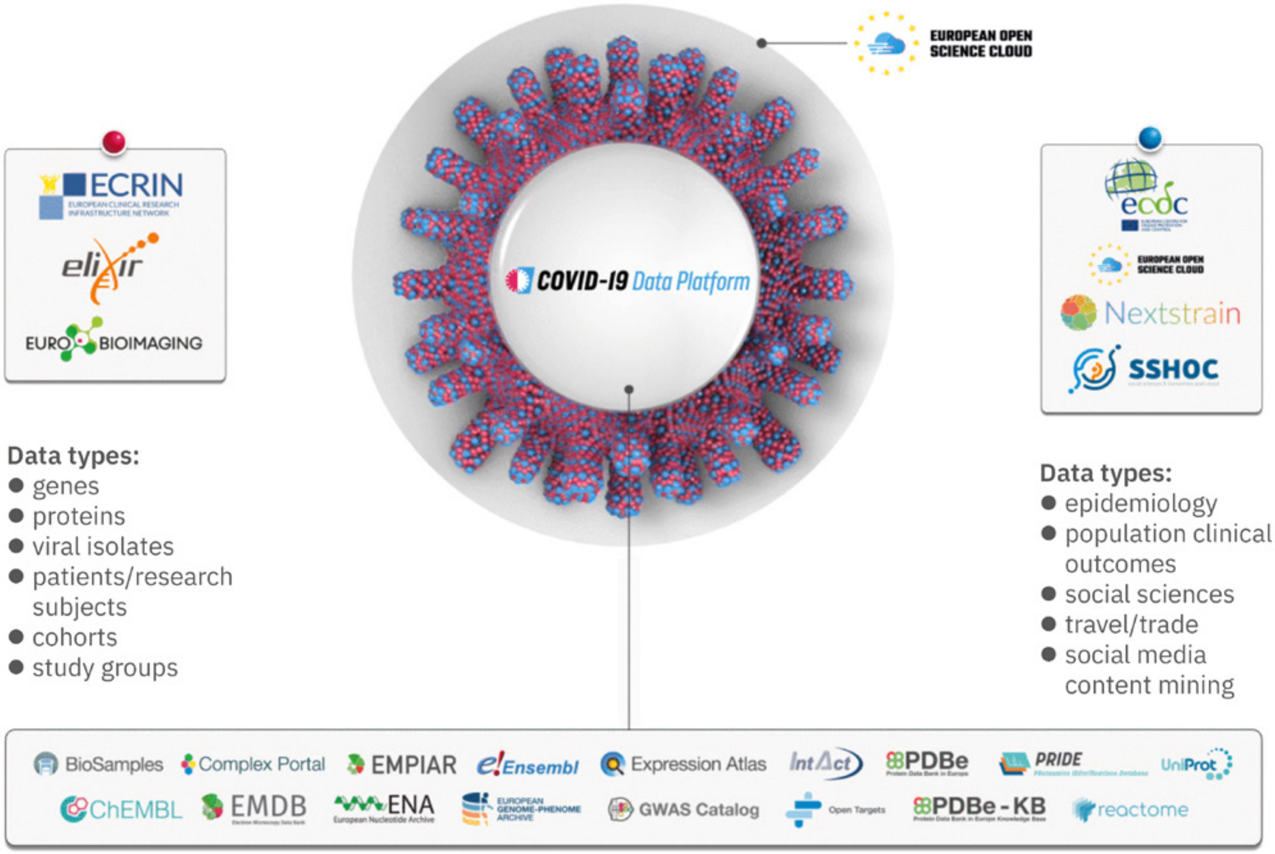
In an institute-wide effort, EMBL-EBI prioritised SARS-CoV-2 data submissions, doubling efforts to get new data rapidly into the public domain, to increase understanding of the virus and the disease.

An ecosystem for data sharing

To accelerate coronavirus research, EMBL-EBI promptly set up the **European COVID-19 Data Platform**, a collaborative online space that supports the rapid collection and comprehensive sharing and analysis of SARS-CoV-2 data generated around the world.

The European COVID-19 Data Platform was set up with the support of the European Commission, as part of the ERAvsCORONA Action Plan. It also relied on the support of ELIXIR – a research infrastructure for biological data, of which EMBL is one Node – and the European Open Science Cloud.

The Platform integrates molecular and genomic data from the virus and its host, into the broader context of healthcare, clinical trials, drug screening and biobanking. In 2021, the Platform continues to broaden in scope and grow in depth of data coverage.



Data types:

- genes
- proteins
- viral isolates
- patients/research subjects
- cohorts
- study groups

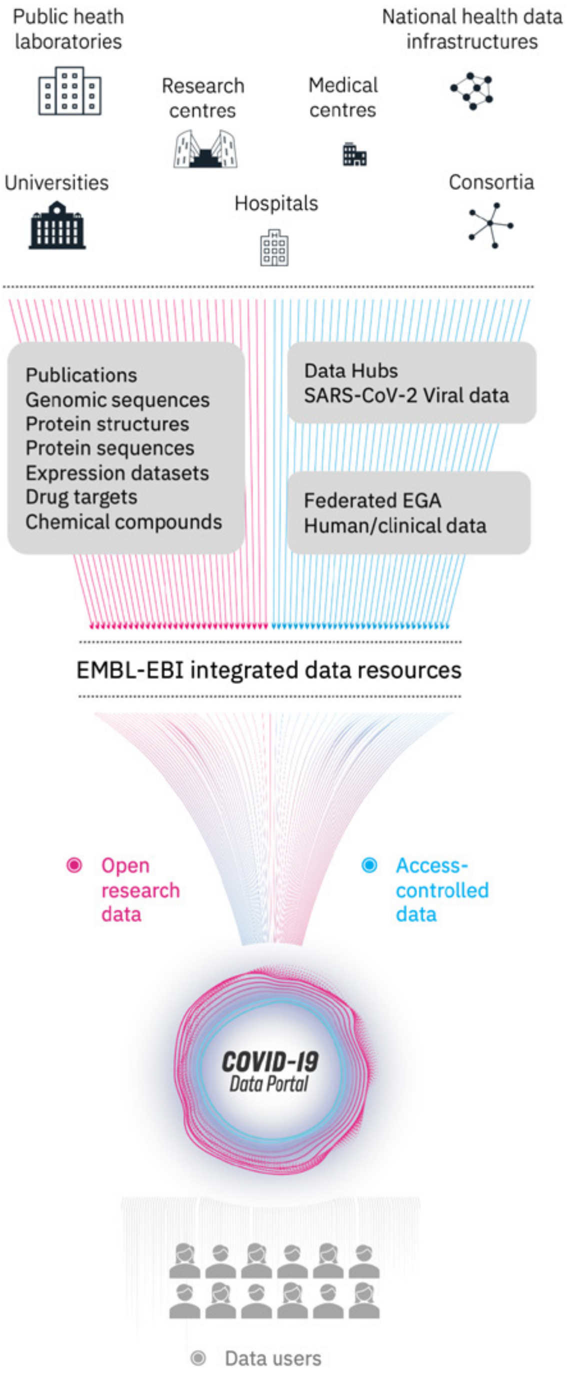
Data types:

- epidemiology
- population clinical outcomes
- social sciences
- travel/trade
- social media content mining

Spencer Phillips

The COVID-19 Data Platform consists of three connected components:

- SARS-CoV-2 Data Hubs, which organise the flow of SARS-CoV-2 sequence data, and provide comprehensive open data sharing for the European and global research communities. In 2020, EMBL-EBI supported the setup of 16 national data hubs.
- The Federated European Genome-phenome Archive, which provides secure, controlled-access sharing of sensitive COVID-19-related datasets from patients and research subjects.
- The COVID-19 Data Portal, which enables researchers to access and share relevant datasets and tools, spanning genomics, proteins, biochemistry, imaging, literature and more.



Caption: Flow of SARS-CoV-2 data into COVID-19 Data Portal

Xenia Sitja

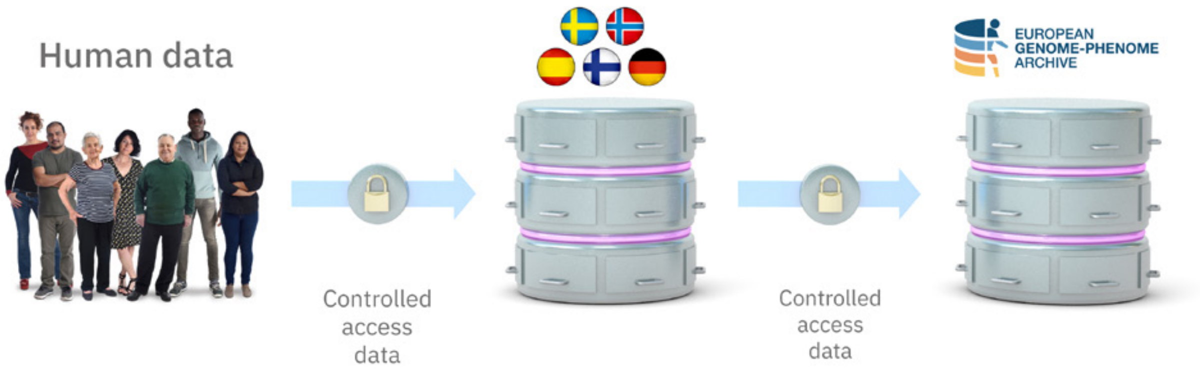
The speed and efficiency with which the Platform was set up was only possible because EMBL-EBI could draw on existing infrastructure, expertise and collaborations. With 20 EMBL-EBI databases feeding different types of molecular data into the **COVID-19 Data Portal**, this is one of the world's most comprehensive data resources for COVID-19 research. The world's largest COVID-19 sequencing initiative, COVID-19 Genomics UK Consortium (COG-UK) is one of the consortiums sending its data to the Portal.

EMBL-EBI also supported Italy, Japan, Norway, Poland, Slovenia, Spain, Sweden and Turkey to set up their own national portals to enable secure data sharing and galvanise national efforts. Our teams provided expertise, toolkits, standards and data models to fast-track data coordination. Many of the national portals have become the central repositories for the COVID-19 efforts within those countries, bringing all the activities together.

In addition to the Portal, researchers can access sequence data and variation analysis through the **SARS-CoV-2 Data Hubs**. This is essential for tracking, analysing and understanding new virus lineages – also called variants. The Data Hubs provide a toolbox for those generating and analysing viral sequence data, including workflows for data validation, sharing, assembly, visualisation and interpretation. Through national campaigns and extensive user support, EMBL-EBI successfully mobilised the majority of the world's raw viral sequence data through the system, and is now operating systematic variant-calling and other analyses on the data.

Many European countries are planning to collect human sequences and associated clinical outcome data in relation to COVID-19, to inform genetic association studies for host factors that determine susceptibility and severity of disease. Thus, there is an urgent demand for secure data sharing in this field.

As part of the European Commission's COVID-19 response through the ELIXIR-CONVERGE project, EMBL-EBI is developing the **Federated European Genome-phenome Archive**, an international network of human data resources with national and regional nodes. We are collaborating with partners in Germany, Finland, Norway, Sweden, and Spain as the first wave of nodes to address the COVID-19 host data sharing challenge, and create a scalable model for how this could work.



Human data mobilisation through federated efforts

Spencer Phillips

To help producers of viral data to share in a quick and seamless way, EMBL-EBI set up a “drag and drop” data submission system. This was particularly important as many research centres, public health laboratories or hospitals have minimal experience or processes for submitting data at such a scale.

COVID-19 Data Portal usage at the end of 2020

3.6 million
web requests

114,000
users
in 175 countries

500,000
data records

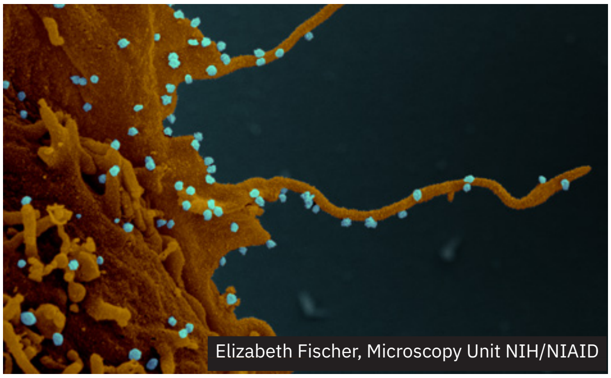
Identifying treatments

An international team of researchers, including the Beltrao Group and the ChEMBL team, have analysed how SARS-CoV-2 hijacks proteins in human cells. They investigated the interactions between viral and human proteins, and identified those human proteins that physically associate with proteins carried by SARS-CoV-2. Another study involving the Beltrao Group has shown how SARS-CoV-2 shifts its host cell’s activity to promote its own replication, and to infect nearby cells. These projects enabled the researchers to identify pharmacological agents that have the potential to be repurposed for treating COVID-19. Among these, 63 compounds are being tested in clinical trials.

