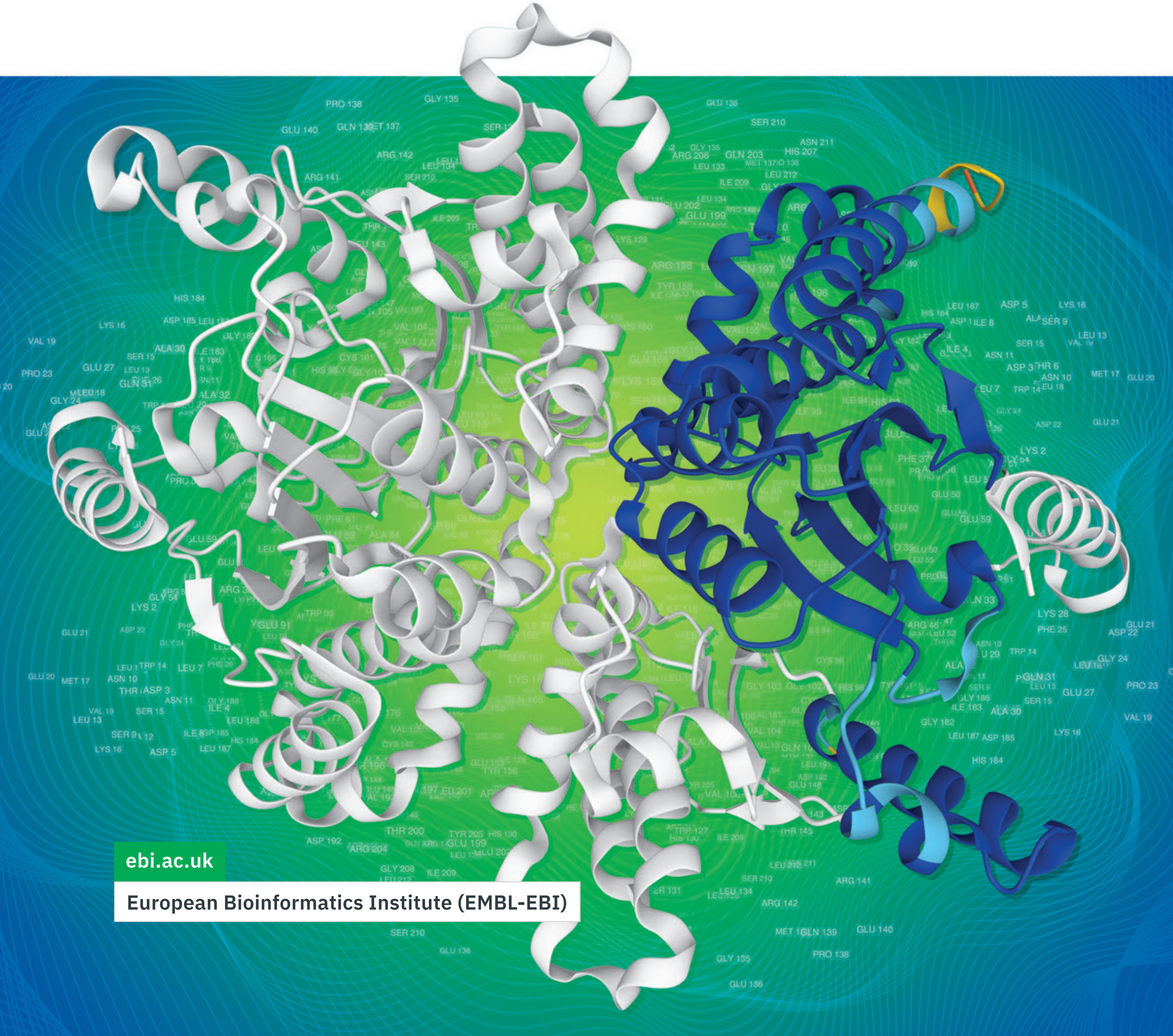
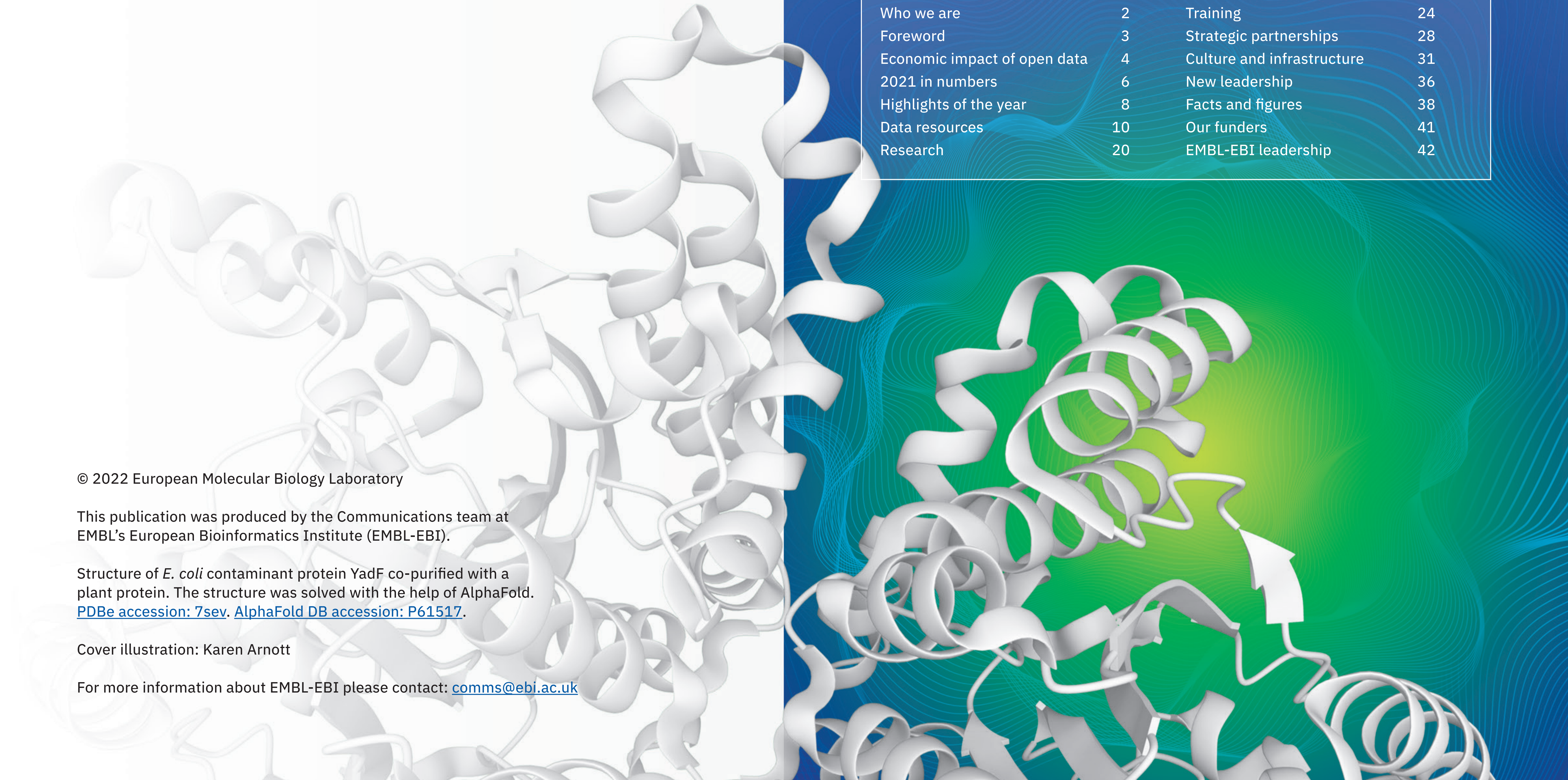
EMBL-EBI

Highlights 2021



ebi.ac.uk

European Bioinformatics Institute (EMBL-EBI)



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

Structure of *E. coli* contaminant protein YadF co-purified with a plant protein. The structure was solved with the help of AlphaFold. PDBe accession: 7sev. AlphaFold DB accession: P61517.

Cover illustration: Karen Arnott

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

Contents

Who we are

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

We are one of the six sites of the European Molecular Biology Laboratory (EMBL), Europe's leading life sciences organisation. EMBL conducts world-class life sciences research, provides training and state-of-the-art research infrastructures. EMBL is an intergovernmental organisation with 27 member states. EMBL's latest scientific programme – Molecules to Ecosystems – seeks to better understand life in context, from molecules to ecosystems.

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.

Foreword

The COVID-19 pandemic has shown without a doubt what the collaboration between scientists, governments and industry can achieve.

Innovation happened at pace – from vaccine development to novel epidemiological models and drug repurposing – highlighting that fundamental research can have the most surprising and useful applications. The pandemic also made it clear that we are living in a world where bioinformatics and data science are crucial to addressing global challenges, from health and disease to food security, and biodiversity conservation.

EMBL's new scientific programme beginning in 2022, called Molecules to Ecosystems, is an ambitious endeavour to bring biology out of the lab and study life in context – from complex ecosystems in oceans to the bacterial communities in the human gut.

The result will be fundamental research that expands what we know about life on earth and that sparks new ways of addressing the challenges our world is facing. Bioinformatics and data science are a vital part of this programme, which sets the stage for innovative multidisciplinary collaborations.

We would like to thank our funders, including EMBL's member states, UK Research and Innovation, the National Institutes of Health, the European Commission and Wellcome, among others, for their continued support and confidence in our work.

We're looking forward to working with existing and new collaborators to continue developing our research efforts and the data infrastructure that underpins scientific discoveries worldwide.