Bouhaddou M. et al. (2020). The Global Phosphorylation Landscape of SARS-CoV-2 Infection.
Cell. DOI: 10.1016/j.cell.2020.06.034

Gordon D.E. et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.
Nature. DOI: 10.1038/s41586-020-2286-9

Gordon, D.E. et al. (2020). Comparative Host-Coronavirus Protein Interaction Networks Reveal Pan-Viral Disease Mechanisms.
Science. DOI: 10.1126/science.abe9403



Elizabeth Fischer, Microscopy Unit NIH/NIAID

SARS-CoV-2 viruses on a cell with filopodia.

The **ChEMBL** team undertook a rapid curation project to integrate cell-based COVID-19 drug screening data, and also contributed their expertise to a number of research projects to identify additional COVID drug repurposing opportunities.

The public-private partnership **Open Targets** developed a new tool, which aids filtering and prioritisation of human and viral proteins as potential drug targets for COVID-19 treatment. Key datasets from publicly available resources are parsed and integrated to inform therapeutic hypothesis generation via a simple user interface. The Open Targets Genetics Portal was also used to provide insight into susceptibility to SARS-CoV-2 infection

Monitoring lineages

The importance of tracking viral sequence data over time and around the world was underscored in the final months of 2020 by the emergence of the novel B.1.1.7 lineage. On behalf of COG-UK, the Gerstung Group and collaborators analysed the increasing prevalence of the novel strain.

Their analysis confirmed that B.1.1.7 is significantly more transmissible than previous variants, and indicated that lockdown measures applied until then were insufficient to contain its spread. The findings were used to inform the policy response to the pandemic by several European governments, and attracted significant media attention.

Vöhringer, H. et al. (2020). Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown. *Virological.*

Understanding the disease

The Marioni Group and collaborators at the Wellcome Sanger Institute, University of Cambridge and University of Newcastle used data from the Human Cell Atlas to identify differences in the immune response to COVID-19, between people with no symptoms, compared to those suffering a more serious reaction to the virus. The team discovered that whereas patients with mild to moderate symptoms, had high levels of B cells and helper T cells, which help fight infection, those with serious symptoms had lost many of these immune cells, suggesting that this part of the immune system had failed in people with severe disease.

Stephenson, E., et al. (2021). The cellular immune response to COVID-19 deciphered by single cell multi-omics across three UK centres. *Nature Medicine. DOI: 10.1038/s41591-021-01329-2*

The Birney Group received funding from UK Research and Innovation (UKRI) to carry out genetic association testing for COVID-19 symptom severity with a particular focus on copy number variation, which is likely to be one of the genetic sources of differences in the wide range of responses to COVID-19 infection. Work on genetic association testing for COVID-19 symptom severity continues using large datasets from Genomics England and the UK Biobank. Results from this project are contributing to an improved understanding of the genetic basis of COVID-19 symptom severity.

The virus at a molecular level

To complement the European COVID-19 Data Platform, EMBL-EBI's data resources rapidly set up dedicated areas and tools for coronavirus research, enabling scientists to delve even deeper into the data.

Genomics

Ensembl has created a COVID-19 genome browser, which enables researchers to visualise annotation of the SARS-CoV-2 genome. Each coding region fragment in the SARS-CoV-2 genome, has been annotated and cross-referenced using EMBL-EBI resources including Rfam and Gene Ontology Annotation (GOA). This gives researchers a better understanding of what is already known about different parts of the viral genome.

In addition, Ensembl provides access to SARS-CoV-2 variation, known problematic sites for resequencing experiments, and a community annotation project highlighting genomic regions of interest and linked publications.

Researchers can monitor the significance of new virus lineages using the Ensembl Variant Effect Predictor tool, which calculates the consequence and possible impact of genomic variation-—a crucial step in the efforts to stifle the pandemic.

The **Expression Atlas** team analysed 41 transcriptomics studies, which were fed into the COVID-19 Data Portal. Of these studies, 19 were from single cell studies and 22 from bulk transcriptomics.

The **Microbiome Informatics** team repurposed its VIRify tool to look for coronaviruses in environmental samples, allowing detection of hosts and potential zoonosis.

The Goldman Group detected systematic errors in SARS-CoV-2 whole-genome sequences that were made public in the early stage of the pandemic. They alerted the community to the types of errors that were occurring, which led to a number of preprints and papers being altered to avoid drawing erroneous conclusions from data artefacts.

Turakhia, Y. et al. (2020). Stability of SARS-CoV-2 phylogenies. *PLoS Genetics. DOI: 10.1371/journal.pgen.1009175*

The Iqbal Group developed new methods for measuring and tracking the quality of SARS-CoV-2 genomes, and applied them to study the full set of data submitted to the ENA.

Proteins

The 29 proteins of SARS-CoV-2 could hold molecular clues about the virus and how to stop it. Just one week after the viral genome appeared in the European Nucleotide Archive, **UniProt** released protein data to the scientific community.

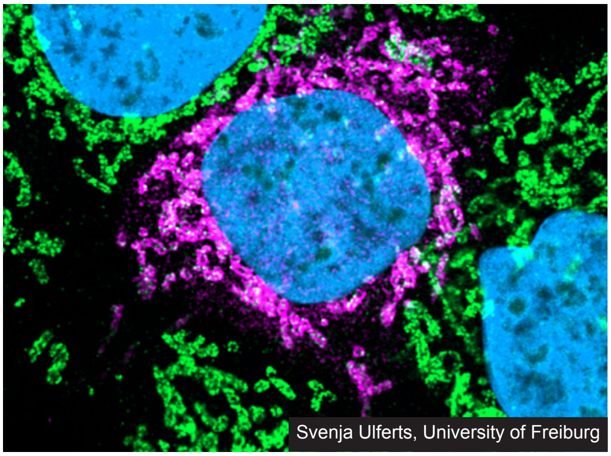
To enable researchers to explore and understand the role of each protein, Uniprot set up a dedicated COVID-19 Portal. In addition to functional annotation from the literature, automatic annotation adds information from a broader taxonomic range of viruses. Links to structures, drugs, interactions, molecular pathways as well as many other resources provide integrated information to help understand the biology and investigate routes to treatment.

In a complementary effort, the PDBe-KB COVID-19 Data Portal enables researchers to easily access and visualise structural data related to over 900 structures of SARS-CoV-2 components. The PDBe team also fast-tracked many planned features and incorporated them into the COVID-19 pages to improve the tools available for COVID-19 structure analysis.

The International Molecular Exchange consortium (IMEx), including the IntAct team at EMBL-EBI, released the first curated coronavirus interactome, a robust dataset consisting of 5,600 molecular interactions focused on proteins from several coronaviruses.

Bioimaging

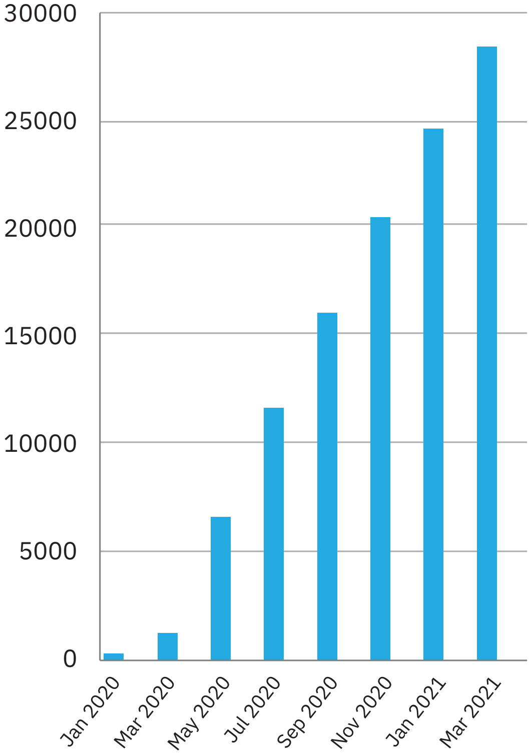
Data relating to many key SARS-CoV-2 studies were deposited to the **Electron Microscopy Data Bank (EMDB)** and the **Electron Microscopy Public Image Archive (EMPIAR)**. The diversity of samples has grown over time, with many entries relating to proteins and enzymes of interest, including the viral spike protein. EMPIAR released its first SARS-CoV-2 dataset on 1 May 2020, in less than 24 hours after the data was submitted.


Svenja Ulferts, University of Freiburg

Human and coronavirus protein interaction

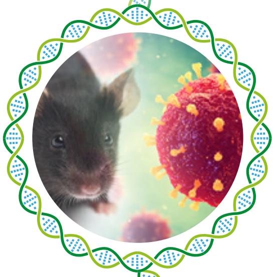
Scientific literature

The increased need for rapid access to research outputs resulted in significant changes in scientific publishing, with preprints taking center stage. In response to this shift, **Europe PMC** gathered COVID-19 preprints, making them available for searching, reading and reuse in a standard XML format. At the end of the year, 22,000 full-text COVID-19 preprints had been indexed in Europe PMC alongside peer-reviewed articles. This represents the largest collection of COVID-19 preprints in the world.



COVID-10 full text preprints available in Europe PMC

Highlights of the year

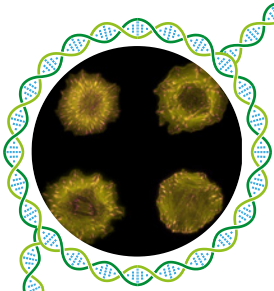


JANUARY

- eQTL Catalogue, a new resource for gene expression, launches - (page 22)
- Parkinson team identify new mutations likely to be responsible for rare childhood diseases (page 30)

FEBRUARY

- Pan-Cancer Analysis of Whole Genomes project shares novel cancer insights, including that cancer mutations occur decades before diagnosis (page 28)
- EMBL-EBI begins to gather and share data on the new SARS-CoV-2 virus as they become available (page 8)



MARCH

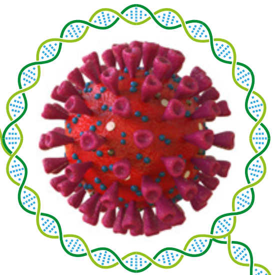
- Petsalaki group publishes insights into the regulators that control cell shape

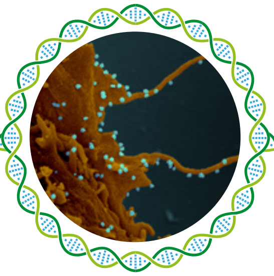
APRIL

- EMBL-EBI launches the COVID-19 Data Platform, a dedicated open data resource for SARS-CoV-2 (page 10)

MAY

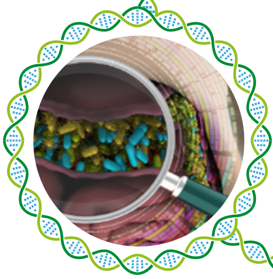
- Goldman group finds that many reported mutations in SARS-CoV-2 genome are actually technical artifacts (page 14)





JUNE

- Beltrao group shows how SARS-CoV-2 hijacks cells and what active compounds could counter viral activity (page 12)
- Flicek group reveals that DNA lesions caused by chemical damage often remain unrepaired for several cell division rounds (page 29)

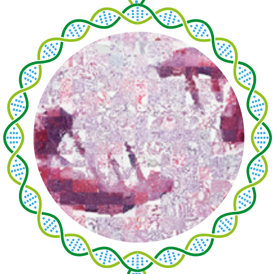


JULY

- EMBL-EBI Director Ewan Birney is appointed EMBL Deputy Director General (page 46)
- Finn group compiles a catalogue of over 200,000 genomes from the human gut microbiome (page 30)

AUGUST

- The genome of the tuatara, an ancient reptile, is sequenced and published in Ensembl
- Birney group finds that the heart trabeculae can influence the risk of heart failure (page 29)

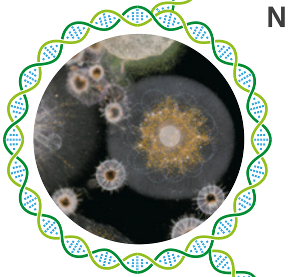


SEPTEMBER

- Gerstung group publishes statistical model that predicts a patient’s risk of developing oesophageal cancer

OCTOBER

- Beltrao group identifies drug targets common to three coronaviruses, and drugs that could be repurposed as COVID-19 treatments (page 12)



NOVEMBER

- Genome Targeting Catalogue, a dedicated data resource for CRISPR experiments, launches (page 22)
- EMBL-EBI joins the AtlantECO project, which aims to understand the impact of climate change on the Atlantic Ocean (page 38)

DECEMBER

- IntAct releases first curated coronavirus interactome (page 15)
- Gerstung group shows B.1.1.7 lineage of SARS-CoV-2 is 30-50% more transmissible (page 13)

2020 in numbers

800 members of staff*

**includes fellows, supernumeraries, trainees and visitors*

from **41** million unique IP addresses

82 million requests to our websites on an average day

3.6 million web requests received by COVID-19 Data Portal

from over **114,000** users in 175 countries

390 petabytes of raw data storage

292 journal papers and preprints published of these

41

are to EMBL only

177

active grants at EMBL-EBI

136

collaborative grants with researchers in 62 countries from 674 institutes

46

training webinars

545,000

unique IP addresses accessed Train online

756

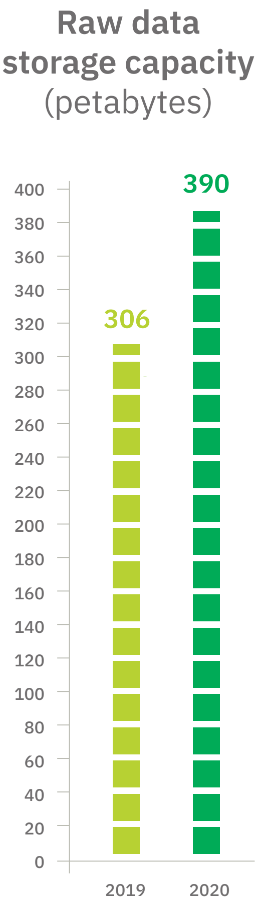
participants in Industry Programme knowledge-exchange workshops

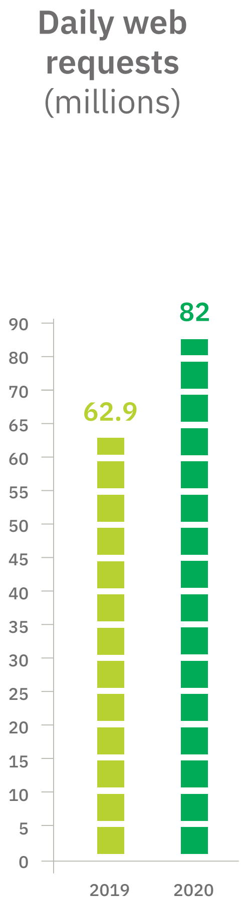
1500

people participated in our public engagement events

Data resources

The COVID-19 pandemic did not deter the scientific community from contributing to open science by submitting data to EMBL-EBI's data resources. In fact, in 2020, we saw an overall increase in data submissions and a significant spike in data usage.





Data usage and submission

On average, EMBL-EBI received more than 82 million web requests per day in 2020. These came from more than 41 million unique IP addresses over the course of the year. The heaviest usage, measured by unique IP addresses, came from the USA (32%), China (8%), the UK (8%), Germany (7%) and India (5%).

EMBL-EBI has also seen the amount of data submitted to our resources continue to grow. For example, the human genome and phenome data deposition grew by over 50%. Imaging data depositions—including electron microscopy data—surpassed all previous annual depositions by 120%, while electron cryo-microscopy data grew by 164%.

To manage this growth, we have increased our raw storage capacity to 390 petabytes at the end of 2020, compared to 307 petabytes in 2019.

Why do IP addresses matter?

Our data resources are open, which means users don't have to sign in. This makes it difficult to estimate the number of users. The number of IP addresses is an indication of the number of users, but not an exact count. The growth in IP addresses in 2020 reflects the remote working context.

What is a web request?

A web request is defined as any time a user or computer algorithm asks for information on our web pages using http. Requests may retrieve an entire webpage or just a single piece of information from an EMBL-EBI data resource.

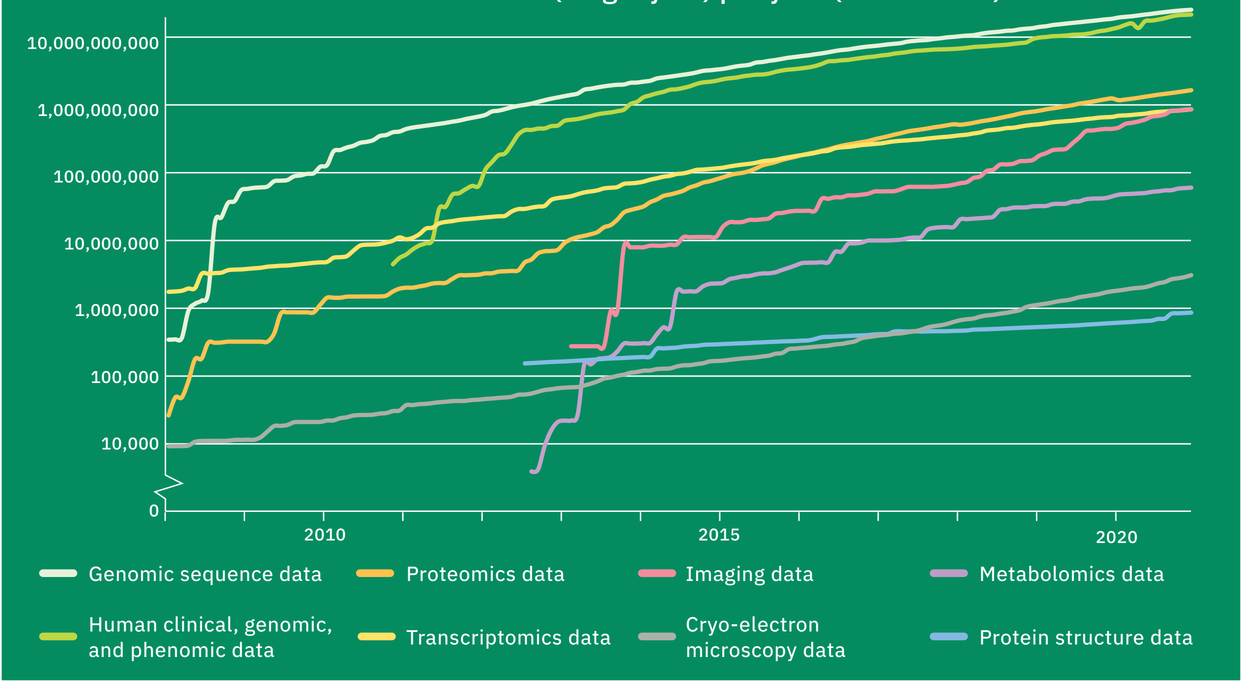
What is a petabyte?

The petabyte is a multiple of the unit byte for digital information. There are approximately 1 million gigabytes in a petabyte – the equivalent of 2000 years worth of music in MP3 format.

DATA GROWTH BY EMBL-EBI SERVICE

The scientific community has continued to submit data to EMBL-EBI's databases throughout the COVID-19 pandemic. Over the course of the year, data deposition for human clinical, genomic, and phenomic data grew by half, and the volume of cryo-electron microscopy data doubled.

Volume of data (megabytes) per year (2008–2020)



New databases

eQTL Catalogue

Genome-wide association studies (GWAS) can advance our knowledge of human health and food safety. They measure the contribution of genetics in determining traits and common diseases. Their results are generally attached to genetic markers, and are thus not actionable by the pharmaceutical industry. Expression QTLs (eQTL) however provide a clear path from genetic markers to potential drug targets. In collaboration with Open Targets, the Sanger Institute and the University of Tartu, EMBL-EBI brought together the largest uniformly processed collection of eQTL results in the eQTL Catalogue. The database will help improve our understanding of gene expression regulation across tissues and cell types.

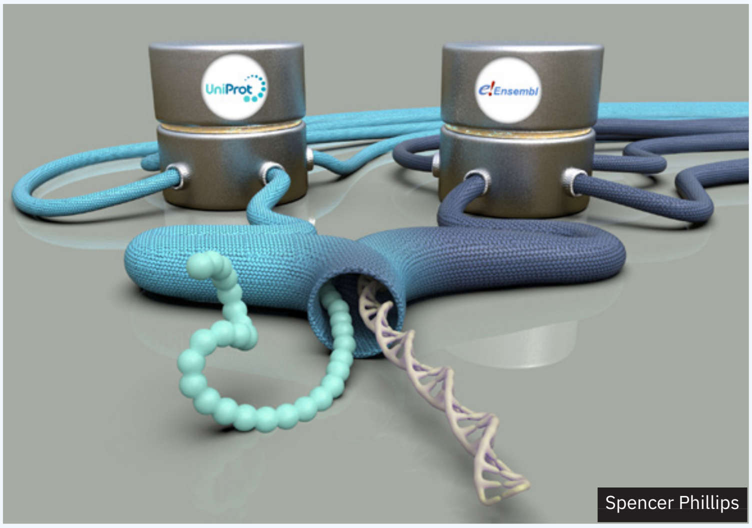
www.ebi.ac.uk/eqtl

Genome Targeting Catalogue

As genome editing is increasingly used in life science research, the need for a central repository for these datasets becomes increasingly clear. In 2020, EMBL-EBI launched the Genome Targeting Catalogue, a public repository of experiments using CRISPR-Cas enzymes, manually-curated from published literature. The repository covers over 40 species, and aims to make CRISPR experiments easier to find and explore.

www.ebi.ac.uk/gtc

In the pipeline: GIFTS portal



Spencer Phillips

In an attempt to bring protein and genomics data together, the Ensembl and UniProt teams launched the Genome Integrations with Function and Sequence (GIFTS). This allows Ensembl annotators and UniProt curators to efficiently work together and resolve discrepancies between these two major databases. GIFTS will be integrated to new Ensembl and UniProt releases, thus ensuring consistency between the two major databases.

www.ebi.ac.uk/gifts/

Ones to watch: Data growth

Bioimaging

EMBL-EBI's bioimage data resources have seen significant growth in 2020. The number of public BioImage Archive datasets increased almost threefold, with significant effort going into improving submission processes. Our other two added-value imaging data resources, the Electron Microscopy Public Image Archive (EMPIAR) and the Electron Microscopy Data Bank (EMDB), also saw record numbers of depositions and growth. EMDB released its 10,000th entry and is on track to reach 20,000 released entries in 2022 – a doubling in just two years compared to the 18 years it took to reach 10,000 entries. This is testament to the central role cryo-EM now plays in structural biology.

Microbiome

The microbiome database MGnify increased its coverage of microbiome data analyses by over 24,000, with more than 400,000 public analyses now available. MGnify also continued to generate metagenome assemblies, which culminated in over 30,000 assembly analysis results. Despite having only performed this service for a little over two years, this represents the largest collection of analysed metagenome assemblies.

Single-cell sequencing

Single-cell RNA sequencing has fueled much discovery and innovation in medicine over recent years. The Single Cell Expression Atlas holds 181 single-cell RNA-Seq studies, across 14 different species. To date, the team has made available the expression profiles of over 4 million cells using open source analysis pipelines. The Atlas now presents marker genes for each cell cluster via heatmaps, as well as human organ anatomograms to enable a better visualisation of cell types and their gene expression.

Metabolomics

Improvements to the MetaboLights guided submission tool enabled the team to respond to the exponential growth in metabolomics data, many of which come from medically-related studies, including COVID-19 multiomics data.

Proteomics

During 2020, the PRoteomics IDEntifications Database (PRIDE) received a record number of submitted datasets—over 440 each month—further strengthening its role as the world-leading proteomics data repository.

Curating the world of proteomics

As a Scientific Curator, Deepti J. Kundu checks and validates new datasets submitted to the PRoteomics IDEntifications Database (PRIDE) - the world's largest public database dedicated to mass spectrometry-based proteomics. With Deepti's help, PRIDE grew by 10,000 datasets between 2019 and 2021, reaching a total of 25,000 at the time of writing.

Before joining EMBL-EBI, Deepti studied botany, psychology and bioinformatics in India, before doing a PhD in Cheminformatics in the Czech Republic.

Deepti J. Kundu

"We do everything we can to make PRIDE intuitive, and to simplify data submission," explains Deepti. "Alongside my biocuration work, I also act as an intermediary between our users and our technical team. I offer support with improvements to the data submission tool and with adding new features that our users need."