Ewan Birney and Rolf Apweiler
EMBL-EBI Directors

Economic impact of open data

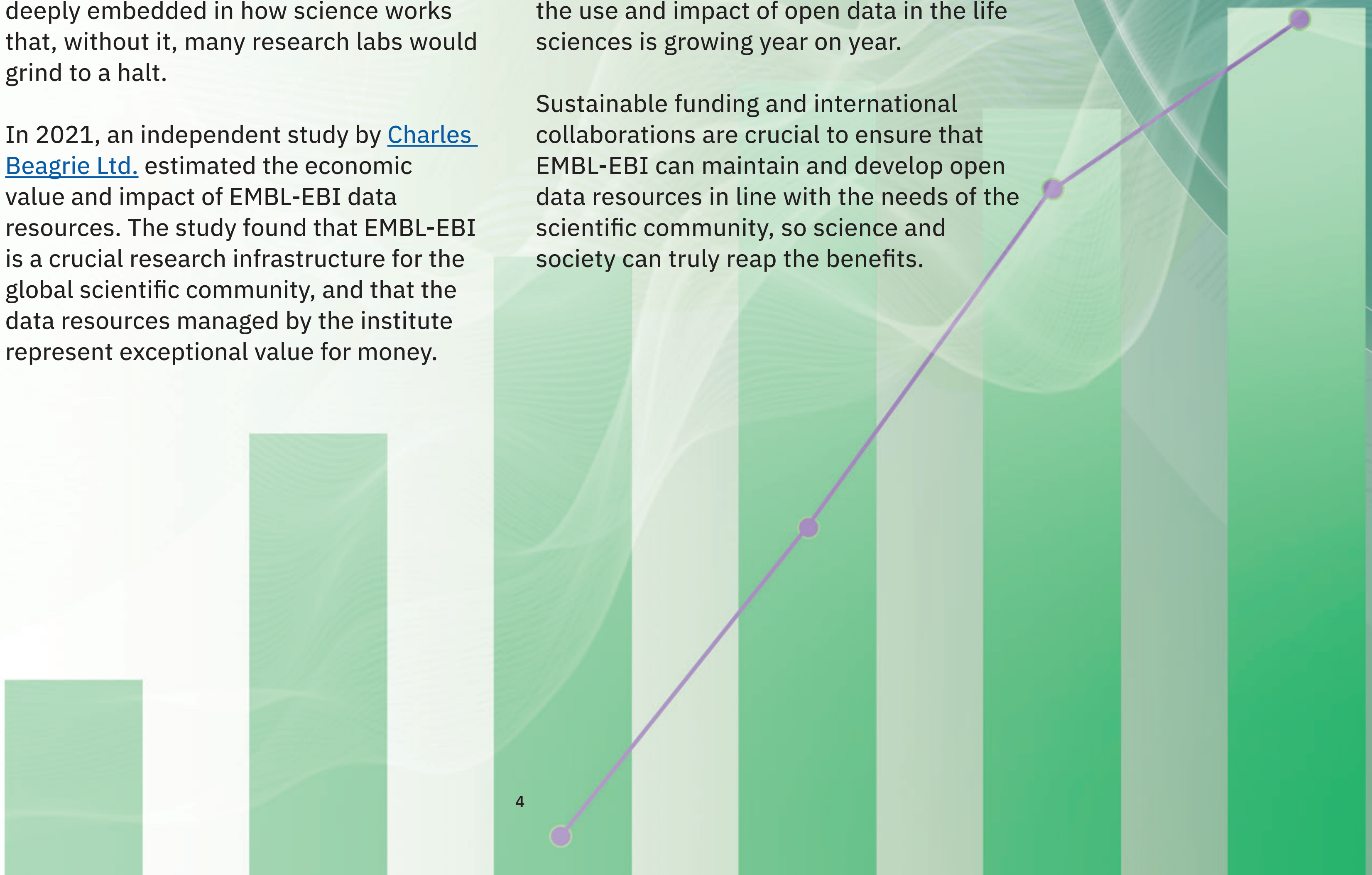
“EMBL-EBI managed data resources represent excellent value for money, and open data is critical for life sciences.”
Neil Beagrie, Director of Charles Beagrie Ltd.

Open data underpin the life sciences, and the open data resources EMBL-EBI manages are essential to the work of millions of scientists around the world. This profoundly collaborative, global data ecosystem is so deeply embedded in how science works that, without it, many research labs would grind to a halt.

In 2021, an independent study by Charles Beagrie Ltd. estimated the economic value and impact of EMBL-EBI data resources. The study found that EMBL-EBI is a crucial research infrastructure for the global scientific community, and that the data resources managed by the institute represent exceptional value for money.

The study used multiple approaches to assess economic value, including an online survey which received more than 4,900 responses. This was the largest study on open data in recent years, and revealed that the use and impact of open data in the life sciences is growing year on year.

Sustainable funding and international collaborations are crucial to ensure that EMBL-EBI can maintain and develop open data resources in line with the needs of the scientific community, so science and society can truly reap the benefits.



“EMBL-EBI resources are the lifeblood of my research group. Without these services, a full two-thirds of my group’s research efforts would suffer dramatically.”
Survey respondent, Belgium

“[Without EMBL-EBI resources] my work would be severely affected. The resources of EMBL and EMBL-EBI allowed me to work even during the lockdown.”
Survey respondent, Italy

The value of EMBL-EBI data resources

Use value

Researchers spend 140 million hours/year using EMBL-EBI data resources. That’s equivalent to **£5.5 billion**.

Critical infrastructure

58% of respondents said they couldn’t have collected the last dataset they used or obtained it elsewhere.

Return on investment

EMBL-EBI data resources contribute to the realisation of future research impacts worth **£2.2 billion** annually.

Read the report

2021 in numbers

837
members
of staff*

from
78
countries

**includes fellows, supernumeraries, trainees and visitors*

497,000
unique IP addresses
accessed our
online training

189
active grants

147
collaborative grants
with researchers in
52 countries from
644 institutes

3,900
people participated
in our 46 public
engagement
activities

over
40
open data
resources

256
journal papers
published

over
20
petabytes of data
deposited into
EMBl-EBI resources

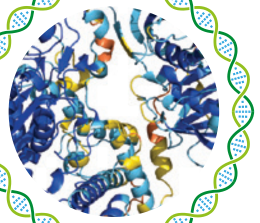
from
41
million unique
IP addresses

107
million requests to
our websites on an
average day

Highlights of the year

AlphaFold Database for protein structure predictions launches (page 12)

3D-Beacons network gathers publicly-available protein structure data in one place (page 17)



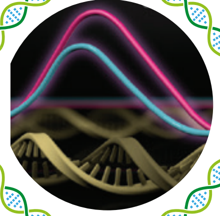
COVID-19 Data Platform continues to grow (page 14)

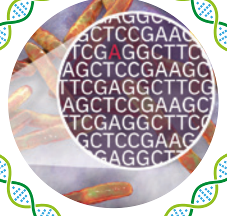
Gerstung group tracks SARS-CoV-2 lineage spread (page 20)

Marioni group analyses differences in immune responses to COVID-19 (page 20)

Leach group identifies drugs that could be repurposed for COVID-19 treatment (page 20)

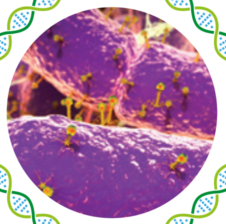
PGS Catalog, an open database for polygenic risk scores, launches (page 17)





Darwin Tree of Life Data Portal, which will host genomic data for thousands of species, launches (page 19)

Iqbal group and CRyPTIC project launch world’s largest database for drug resistance in tuberculosis (page 21)