"Every day brings a new challenge but the best thing is that our team loves solving problems so there's always someone to consult with."

Upside of 2020: "Working from home meant I could spend more time with my family, especially my daughter. The institute was very supportive of staff with childcare responsibilities."

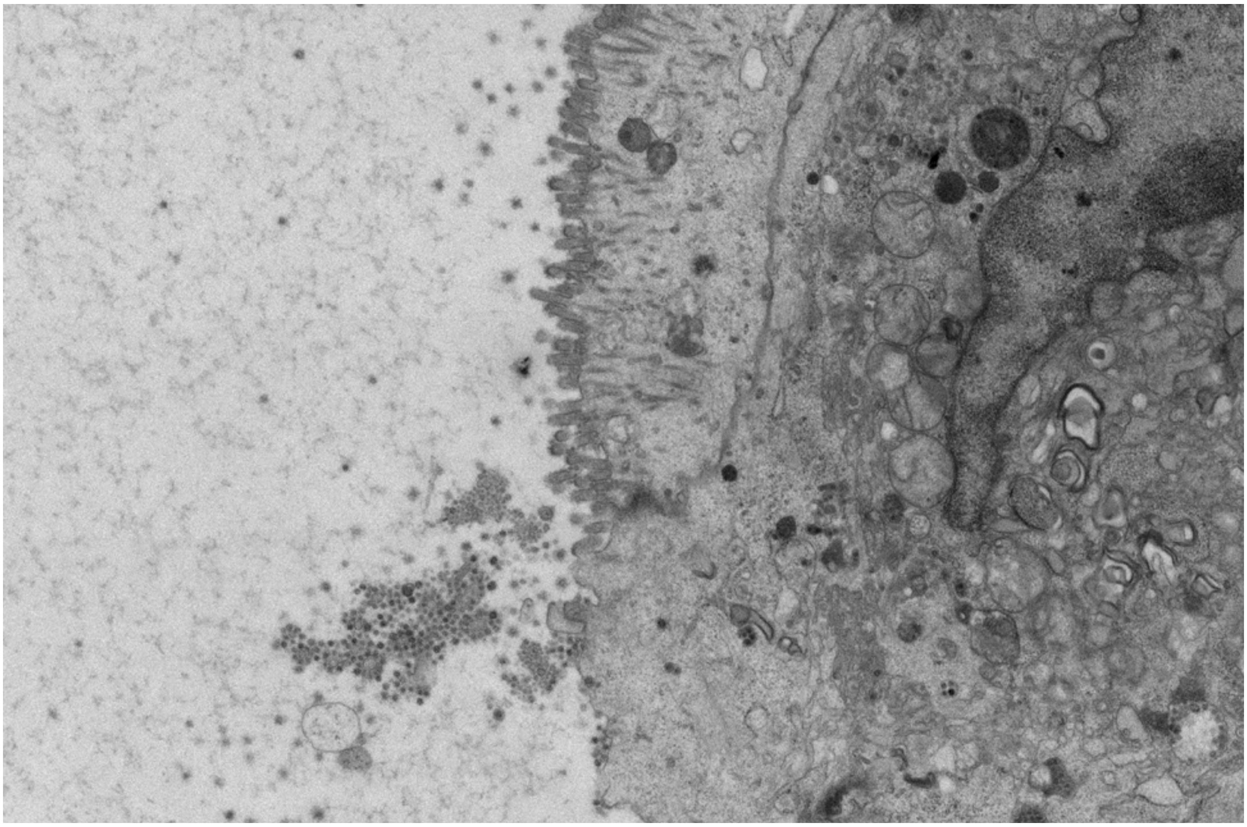
New capabilities

Genomics

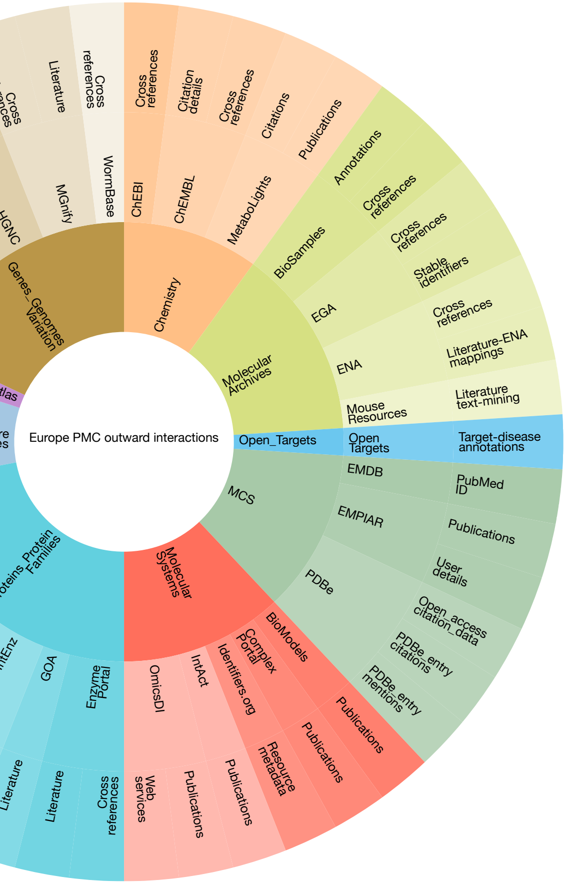
To support large-scale biodiversity studies like the Darwin Tree of Life and the Vertebrate Genomes Project, the teams have launched **Ensembl Rapid Release**—a bi-weekly release of vertebrate genomes. This significantly speeds up the dissemination of these data to the scientific community. The team contributes its expertise in annotation and comparative genomics to add value to the data. In 2020 they annotated nearly 250 genomes, a near-doubling from the previous year. The work will continue to scale up in the coming years.

Bioimaging

In response to the increase in imaging data submissions, EMBL-EBI is aligning its BioImage Archive with community needs and wider trends. A new light microscopy data submission template now enables users to annotate their datasets in a more complete manner, in alignment with the Image Data Resource (IDR). A new data acquisition pipeline for X-ray tomography and super-resolution structured illumination microscopy datasets was also developed.



SARS-CoV-2 productively infects human gut enterocytes



Text mining

Our life science literature database Europe PMC mines all incoming content for key biological entities such as genes, proteins and diseases. This is a major undertaking for millions of records. For example, the EMERALD research project has developed a new machine learning model for the automatic identification of biome keywords in metagenomics articles. Because concepts such as biome material, host state or body site can be shared as structured metadata, they can be reused in resources such as MGnify to give a richer context for computational analysis across multiple microbiome datasets, rather than remaining buried in the text of articles.

FAIR data

EMBL-EBI continues to champion FAIR principles, to make research data Findable, Accessible, Interoperable and Reproducible (FAIR). To do so, we are working with data producers to improve the process of submitting to our databases, improving the usability of our data resources, and creating new training material.

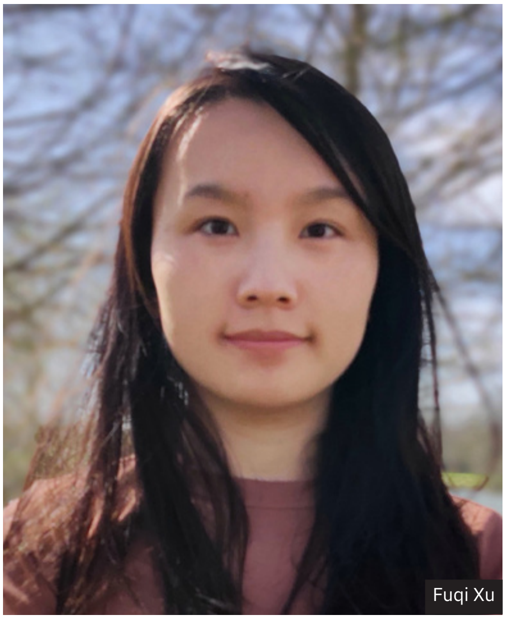
Europe PMC changed its manuscript submission system in line with Plan S, an international initiative to ensure scientific publications that result from public grants are published in compliant open access journals or platforms. From January 2021, researchers have been able to make final manuscripts freely available in Europe PMC with a CC-BY license and zero month embargo.

Retiring resources

In 2021 the ArrayExpress infrastructure for array data from functional genomics studies will be shut down, and BioStudies will become the only system managing existing and incoming expression and other functional genomics datasets.

A passion for FAIR data

Fuqi Xu is a Bioinformatician who works with metadata standards and ontologies, and dedicates a lot of her time to making data FAIR (Findable, Accessible, Interoperable and Reusable), supporting the change of data management culture in the scientific community.


Fuqi Xu

Fuqi studied biological sciences at Nanjing University in China before specialising in bioinformatics at the Karolinska Institute.

She now works on the FAIRplus project, which develops processes, guidelines, tools, and indicators to make data FAIR. "Such 'FAIRification' solutions can be implemented in both academia and industry by organisations of all sizes," explains Fuqi.

"Within the FAIRplus project, we work with pharma companies to understand their needs and develop FAIR solutions for the Innovative Medicines Initiative project. We're currently developing a FAIR Cookbook - a useful guide to putting the FAIR principles in practice."

Upside of 2020: "With many events going virtual, knowledge exchange was at our fingertips and conferences became more accessible to researchers worldwide."

Research highlights

Understanding cancer

EMBL-EBI scientists contributed to the Pan-Cancer Analysis of Whole Genomes (PCAWG) project: a collaboration of more than 1,300 scientists and clinicians from 37 countries, who conducted the most comprehensive study to date of whole cancer genomes. PCAWG has significantly improved our fundamental understanding of cancer and marked out new directions for its diagnosis and treatment.

The Gerstung Group and collaborators demonstrated that cancerous mutations can occur decades before diagnosis. They analysed the whole genomes of over 2,600 tumours from 38 different cancer types to determine the chronology of genomic changes during cancer development. Their findings could have significant implications for cancer diagnosis and monitoring.

GERSTUNG, M. et al. (2020). The evolutionary history of 2,658 cancers. *Nature. DOI:10.1038/s41586-019-1907-7*

Spencer Phillips, Moritz Gerstung

DNA mutations – a mosaic of tumour images

As part of PCAWG, the Brazma Group and collaborators analysed genomic and transcriptomic data from over 1,000 donors and more than 25 cancer types. They created the largest catalogue to date of cancer-specific RNA alterations, and proposed a new classification of gene fusions in cancer. The study greatly improved our knowledge of gene expression changes in cancer, opening avenues for identifying new cancer drug targets and therapies.

PCAWG Transcriptome Core Group et al. (2020). Genomic basis for RNA alterations in cancer. *Nature. DOI: 10.1038/s41586-020-1970-0*

Myeloproliferative neoplasms are blood cancers that are thought to originate through the acquisition of a driver mutation in a hematopoietic stem cell (HSC). The Cortes-Ciriano Group and collaborators shed light on how and when this unusual mutation occurs. These findings have implications for early detection and monitoring of patients.

VAN EGEREN, D. et al. (2021). Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in JAK2-mutant myeloproliferative neoplasms. Cell Stem Cell. DOI: 10.1016/j.stem.2021.02.001

The identification of genes that are essential for the survival and fitness of cancer cells is an important field in cancer research. Finding genes that are essential only for the survival of cancer cells – but not the survival of normal cells – can be an excellent therapeutic approach. The Petsalaki Group developed a tool and database to navigate genome-scale essentiality data from a large number of studies. They used it to identify a potential target for melanoma, which currently has few therapeutic options.

Sharma, S. et al. (2020). CEN-tools: an integrative platform to identify the contexts of essential genes. Molecular Systems Biology. DOI: 10.15252/msb.20209698

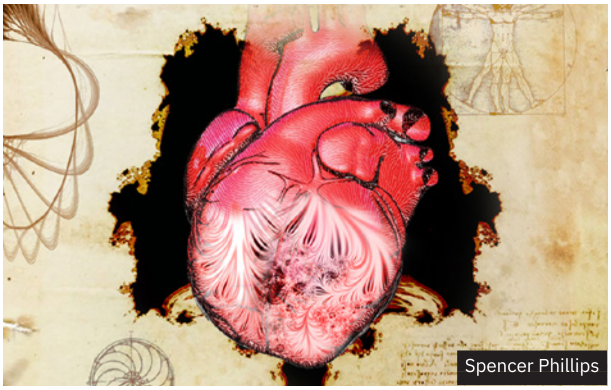
The Flicek Group and collaborators examined the evolution of tumours in mice following chemical damage. Their results showed that DNA lesions caused by chemical damage are not eliminated immediately, but are passed on – unrepaired – through several rounds of cell division.

AITKEN, S.J. et al. (2020). Pervasive legion segregation shapes cancer genome evolution. Nature. DOI: 10.1038/s41586-020-2435-1

A closer look at the heart

Research from the Birney Group and collaborators shed light on questions asked by Leonardo da Vinci 500 years ago. They used machine learning and genetic analyses to examine the structure of the inner surface of the heart using 25,000 MRI scans from UK Biobank. They found that the complex network of muscle fibres lining the inside of the heart, called trabeculae, allows blood to flow more efficiently and can influence the risk of heart failure.

MEYER, H.V. et al. (2020). Genetic and functional insights into the fractal structure of the heart. Nature. DOI: 10.1038/s41586-020-2635-8



Spencer Phillips

The heart and its trabeculae, first described by Leonardo da Vinci.

The Marioni Group used complementary approaches, from single-cell transcriptomics to live imaging and genetic labelling, to identify a novel progenitor population in the heart—a new type of heart cell. The study provides detailed insights into the formation of early cardiac cell types, with particular relevance to cell-based cardiac regenerative therapies.

Tyser R.C.V., et al. (2021). Characterization of a common progenitor pool of the epicardium and myocardium. Science. DOI: 10.1126/science.abb2986

Exploring health and disease

As part of the International Mouse Phenotyping Consortium (IMPC), EMBL-EBI researchers categorised which genes are essential for supporting life. The study compared mouse data from IMPC with human cell lines provided by the Broad Institute to create categories indicating how crucial a gene is to producing viable life. The results from this study could help researchers identify mutations responsible for rare childhood diseases.

CACHEIRO P. et al. (2020). Human and mouse essentiality screens as a resource for disease gene discovery. *Nature Communications. DOI: 10.1038/s41467-020-14284-2*

In a deep exploration of the human gut microbiome, the Finn Group and MGnify colleagues generated the Unified Human Gastrointestinal Gut catalogue, consisting of over 200,000 bacterial genomes, representing approximately 5,000 species. 70% of these genomes are yet to be experimentally isolated and sequenced. The data is freely accessible through MGnify.

ALMEIDA, A. et al. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnology. DOI: 10.1038/s41587-020-0603-3*

Bacteria are the most abundant cellular life form on the planet, critical to every ecosystem. Unlike humans, whose genomes are incredibly similar, bacteria can have very different genomes even within a single species. This flexibility has made it very hard to compare bacterial genomes. The Iqbal Group solved this problem, showing how, by using a reference built from an ensemble of previous genomes, it was possible to compare many bacterial genomes and cleanly access mutations. This is just the first step on a road to enabling many future studies of bacterial genomics.

Colquhoun, R et al. (2020). Nucleotide-resolution bacterial pan-genomics with reference graphs. *bioRxiv. DOI: 10.1101/2020.11.12.38037*

For the past 40 years, the HUGO Gene Nomenclature Committee (HGNC) has championed standardised gene naming in human genetics, fostering a vibrant research environment where knowledge is communicated and aggregated efficiently. With the increasing relevance of genomics to healthcare, the need for a consistent language to refer to genes has become even more crucial. In 2020, HUGO published its guidelines for naming genes.

Bruford, E.A. et al. (2020). Guidelines for human gene nomenclature. *Nature Genetics. DOI: 10.1038/s41588-020-0669-3*

Exploring the molecular mechanisms of ageing

Most research about ageing is focused on increasing life span, but Melike Donertas, a recent PhD graduate at EMBL-EBI, wants to delve deeper.

Melike Donertas

"I'm interested in the compression of comorbidities—preventing or postponing multiple diseases so we can continue to live healthy lives well into old age. During my PhD, I explored a number of topics: drug repurposing for ageing, the molecular mechanisms of ageing, and age-related diseases."

Melike is also part of the International Society for Computational Biology's Student Council, and, as such, is actively involved in organising bioinformatics webinars in Turkey, her home country. "It's so exciting to see that great bioinformatics and computational biology research are happening everywhere!"

The next step in Melike's academic career is a postdoc at the Max Planck Institute for Biology of Ageing, where she will be focusing on the links between the gut microbiome and ageing.

Upside of 2020: "I've seen that working from home is manageable and can be quite enjoyable—even in the last year of your PhD!"

Training and outreach

The EMBL-EBI Training Programme aims to deliver world-leading training in bioinformatics and scientific service provision to the research community. We aim to empower scientists at all career stages to make the most of biological data, and to strengthen bioinformatics capacity across the globe. The team is part of EICAT – the EMBL International Centre for Advanced Training.

Going virtual

In a rapid response to the pandemic, EMBL-EBI quickly made a range of its practical courses virtual. The extensive webinar programme and online tutorials provided immediate training opportunities for the research community as soon as lockdown took hold.

In 2020, 46 webinars took place, with 1,277 participants. This was more than a doubling of uptake compared with previous years.

In addition, 545,000 unique IP addresses accessed EMBL-EBI's online tutorials through the Train online portal. Our first virtual course made use of our new virtual training room in the commercial cloud. This significantly increases our capacity to run multiple courses in parallel, and to provide training around the globe more seamlessly.

We have a large and ever-growing catalogue of training content. This includes 176 online courses and pre-recorded webinars, and an impressive 1.6 terabytes of materials from face-to-face courses.

EMBL-EBI Training in 2020

46
webinars with
1,277
participants

176
online courses and pre-recorded webinars available

545,000
unique IP addresses accessed Train online portal

A real drive to improve

A new job can be daunting, but Mahfouz Shehu, a Full Stack Developer at EMBL-EBI faced an additional challenge: he's never seen the office or met his colleagues. Mahfouz is just one of 100 new staff remotely on-boarded by the institute in 2020.

Mahfouz Shehu

Originally from Nigeria, Mahfouz studied computer science at Coventry University in the UK. At EMBL-EBI, he builds and maintains websites. With a suite of over 40 data resources and tens of projects, the institute keeps Mahfouz and his colleagues in the Web Development team very busy.

"The role is extremely varied and engaging," says Mahfouz. "We work with different languages and systems, and there's a real drive for continuous improvement. Recently, my team completely redesigned EMBL-EBI's Training website, an essential tool for any researcher doing bioinformatics and data analysis. We improved the User Interface and the User Experience, and refined the search function, so users can easily find the right courses for them."

Upside of 2020: "Spending more time at home allowed me to focus on gaining new skills. I also got to speak to my family more often."

A new look and feel

The EMBL-EBI Training website has had a complete makeover and now provides seamless access to the different types of training on offer.

Two channels—live and on-demand—direct users though the offering, and all the training activities have been redesigned to optimise usability. The new website enables potential learners to find the right form of training for them, thanks to the integration of a powerful search function. Over 171,000 unique users accessed online tutorials in the new site in the first few months since launch.

The benefits of the new site include:

- Quick and easy ways to find relevant bioinformatics training
- Flexible, free-text search based on EMBL-EBI search
- The ability to combine search and filters to quickly identify courses
- Modern, clean look and feel with simple navigation

Collaborative training

EMBL-EBI works closely with its collaborators to offer relevant and up-to-date training for a range of communities. Through externally-funded projects, EMBL-EBI trainers delivered many virtual training sessions for initiatives such as BioExcel, CABANA and the Global Alliance for Genomics and Health (GA4GH).

The Massachusetts Institute of Technology (MIT) in conjunction with EMBL-EBI delivered two metagenomics bioinformatics courses, using a fully-remote model. This included containerised versions of EMBL-EBI tools and software to allow hands-on sessions. For many years, demand for training in metagenomics has far outpaced the number of courses we can deliver face-to-face. We are hopeful that virtual courses will start to bridge this gap in metagenomics and microbiome training.

The Uhlmann group was part of the organisation committee of the From Images to Knowledge (I2K) 2020 workshop for the bioimaging community. The learning-oriented event featured hands-on tutorials and opportunities for spontaneous networking. The feedback was overwhelmingly positive and the team has since been contacted by several large conferences to provide advice.

For the third successive year, EMBL-EBI was delighted to host Masters students from the prestigious Grandes Ecoles in France, under the rolling internship programme generously funded by the French Embassy in the UK.

Public engagement and communication

In the face of the pandemic, it was more important than ever for EMBL-EBI to raise enthusiasm and interest in science. Although the methods had to adapt, public engagement activity continued in 2020. Our staff and students engaged with more than 1,500 members of the public during 29 different events, from classroom talks and science comedy to remote engagement, such as virtual Cafe Sci and a series of "in conversation with" events for the anniversary of the Human Genome Project. In collaboration with colleagues from the European Learning Laboratory for the Life Sciences (ELLS), we supported professional development for almost 200 teachers.

Because of the COVID-19 pandemic, the 2020 PDB Art exhibition was developed as a virtual experience, allowing people from around the world to view the artworks created by local school students. The opening of the virtual exhibition also took place online, with invited speakers, videos from the student artists, and a guided tour.

In 2020, EMBL-EBI was mentioned in the press 1,115 times, across 67 countries—a doubling of 2019 coverage. This included major media outlets, such as BBC News, Sky News, The Financial Times, Publico, EuroNews, Der Standard. The institute's social media follower base continued to grow across platforms.

Video became a central communications tool, and was successfully used to communicate internally at key points of the pandemic, and externally via social media channels.

In the first few months of the pandemic, the institute also supported UK Research and Innovation with the development of the Coronavirus explained website, which aimed to provide accessible and authoritative scientific information about the virus and the disease. The website was successfully handed over to UKRI's team in July 2020.

Tahli Turner, Viewbank College, Melbourne, Australia

Rhodopsin is a biological pigment found in the retina, and enables vision in low light conditions.

Innovation and translation

Industry programme

EMBL-EBI’s Industry Partnerships team responded to the challenge of the COVID-19 pandemic by making its scheduled activities fully virtual. Digital engagement tools helped maximise audience involvement and participation.

The team successfully delivered three quarterly meetings and six knowledge exchange workshops for Industry Programme members. Compared with in-person meetings in 2019, the average workshop participation increased twofold. A total of 756 industry delegates, from all 25 pharmaceutical and agritech member companies attended.

Two companies joined the EMBL-EBI Industry Programme: Japanese pharmaceutical company Eisai, and Lonza, a large contract development and manufacturing organisation, which represents a new sector for the programme.

To further increase engagement with the private sector, the team set up a dedicated quarterly newsletter featuring news about EMBL-EBI research, data resources and updates on the programme of activities tailored to industry users.

iStock

Open Targets

The public–private partnership Open Targets, which includes EMBL-EBI, the Wellcome Sanger Institute, and pharmaceutical companies GSK, Bristol Myers Squibb, and Sanofi, uses human genetics and genomics data to identify and prioritise drug targets.

Open Targets established virtual ways of staying connected, including 15 training sessions for 475 participants across nine countries. Despite delays caused by lab closures, the Open Targets Validation Lab began its first experiments to validate oncology targets.

The team also considerably enhanced its use of human genetics data for drug target identification by incorporating a novel machine learning model into the Open Targets Genetics Portal, which identifies the most likely underlying causal gene from GWAS data. This work was greatly assisted by collaboration with the GWAS Catalog, and by the release of the eQTL Catalogue. The Genetics Portal evidence now feeds directly into the Open Targets Platform to create an integrated ecosystem of tools.

Adobe stock

Human and environmental health applications

The Ensembl Variant Effect Predictor (VEP) Genotype to Phenotype (G2P) software is now being used in a clinical setting for genetic diagnoses, saving clinical scientists significant time. The Greater Manchester Genomic Medicine Centre confirmed that VEP-G2P increases the precision of genetic testing for eye inherited disorders, while at NHS Lothian in Scotland there is now a whole exome sequencing service for patients with severe developmental delay, which uses VEP-G2P.

Hannah Currant, a recent PhD graduate from the Birney Group collaborated with Moorfields Eye Hospital in London to identify genetic associations using imaging data from UK Biobank. Her research discovered novel gene associations related to retinal morphology, and has highlighted areas for potential improvement during routine screening for common eye disorders such as glaucoma.

In an effort to improve scientific knowledge of proteins linked to Alzheimer's Disease, UniProtKB updated almost 300 proteins, concomitant with Genome Ontology annotation. With the IMEx consortium, they produced a disease-specific molecular interaction network. This was made possible by funding from NIH and Alzheimer's Research UK, and the data are now available through a specific Uniprot diseases website.

The Transform-MinER software developed by Jon Tyzack, a postdoc in the Thornton Group at EMBL-EBI, has received funding from EMBL Technology Development for further development and experimental validation. Transform-MinER enables high-throughput discovery of enzymes that can catalyse novel reactions for the environmentally friendly production of drugs and chemicals.