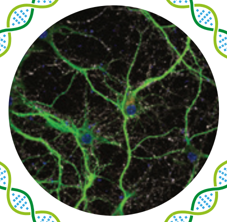


Thornton group uses UK Biobank data to investigate genetics of age-related diseases (page 21)

Finn group identifies 140,000 virus species in the human gut and hundreds of new bacteria species on human skin (page 22)

Data Use Ontology for streamlining usage of biomedical data in research launches (page 29)

REMBI metadata guidelines for bioimages published to improve microscopy data reuse and analysis (page 16)



Proteomics data usage and interoperability shows potential for biomedical research (page 18)

CABANA project for increasing bioinformatics capacity in Latin America wraps up (page 25)

Training Competency Hub increases focus on FAIR training (page 24)

First Bioinformatics for BioBusiness event for small and medium companies takes place (page 28)

Strategic partnerships with cloud providers commence (page 34)

Transition to new EMBL visual framework (page 34)

Data resources

EMBL-EBI manages the world's most comprehensive suite of open data resources for the life sciences. Our 40 data resources and dozens of tools span genetics, genomics, proteins, chemistry, literature data and more.

Chemicals, molecules and drug discovery

ChEBI
ChEMBL
MetaboLights
Open Targets
SureChEMBL

Genes, genomes and RNA

ArrayExpress
Ensembl
European Nucleotide Archive
Expression Atlas
HGNC
MGnify
Rfam
RNAcentral
VectorBase
WormBase

Proteins

AlphaFold DB
Enzyme Portal
InterPro
PDBe
PDBe-KB
Pfam
PRoteomics IDEntification
UniProt
UniProtKB

Imaging and cellular structure

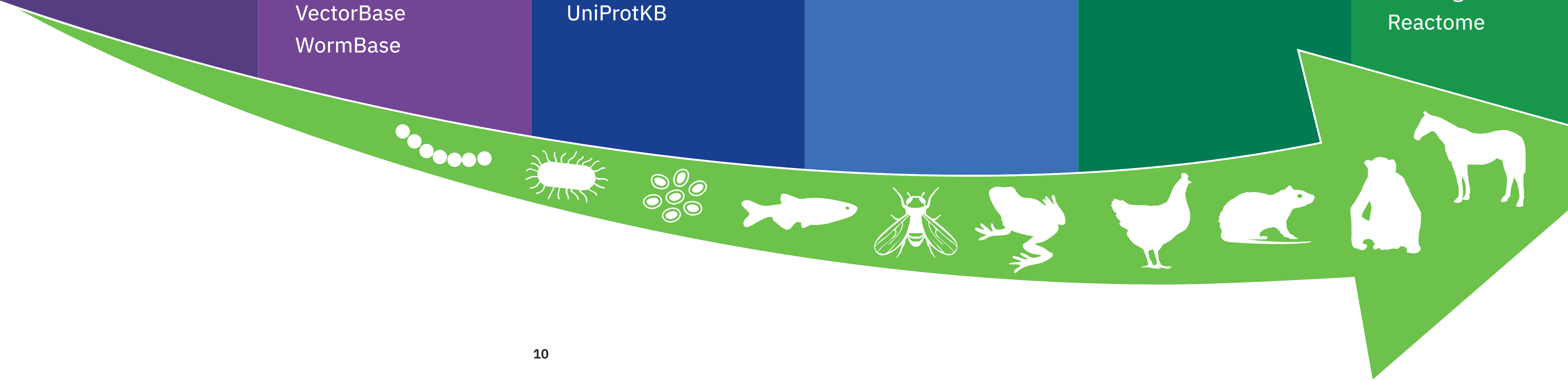
BioImage Archive
Electron Microscopy Data Bank
Electron Microscopy Public Image Archive