In the pipeline:

PDBe supported data sharing in large projects and emerging communities, in particular, toxicology data in EU-ToxRisk and a pilot for protein design data capture.

ChEMBL is part of the EUbOPEN project, a major public-private initiative that aims to develop high-quality chemical tool compounds for 1000 proteins. This will empower academia and industry to explore disease biology and unlock the discovery of new drug targets and treatments.

The Marioni Group has begun a collaboration with GSK on single-cell data analytics, while the Finn Group and Unilever have extended their collaboration, which relates to the skin microbiome.

EMBL-EBI is a partner in the ambitious AtlantECO project, set to map the microscopic organisms in the Atlantic ocean, including microbes found on plastic litter. All the data will be made available in the ENA and MGnify data resources.

EMBL-EBI provides data coordination and bioinformatics expertise to three projects focusing on accelerating genome to phenome research for farmed animals: AQUA-FAANG (fish), BovReg (cattle), and GENE-SWitCH (pig and chicken).

Genomic medicine

Human genomics is going through a paradigm shift whereby the majority of the genetic data available for research will be produced by healthcare.

EMBL-EBI continues to support EMBL member states with the implementation of national genomic programmes, offering strategic advice, reference data, training and support to join the Federated European Genome-phenome Archive, set to launch in 2021. This is an international cross-border federated network of human data resources with national and regional nodes. EMBL-EBI is already collaborating with partners in Germany, Finland, Norway, Sweden, and Spain.

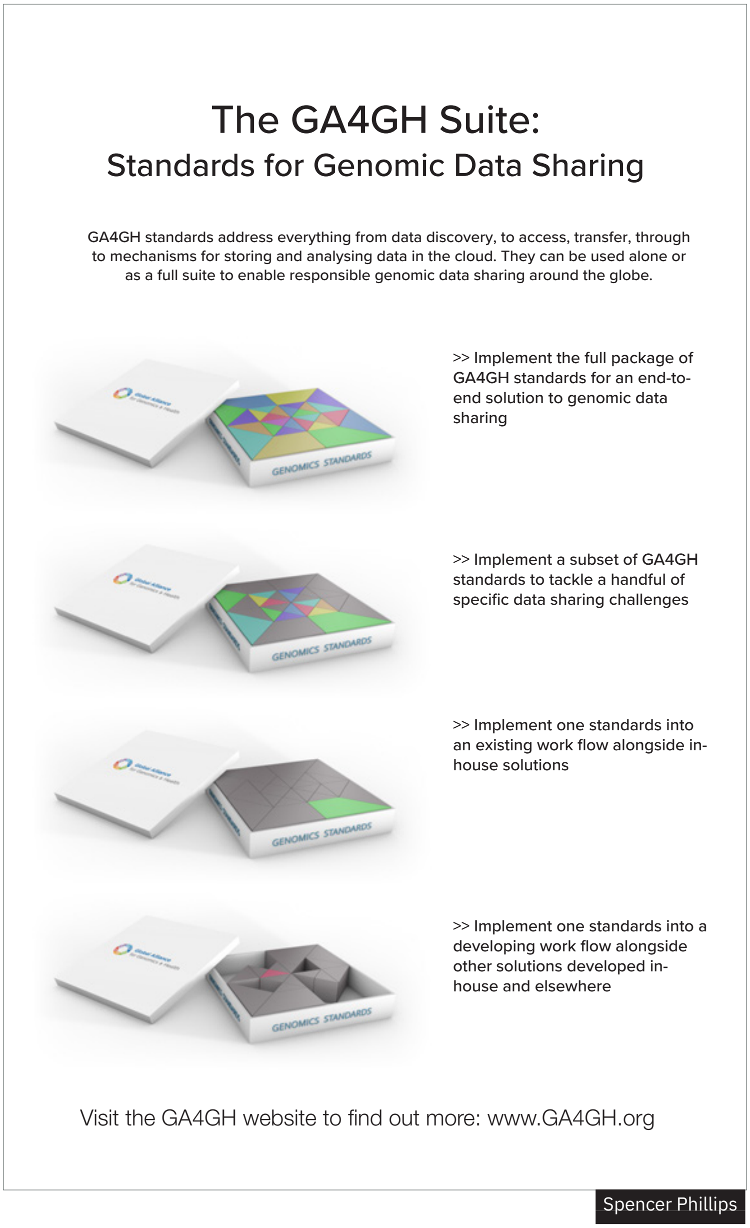
Before the pandemic, EMBL-EBI hosted an inbound delegation from INSERM, the French National Institute for Health and Medical Research, to explore possibilities for bioinformatics knowledge transfer to embed the French Plan for Genomic Medicine 2025. EMBL-EBI Director, Ewan Birney, also delivered a presentation on genomic medicine to the Institute Regina Elena in Rome, which focuses on cancer research. EMBL-EBI staff also delivered a virtual training course to Danish researchers covering data collection, analysis and interpretation for genomic medicine.

EMBL-EBI continued to support the Global Alliance for Genomics and Health (GA4GH), the policy-framing and technical standards-setting organisation, which seeks to enable responsible genomic data sharing within a human rights framework.

The GA4GH Connection Demos are real-world implementations of GA4GH standards across multiple institutions. EMBL-EBI helped develop the Horizontal demo, which reliably produces the same research results regardless of biomedical platform. We also supported the development of Application Programming Interfaces (APIs) to provide a standard format for listing genomics web services along with their metadata, and provided one of the first real-world implementations of GA4GH Researcher Passports.

Since 2020 all new submissions to the European Genome-phenome Archive include Data Use Ontology (DUO) terms to streamline datasets access. EMBL-EBI further contributed to the Machine Readable Consent Guidance, which embeds DUO terms in consent forms and was approved as a GA4GH standard. Over 200,000 DUO annotations world-wide have been made to date. These facilitate more rapid discovery and access to research data across organisations.

To improve data exchange and interoperability, EMBL-EBI's BioSamples database exports metadata from 17 million samples in the GA4GH Phenopackets format. In 2020, additional efforts were deployed to enrich this export by automated curation focusing on human disease annotations.



The GA4GH Suite:

Standards for Genomic Data Sharing

GA4GH standards address everything from data discovery, to access, transfer, through to mechanisms for storing and analysing data in the cloud. They can be used alone or as a full suite to enable responsible genomic data sharing around the globe.

>> Implement the full package of GA4GH standards for an end-to-end solution to genomic data sharing

>> Implement a subset of GA4GH standards to tackle a handful of specific data sharing challenges

>> Implement one standards into an existing work flow alongside in-house solutions

>> Implement one standards into a developing work flow alongside other solutions developed in-house and elsewhere

Visit the GA4GH website to find out more: www.GA4GH.org

Spencer Phillips

Technology infrastructure

EMBL-EBI continues to undertake the modernisation of its processes and technology infrastructure. These constitute the backbone on which our open data resources, tools and training operate.

The technical teams were instrumental in enabling the transition of EMBL-EBI from a largely office-based institute to complete remote working for all 850 staff in a very short timeframe.

The teams have also supported the set-up and growth of the COVID-19 Data Portal in record time. The Portal uses the EBI Search which indexes most of the data stored by the institute. The design of the Portal is based on the EMBL Visual Framework 2.0, and is continuously updated to improve user experience, making the exploration and analysis of data seamless.

Smarter, greener data centre

The institute was in the process of a major data centre move when the first UK lockdown started. Over 100 racks of equipment were being transferred to a highly resilient data centre, powered by 100% certified green energy and innovative cooling technology.

This move is part of a long-term plan to improve and diversify our data storage capabilities, and expands our off-campus data centre space so we can bring data analysis and service hosting to the same location. The new infrastructure has increased networking capacity between EMBL-EBI's data centres and to the external world while providing us with the physical, cooling and electrical capacity to continue to expand, to support our activities. The second phase of the migration will take place in 2021.

Kao Data

New EMBL-EBI data centre

Where data flows 24/7

Like a giant digital library, EMBL-EBI's public databases have been growing year on year. With petabytes of data now accessed by millions of users around the world, the institute's database administrators play a more important role than ever. Sri Attili is one of them.

Sri Attili

"My role is to make sure EMBL-EBI's biological databases are up and running at all times, that the data is securely stored, easy to access, and that proper back-ups are in place. It's a large, complex database infrastructure, and there is no margin of error."

Sri's energy and enthusiasm for learning new skills has meant that in just two years, she's already mastered two new database management software systems.

"I love the diverse environment of EMBL-EBI. We have colleagues from all over the world and it's amazing to learn about everyone's culture. People are very friendly and open-minded."

Upside of 2020: "The pandemic came at a terrible time, as my young son was in India when the first lockdown happened. When I finally managed to return there, it was a huge relief, and EMBL-EBI was very supportive, allowing me to work from abroad for a few months."

Better storage and archiving

EMBL-EBI's current archiving system, called FIle REplication (FIRE) provides storage for key EMBL-EBI data resources including ENA, EGA, EMPIAR and 1000 Genomes. Use of FIRE continues to grow rapidly, doubling in size over the last two years. By September 2020, the average ingress of new data to the FIRE archive had increased to 1.6 petabytes per month.

Negotiations with public cloud providers have progressed during the year and, once complete, will enable EMBL-EBI to expand its internal IT infrastructure from the public cloud. This will provide greater capacity and specialised capabilities that can be offered to research and service teams in 2021.

Embassy Cloud provides private, secure, virtual-machine-based workspaces within the EMBL-EBI infrastructure, where collaborators can make use of their own customised workflows, applications and datasets. A major review of Embassy Cloud was completed in 2020 taking in technology developments and user feedback. The resulting plan for the next generation of Embassy cloud will be implemented in 2021.

Improving user experience

EMBL-EBI technical and service teams work very closely with the user community to continuously improve user experience. The aim is to make EMBL-EBI's data resources, tools and software easy to use and understand by the diverse range of users, including new communities accessing bioinformatics tools, such as clinicians, pharmacists, geneticists etc.

EMBL-EBI is continuing the transitioning of its corporate website to the EMBL Visual Framework 2.0, improving the look and feel, as well as the functionalities for users and content owners alike. The new framework will empower teams across EMBL to create and control their own content alongside centrally managed views of team members, group publications, open vacancies, news and events.

Working in collaboration with the Digital and IT teams at EMBL, microsites are just one part of a complex and ambitious project that pulls together data from multiple external systems to provide site owners from all six EMBL sites the ability to drag and drop widgets from the EMBL's Visual Framework 2.0 into their microsites.

In the pipeline:

We are expanding the use of public cloud at EMBL-EBI through support to individual teams and training materials.

We commissioned a new high performance computing (HPC) cluster to provide a new facility for large-scale compute with optimised storage for service and research teams.

Our adoption of IT Service Management has advanced establishing processes that will be implemented in a new tool that will improve the consistency and responsiveness of services to our users when implemented in 2021.

Work continues on the development of a refactored framework for sequence analysis, called the JD2 (Job Dispatcher 2) for giving access to tools in an easier, FAIR and unrestricted fashion.

Adobe stock

Administration

EMBL-EBI's Administration teams were instrumental in the institute's response to the pandemic. They not only enabled the institute's activity to function with minimal disruption, but also continued to support the development of EMBL-EBI's data resources.

Business continuity

EMBL-EBI's Business Continuity Planning team produced a business continuity plan to help the institute navigate the uncertainties it faced in 2020 because of the coronavirus pandemic. This included the closure of EMBL-EBI buildings, the move to remote working, and the partial reopening of buildings later in the year. By working closely with the EMBL and Wellcome Genome Campus crisis management teams, they developed practical guidance and safety procedures to ensure safe working conditions for all staff working remotely and on campus.

The team met on a regular basis throughout the year, and sent updates and communications to EMBL-EBI staff about the changing situation and the support available. They also completed five staff surveys to identify and address staff needs during the pandemic. With EMBL-EBI staff working remotely during most of 2020, there was a heavy focus on internal communications, to ensure staff were informed and well-supported during the different stages of the pandemic.

EMBL-EBI Administration also continued preparations for Brexit and supported staff and new starters with settled status applications and immigration during this difficult period.

To enable staff to work from home during the COVID-19 pandemic, the Facilities team addressed over 550 homeworking checklists, supplying more than 120 desks and 140 office chairs. The HR team supported the management, staff and their families with the uncertainties and stresses of both remote working and Brexit preparations.