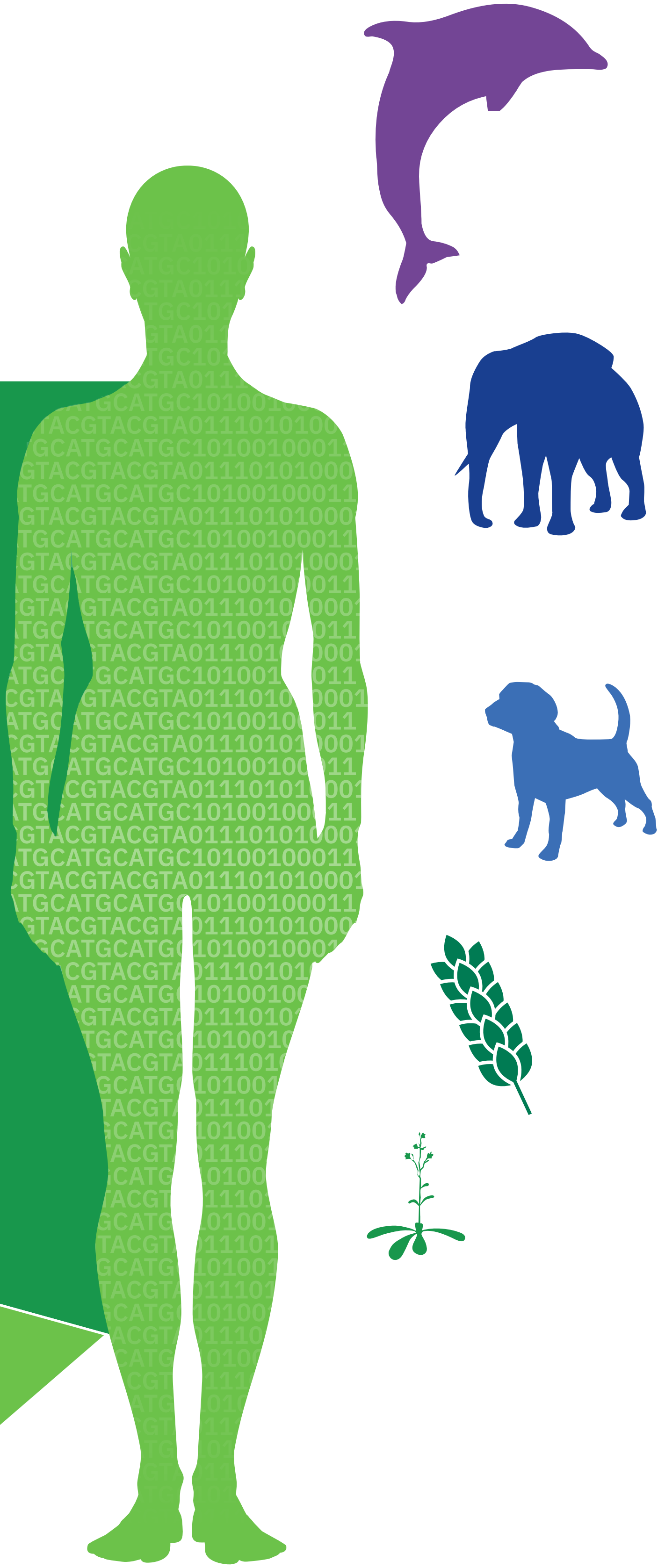
Genetic variation and disease data

COVID-19 Data Platform
European Genome-phenome Archive
European Variation Archive
Mouse informatics

Literature and knowledge management

BioModels
BioSamples
BioStudies
Complex Portal
Europe PMC
GWAS Catalog
IntAct
OmicsDI
Ontologies
Reactome





AlphaFold ushers in a new era in biology

A protein's 3D structure is closely related to its function, so knowing its shape is essential for understanding how it works, and why sometimes things go wrong. But predicting a protein's structure is no easy task. Since the 1970s, this has been one of the biggest unanswered questions in biology, with many experimental and computational scientists dedicating their careers to it.

This all changed when artificial intelligence (AI) company DeepMind released its ground-breaking AlphaFold system, which uses deep learning to predict protein structures more accurately than any previous computational methods.

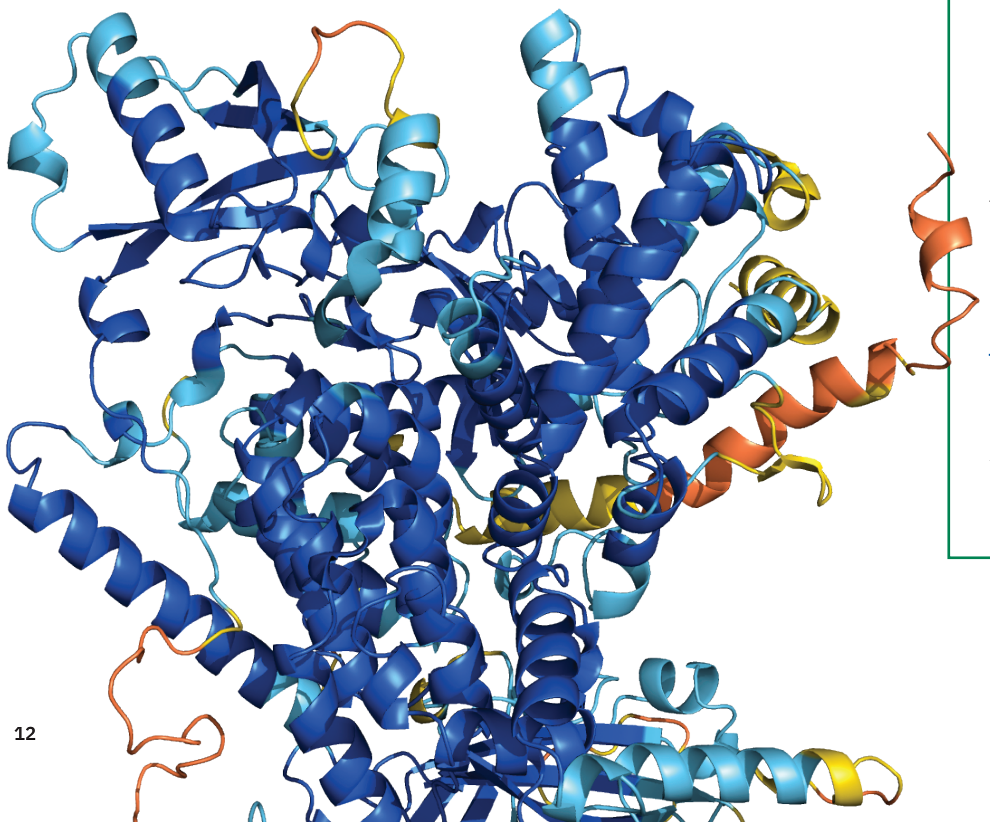
The real gains started when DeepMind teamed up with EMBL-EBI to make its predictions freely and openly available to all through the AlphaFold Protein Structure Database.

Hailed by Forbes as "the most important achievement in AI – ever", AlphaFold is a perfect example of the virtuous cycle of open data. The system was 'trained' on open data generated by experimental scientists. These data are freely available in public databases, such as those co-hosted by EMBL-EBI (PDB, UniProt and MGnify).

By making AlphaFold's predictions freely available to all, DeepMind opened up research avenues for scientists everywhere in areas such as drug discovery, neglected diseases, improving crops, designing new enzymes with new functions like processing waste or degrading plastics.

Building on decades of expertise in making the world's biological data available, EMBL-EBI is working with DeepMind to grow the database to 130 million predictions. We're dedicated to ensuring the predictions are Findable, Accessible, Interoperable and Reproducible (FAIR) so researchers everywhere can exploit them to drive discovery.