EMBL-EBI Grants and Research Management teams supported £9.5 million of COVID-19 grant applications whilst also securing and managing £1.6 million awarded to compensate for projects disrupted by the pandemic.

Our people

Recruitment continued after a short pause in the spring, and 131 new starters joined EMBL-EBI in 2020, chosen from over 8,500 applicants. The onboarding process was adapted to enable new starters to join their teams remotely. Necessary equipment for new starters was shipped to destinations around the world, including Sri Lanka, Denmark, India and Canada.

Staff in 2020 FTE (full-time equivalent)

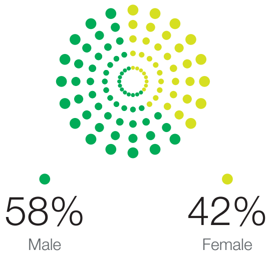
697

618
Staff members

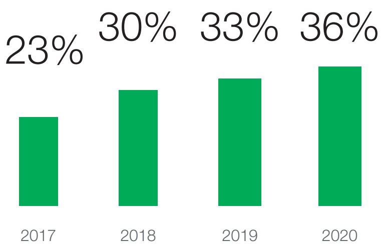
42
Postdoctoral fellows

37
Predoctoral fellows

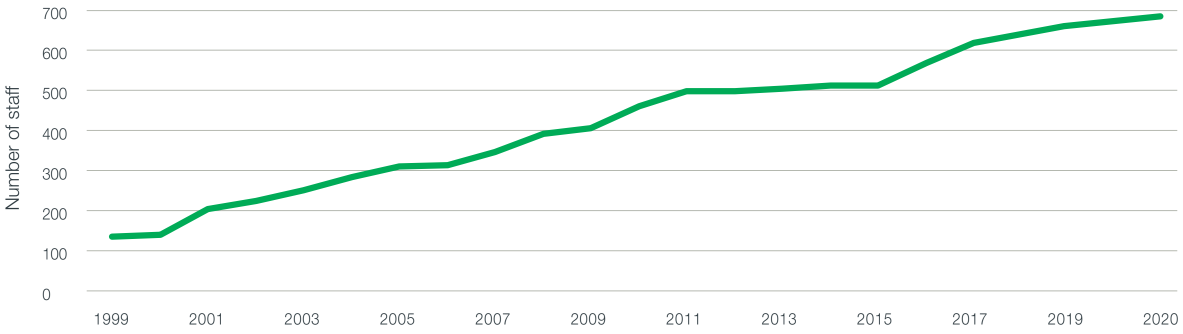
Gender distribution of staff in 2020



Senior roles held by women



Staff growth at EMBL-EBI, 1999 - 2020



New leadership

Ewan Birney was appointed Deputy Director General of EMBL. He continues as EMBL-EBI Director and as a Research Group Leader.

Tim Dyce joined EMBL-EBI as Head of Infrastructure Services, focusing on the smooth running of the institute's data centres, network and storage. With an ever increasing amount of data to process, store and transmit, these foundational services remain crucial for the institute and the wider scientific community.

Andrew Leach took up the role of Head of Industry Partnerships at EMBL-EBI. He continues as Head of Chemical Biology and as a Research Group Leader.

Geetika Malhotra is the new Head of Web Development, and her team provides a central source of web design and development, as well as user experience expertise for the institute.

John Marioni was named the new Head of Research at EMBL-EBI, and his role covers the development and implementation of a new vision for the institute's research efforts.

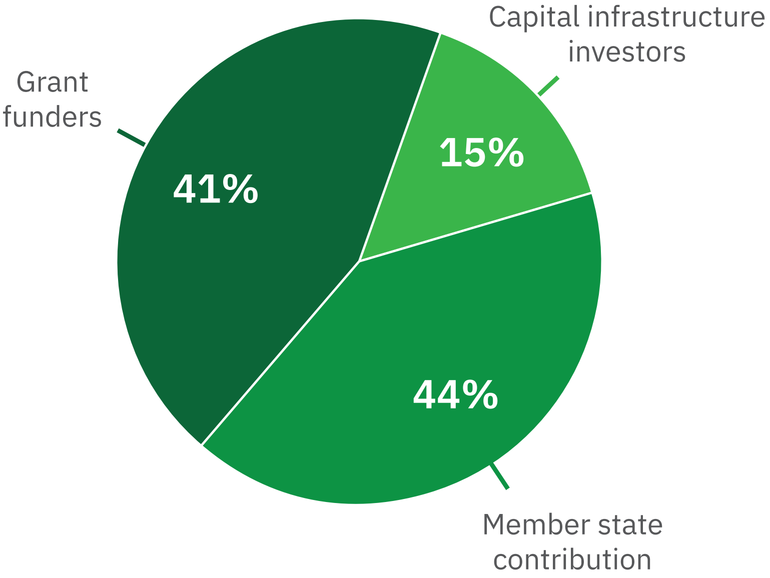
Ellie McDonagh joined Open Targets as Informatics Science Director. She focuses on bringing new tools and data into the Open Targets Platform and Genetics Portal. She also works with new data providers and collaborators to create innovative ways to aggregate and visualise data in a useful way.

Gemma Wood is EMBL-EBI's new Head of Communications. She coordinates public engagement, internal and external communications for the institute.

Financial figures

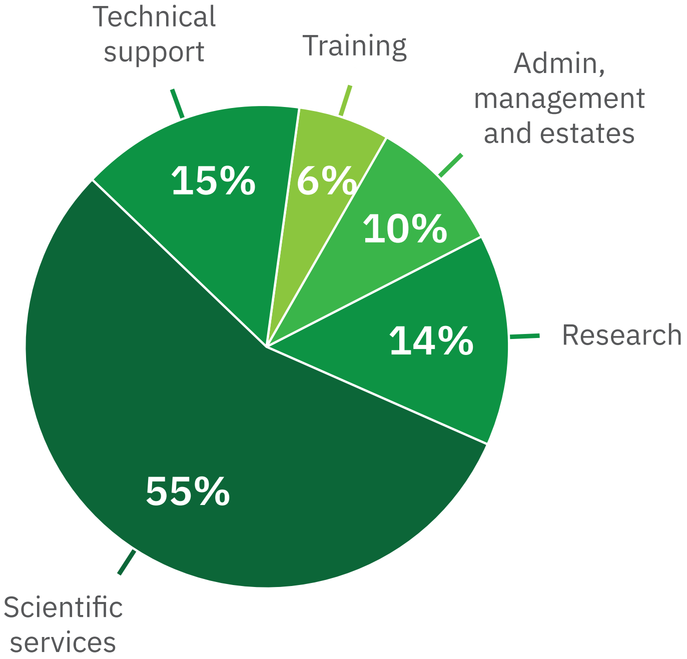
We are grateful for the continued support of our member states and other funding bodies, which in 2020 helped us maintain our data resources, conduct vital research and training, as well as respond to the requirements of the international scientific community.

EMBL-EBI receives its funding through three main funding channels. The operating expenditure is funded by either EMBL member state contributions (44%) or from external funding bodies via grant awards (41%). EMBL-EBI also receives significant capital awards for investment in buildings and data infrastructure (15%).



Operating expenditure

The total operating expenditure of EMBL-EBI in 2020 was €82.7m. Scientific Services accounted for 55% of the expenditure in support of EMBL-EBI's data resources, with a further 14% allocated to research and 6% on external training. Approximately 15% of costs were on technical support which maintains and develops the technical and IT infrastructure, and 10% on management, administration and estate costs.



Capital investment

In 2020, the establishment of EMBL-EBI's Data Infrastructure for Life Sciences programme began. The four-year programme oversees £70 million of capital investment, of which £44.5 million from the UK Research and Innovation (UKRI) Strategic Priorities Fund. This will enable the expansion of storage, compute and networking facilities as the data resource demand continues to grow, and the establishment of a central, open data resource for biological images – the BioImage Archive.

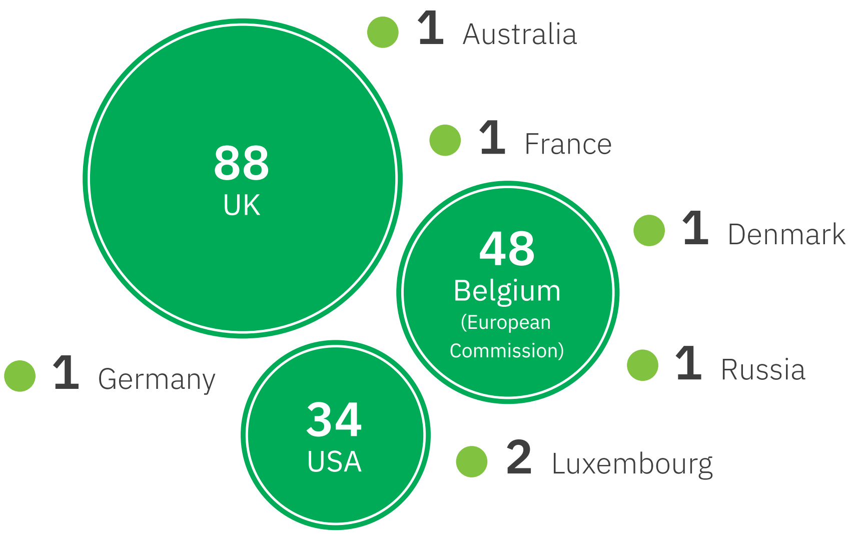
Further UK Government, Wellcome and Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) capital investment has been awarded for office accommodation on the Wellcome Genome Campus. Modular accommodation for 100 EMBL-EBI staff was ready for occupation in spring 2021, and long-term office space for 250 staff and collaborators will be ready by 2024.

Expenditure relating to these capital investments amounted to €14.7m during the 2020 EMBL financial year.

Strategic partners and funders

EMBL-EBI works closely with strategic partners and funders from across the globe. Alongside funding from EMBL member states, in 2020 we had 27 funders providing support for 177 projects. This included six new funders: Connective Tissue Oncology Society (USA), Engineering and Physical Sciences Research Council (UK), Economic and Social Research Council (UK), NF1 Research Initiative (USA), National Institute for Health Research (UK), Sarcoma Foundation of America (USA).

We are extremely grateful to all of our funders for their continuous investment and support. See the full acknowledgement list on page 51.



Number of grants received by EMBL-EBI in 2020 by funding body location

Our governance

EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an inter-governmental organisation with over 20 member states, two associate member states and two prospect member states. EMBL is led by a Director General, Edith Heard, appointed by the EMBL Council.

The EMBL Council is composed of representatives from all member states of the Laboratory and determines its policy in scientific, technical and administrative matters by giving guidelines to the Director General. The Council ensures that the financial requirements of the agreement establishing EMBL, and of the agreements with host member states are complied with.

In 2020, EMBL-EBI was led by joint Directors Rolf Apweiler and Ewan Birney, with the support of two Associate Directors of Services, Paul Flicek and Johanna McEntyre, the Head of Research, John Marioni, and the Head of Administration and Operations, Rachel Curran.