> ***"With this resource freely and openly available, the scientific community will be able to draw on collective knowledge to accelerate discovery, ushering in a new era for AI-enabled biology."*** *– Sir Paul Nurse, Nobel Laureate for Physiology or Medicine, Director of the Francis Crick Institute*



Structure prediction for Trypanosoma cruzi protein. The parasite causes Chagas disease. Credit: AlphaFold. Accession: Q4DCH1

ALPHAFOLD DATABASE IN 2021

992,000	48	365,000	15,000
protein structure predictions	species	unique users	views of first AlphaFold webinar

SAMEER VELANKAR

AlphaFold database – a look behind the scenes

"AlphaFold has not only reinvigorated structural biology but will have far reaching impact on life science research in the coming decades," said Sameer Velankar, Team Leader at EMBL-EBI.

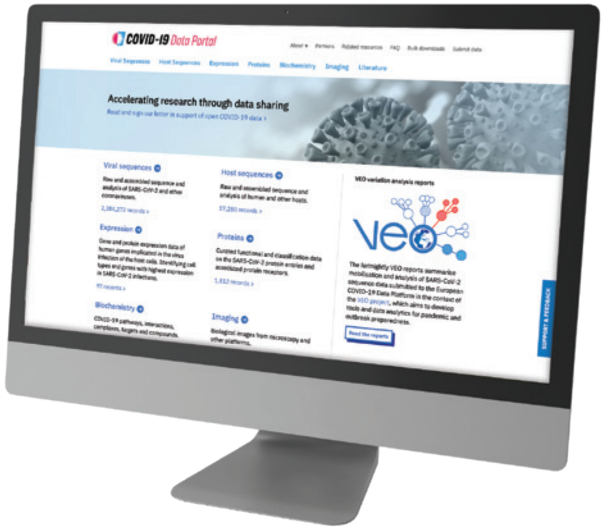
Sameer was one of the people who made the collaboration between EMBL-EBI and DeepMind a success. Thanks to the dedication of Sameer and his team, almost one million protein structure predictions have already been made accessible through the AlphaFold Protein Structure Database.

Sameer studied structural biology at the Indian Institute of Science in Bengaluru, India before doing a postdoctoral degree in Oxford, UK. At EMBL-EBI he leads the Protein Data Bank in Europe (PDBe) team.

Sameer is an advocate of open data and collaborative science. "It's wonderful to see scientists building on AlphaFold and using its predictions to advance their work. I expect many new developments in the coming months and years."

Tracking the pandemic in real time

The second year of the COVID-19 pandemic signalled a shift to using open data for monitoring the spread of the viral lineages circulating worldwide. As an established source of SARS-CoV-2 open data, EMBL-EBI's COVID-19 Data Platform continued to increase in size and improve the depth of analysis it facilitates.



COVID-19 Data Portal

About ▾ Partners Related resources FAQ Bulk downloads Submit data

Viral Sequences Host Sequences Expression Proteins Biochemistry Imaging Literature

Accelerating research through data sharing

Read and sign our letter in support of open COVID-19 data ›

Viral sequences

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

2,384,272 records ›

Host sequences

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

17,280 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.

97 records ›

Proteins

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

1,612 records ›

Biochemistry

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

Imaging

Biological images from microscopy and other platforms.

VEO variation analysis reports

The fortnightly VEO reports summarise mobilisation and analysis of SARS-CoV-2 sequence data submitted to the European COVID-19 Data Platform in the context of the VEO project, which aims to develop tools and data analytics for pandemic and outbreak preparedness.

Read the reports

SUPPORT & FEEDBACK

> *"The COVID-19 Data Platform has gone from strength to strength with increased user engagement and new features and tools; we're now developing it for future pandemic preparedness."*
> **Marianna Ventouratou, Platform Manager**

COVID-19 DATA PORTAL PROVIDES OPEN ACCESS TO 11 MILLION RECORDS

- Viral and host genomic sequences
- Gene expression
- Protein structure and function
- Biological pathways and interactions
- Drug targets and compounds
- Imaging
- Literature

National and international collaborations

Insights revealed by the COVID-19 Data Platform and research at EMBL-EBI were used to highlight the spread of the Delta and Omicron lineages to national governments in a series of meetings and consultations.

As the benefits of national and centralised open data portals became clear, we continued to support researchers in different countries with SARS-CoV-2 data coordination. In 2021, two more national COVID-19 Data Portals were set up in Turkey and Greece, alongside the existing ones in Estonia, Italy, Japan, Norway, Poland, Slovenia, Spain and Sweden.

Improved search and analysis

We increased the functionalities of the Platform by introducing new search and filtering tools, such as a new API, which enables researchers to adapt search results to their specific use case. For ease of use and reference, the World Health Organisation's nomenclature search was also added to the Platform.

A new tool allows bulk downloads outside of the web browser in a range of formats. This enables researchers to keep their local data in sync with the data available in the portal, making local data analysis faster and more flexible.

Preparing for future pandemics

A new wastewater surveillance working group was set up to improve global efforts to use viral RNA in wastewater samples to track case trends. The working group, comprising specialists from EMBL-EBI and ten European countries, are collaborating to define a systematic approach to data collection and harmonisation, and to propose bioinformatics solutions for wastewater surveillance, which could be crucial in future pandemics.

Building on the COVID-19 Data Portal, the new BeYond-COVID (BY-COVID) initiative, led by ELIXIR, aims to improve European readiness for future pandemics, enhance genomic surveillance and rapid response capabilities. EMBL-EBI is contributing its data coordination expertise to this ambitious initiative set to generate valuable new insights on infectious disease.