Our funders

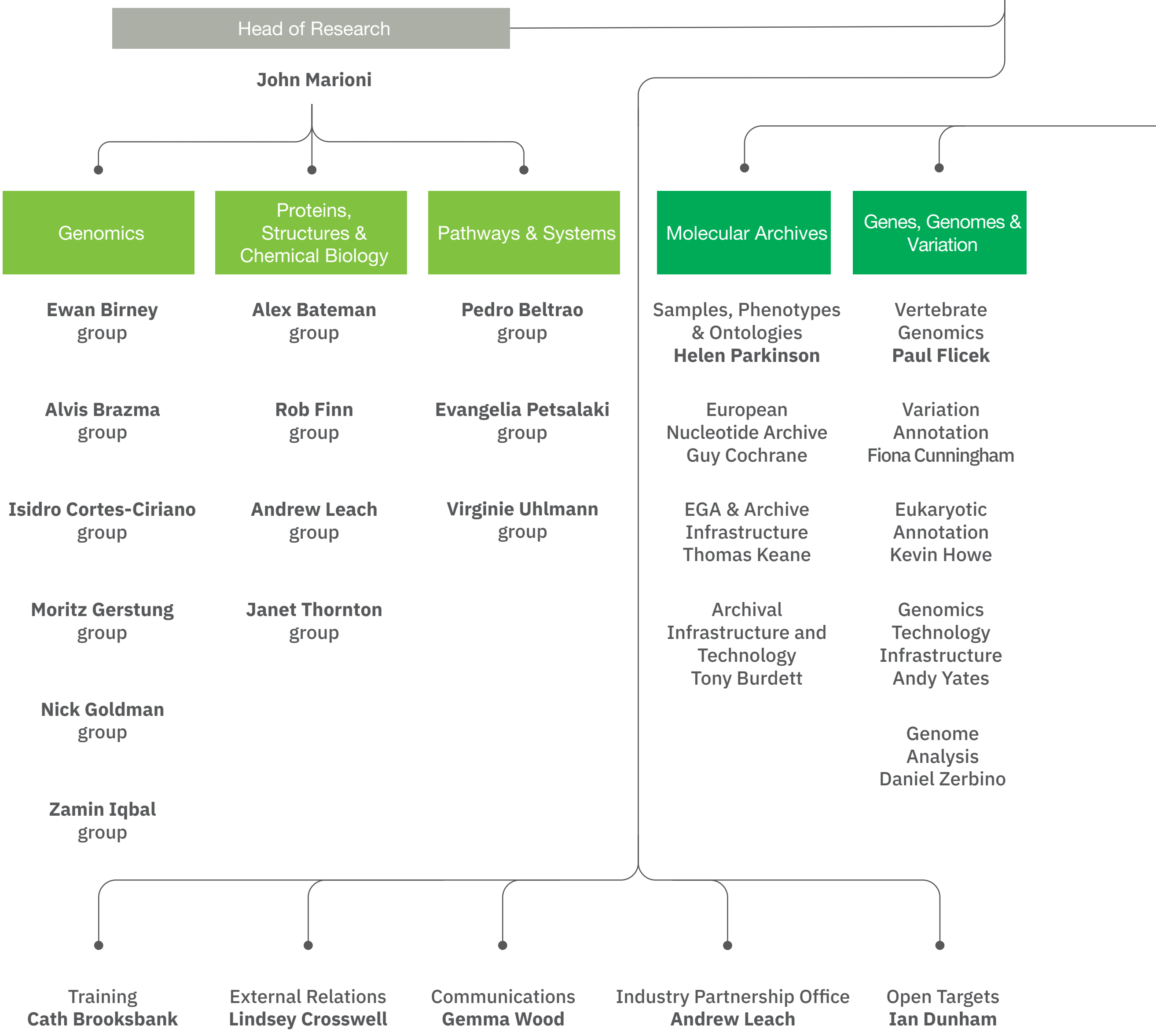
- Alzheimer's Research UK
- Biotechnology and Biological Sciences Research Council
- British Council
- Cancer Research UK
- Connective Tissue Oncology Society
- Chan Zuckerberg Institute
- European Commission
- European Molecular Biology Organization
- Engineering and Physical Sciences Research Council
- Economic and Social Research Council
- Foundation for the National Institutes of Health
- Fonds National de la Recherche
- French Embassy
- Bill & Melinda Gates Foundation
- Gordon and Betty Moore Foundation
- Medical Research Council
- NF1 Research Initiative
- National Institutes of Health
- National Institute for Health Research
- Novo Nordisk
- National Science Foundation
- Nvidia
- Russian Foundation for Basic Research
- Sarcoma Foundation of America
- Save the Tasmanian Devil Program
- The Genetics Society
- Wellcome

List of acronyms

- **API**
  Application Programming Interface
- **AIT**
  Archival Infrastructure and Technology
- **BBSRC**
  Biotechnology and Biological Sciences Research Council
- **COG-UK**
  COVID-19 Genomics UK Consortium
- **DCP**
  Data Coordination Platform
- **DSP**
  Data Submission Portal
- **CINECA**
  Common Infrastructure for Cohorts in Europe, Canada and Africa
- **CRISPR**
  Clustered regularly interspaced short palindromic repeats
- **DNA**
  Deoxyribonucleic Acid
- **DToL**
  Darwin Tree of Life
- **EGA**
  European Genome-phenome Archive
- **EM**
  Electron Microscopy
- **EMBL**
  European Molecular Biology Laboratory
- **EMBL-EBI**
  EMBL's European Bioinformatics Institute
- **EMDB**
  Electron Microscopy Data Bank
- **EMPIAR**
  Electron Microscopy Public Image Archive
- **ENA**
  European Nucleotide Archive
- **ERC**
  European Research Council
- **EVA**
  European Variation Archive
- **FAANG**
  Functional Annotation of Animal Genomes
- **FTE**
  Full-Time Equivalent
- **GA4GH**
  Global Alliance for Genomics and Health
- **GWAS**
  Genome-wide association study
- **GB**
  Gigabyte
- **GCRF**
  Global Challenges Research Fund

- **GWAS**
  Genome-Wide Association Studies
- **HCA**
  Human Cell Atlas
- **IP**
  Internet Protocol
- **IMEx**
  International Molecular Exchange Consortium
- **INSERM**
  French National Institute of Health and Medical Research
- **IT**
  Information Technology
- **LFCF**
  Large Facilities Capital Fund
- **LMIC**
  Low-to-Middle Income Countries
- **MANE**
  Matched Annotation from the NCBI and EMBL-EBI
- **MIT**
  Massachusetts Institute of Technology
- **NCBI**
  National Centre for Biotechnology Information
- **NMR**
  Nuclear magnetic resonance
- **OCT**
  Optical coherence tomography
- **PB**
  Petabytes
- **PDBe**
  Protein Data Bank in Europe
- **PDBe-KB**
  PDBe-Knowledge Base
- **QTL**
  Quantitative trait loci
- **RNA**
  Ribonucleic Acid
- **SPF**
  Strategic Priorities Fund
- **TB**
  Terabyte
- **UKRI**
  UK Research and Innovation
- **USCS**
  University of California Santa Cruz
- **UX**
  User Experience

EMBL-EBI leadership



Head of Research

John Marioni

Genomics

Ewan Birney group

Alvis Brazma group

Isidro Cortes-Ciriano group

Moritz Gerstung group

Nick Goldman group

Zamin Iqbal group

Proteins, Structures & Chemical Biology

Alex Bateman group

Rob Finn group

Andrew Leach group

Janet Thornton group

Pathways & Systems

Pedro Beltrao group

Evangelia Petsalaki group

Virginie Uhlmann group

Molecular Archives

Samples, Phenotypes & Ontologies
Helen Parkinson

European Nucleotide Archive
Guy Cochrane

EGA & Archive Infrastructure
Thomas Keane

Archival Infrastructure and Technology
Tony Burdett

Genes, Genomes & Variation

Vertebrate Genomics
Paul Flicek

Variation Annotation
Fiona Cunningham

Eukaryotic Annotation
Kevin Howe

Genomics Technology Infrastructure
Andy Yates

Genome Analysis
Daniel Zerbino

Training
Cath Brooksbank

External Relations
Lindsey Crosswell

Communications
Gemma Wood

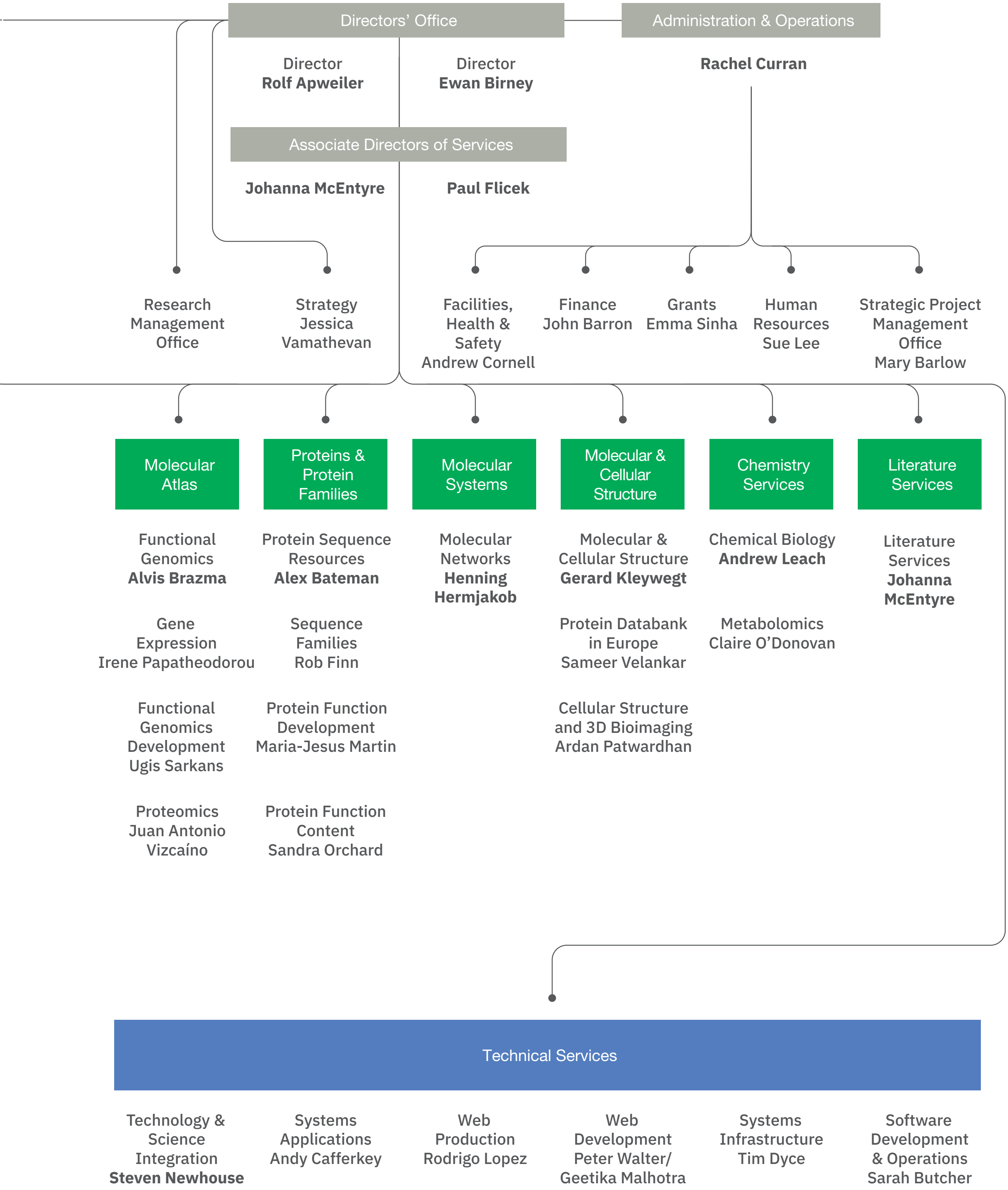
Industry Partnership Office
Andrew Leach

Open Targets
Ian Dunham

🟩 RESEARCH GROUPS

🟩 SERVICE TEAMS

🟦 TECHNICAL SERVICES



Directors' Office

Director
Rolf Apweiler

Director
Ewan Birney

Administration & Operations

Rachel Curran

Associate Directors of Services

Johanna McEntyre

Paul Flicek

Research Management Office

Strategy
Jessica Vamathevan

Facilities, Health & Safety
Andrew Cornell

Finance
John Barron

Grants
Emma Sinha

Human Resources
Sue Lee

Strategic Project Management Office
Mary Barlow

Molecular Atlas

Functional Genomics
Alvis Brazma

Gene Expression
Irene Papatheodorou

Functional Genomics Development
Ugis Sarkans

Proteomics
Juan Antonio Vizcaíno

Proteins & Protein Families

Protein Sequence Resources
Alex Bateman

Sequence Families
Rob Finn

Protein Function Development
Maria-Jesus Martin

Protein Function Content
Sandra Orchard

Molecular Systems

Molecular Networks
Henning Hermjakob

Molecular & Cellular Structure

Molecular & Cellular Structure
Gerard Kleywegt

Protein Databank in Europe
Sameer Velankar

Cellular Structure and 3D Bioimaging
Ardan Patwardhan

Chemistry Services

Chemical Biology
Andrew Leach

Metabolomics
Claire O'Donovan

Literature Services

Literature Services
Johanna McEntyre

Technical Services

Technology & Science Integration
Steven Newhouse

Systems Applications
Andy Cafferkey

Web Production
Rodrigo Lopez

Web Development
Peter Walter/
Geetika Malhotra

Systems Infrastructure
Tim Dyce

Software Development & Operations
Sarah Butcher

European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom

🌐 www.ebi.ac.uk
☎ +44 (0)1223 494 444
✉ comms@ebi.ac.uk

@emblebi
/EMBLEBI
/EBImedia
/company/ebi/

EMBL-EBI is a part of the European Molecular Biology Laboratory.

A digital version of this publication is available on
www.ebi.ac.uk/about/our-impact

EMBL member states and associate member states: Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom, Argentina, Australia

Prospect member states: Estonia, Latvia