COVID-19 DATA PORTAL

6.4 million
web requests since launch

248,000
unique users from 191 countries

11 million
SARS-CoV-2 data records

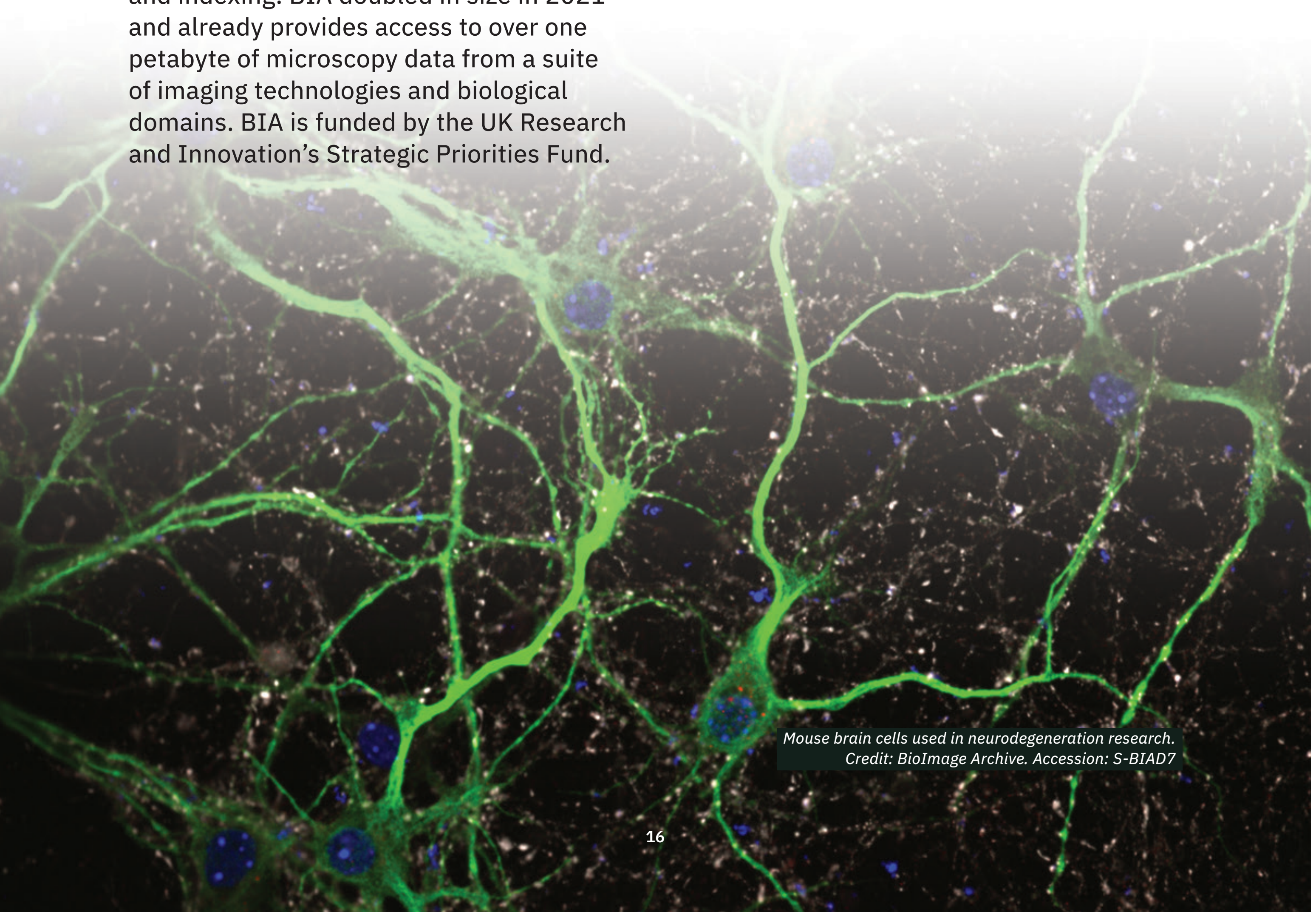
The foundations of a bioimaging revolution

Bioimaging has advanced in leaps and bounds in recent years, giving researchers a closer look at how life works at a cellular, molecular and even atomic level. But maximising the potential of bioimaging data depends on robust data sharing infrastructure and standards.

In 2021, EMBL-EBI and collaborators continued to develop the BioImage Archive (BIA), the open data resource for biological imaging data associated with peer-reviewed publications. BIA anchors an entire ecosystem of related databases, ensuring efficient data storage, crosslinks and indexing. BIA doubled in size in 2021 and already provides access to over one petabyte of microscopy data from a suite of imaging technologies and biological domains. BIA is funded by the UK Research and Innovation's Strategic Priorities Fund.

In an effort to make bioimaging data easier to find and reuse, researchers from over 30 institutions including EMBL-EBI published the Recommended Metadata for Biological Images (REMBI). REMBI provides metadata standards across biological imaging domains supporting FAIR sharing of image data.

As the volume and quality of bioimages increases – and the technologies become more accessible through initiatives such as EMBL's Imaging Centre or Euro-BioImaging – the scientific community needs to cooperate to openly share the data to maximise reuse.



Mouse brain cells used in neurodegeneration research. Credit: BioImage Archive. Accession: S-BIAD7

Assessing inherited disease risk

Polygenic risk scores (PGS) are a new clinical tool for assessing a person's inherited risk for diseases such as Type 2 diabetes or coronary heart disease. The score provides an estimate of an individual's risk for the disease, based on their DNA.

Despite the rise in PGS studies, there are inconsistencies in how the scores are calculated and reported. Lack of adherence to standards has hindered the translation of this important tool into healthcare.

To bridge the gap and support the adoption of PGS scores when appropriate in clinical settings, EMBL-EBI teamed up with the University of Cambridge to create the PGS Catalog, an open database of published polygenic risk scores. The teams also collaborated with NHGRI to create a minimal information framework for PGS, published in the journal Nature[1], which helps promote the validation of scores as well as the reproducibility of the data. This will be key for PGS usage in the clinic.

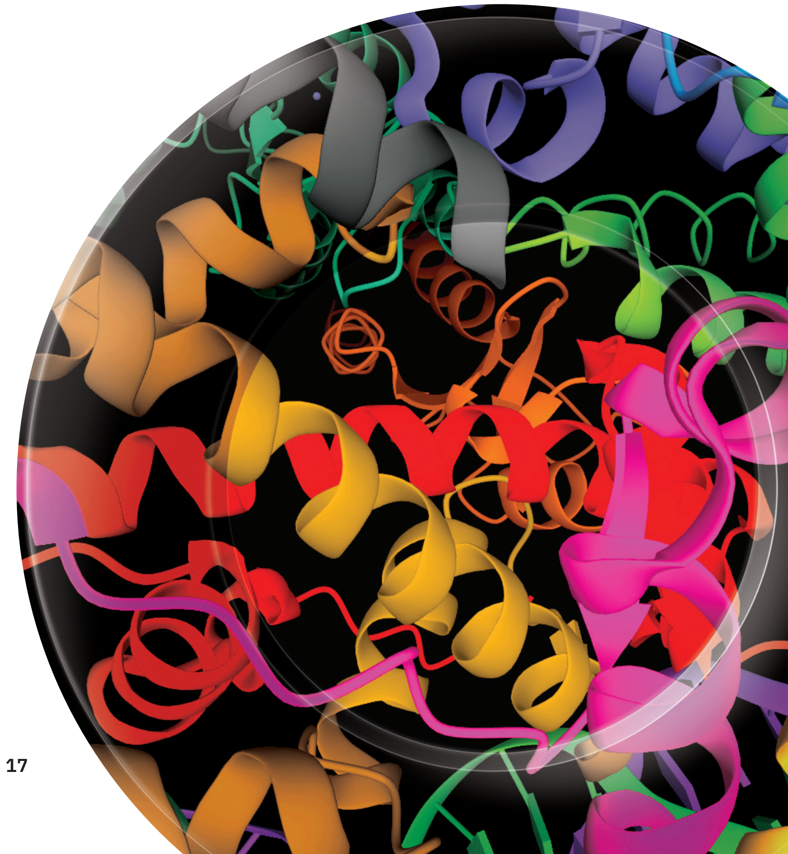
Protein structure data in one place

Another useful development for protein structures was the launch of the 3D-Beacons Network. This community-driven initiative acts as a one-stop shop for experimentally-determined and computationally-predicted protein structures from different providers.

Before the network launched, scientists had to search a range of data resources to find out if any structure models existed for their protein of interest. With the 3D-Beacons Network, researchers can now find all the available models from different providers in a standardised format, with just one search.

The list of participating data providers is impressive, with the Protein Data Bank in Europe (PDBe), AlphaFold DB, SWISS-MODEL, Protein Ensemble Database, Small Angle Scattering Biological Data Bank, Genome3D, and PDBe Knowledge Base (PDBe-KB) joining in. The Central Hub for the Network is located at EMBL-EBI.

Having all protein structure data in one place saves researchers time and makes it much easier to access these models to get valuable insights into health and disease, identifying drug targets and more.



[1] Wand, H. *et al.* (2021). Improving reporting standards for polygenic scores in risk prediction studies. Nature. DOI: 10.1038/s41586-021-03243-6

Unlocking proteomics data

Proteomics is the large-scale study of proteins, and the proteome is a set of proteins produced by an organism or a system. Proteomics helps answer key biological questions about health and disease, but can be challenging to study because the proteome is much more complex than the genome and it isn't constant – it changes from cell to cell and over time.

The PRIDE data resource at EMBL-EBI is the world-leading proteomics data repository, storing more than 80% of all public proteomics data worldwide. It also plays a leading role in the International ProteomeXchange Consortium, which coordinates proteomics data provision worldwide.

As more proteomics data becomes publicly-available, its reuse also increases dramatically, for example for big data methods. EMBL-EBI is focusing on disseminating and integrating proteomics data from PRIDE into added-value data resources, namely UniProt (phosphorylation data), Expression Atlas (protein expression), Ensembl (proteogenomics data) and MGnify (metaproteomics data). The aim is to make proteomics data useful and accessible to all life scientists for a range of applications.

Genomic data for biodiversity

In advance of the UN's Biodiversity Conference (COP15), and with megaprojects such as the Darwin Tree of Life (DToL) and Earth BioGenome Project sequencing hundreds of new species every day, the question about how genomic data are made available to the scientific community is more pertinent than ever.

To address community calls for richer metadata, the International Nucleotide Sequence Database Consortium, which includes EMBL-EBI's European Nucleotide Archive has made spatio-temporal metadata mandatory for new submissions. The move aims to enrich the scientific value of the data, especially for researchers working in biodiversity, ecology and infectious disease.

In 2021, the Darwin Tree of Life Project (DToL), which aims to sequence all eukaryotic species in the UK and Ireland, also made significant progress. Using a new and improved rapid release function, we annotated and released genomes for over 90 species, including many butterflies and moths, in Ensembl and the DToL Data Portal – the gateway to all the data generated by the project. The Portal is developed and maintained by EMBL-EBI, and boasts a new phylogeny browser that allows users to see at a glance what data is available for a particular clade, family, or genus. The DToL dataset, which is openly available to all, is set to uncover significant insights in the field of biodiversity.

MICHAEL PARKIN

Better access to scientific research

Growing up, Michael Parkin enjoyed science so much that for a long time he couldn't choose between physics, chemistry and biology. The scales finally tipped when Michael decided to study Biology and Chemistry at Durham University (UK).

After a brief spell in scientific publishing, Michael joined the Literature Services team at EMBL-EBI, where he sources new content for Europe PMC, the open-access repository containing millions of biomedical research articles.

"I retrained as a Data Scientist while at EMBL-EBI, and now focus on making scientific publications easier to find and explore," Michael explained. "I also make research machine-findable so funders and universities can track what research they've enabled. During the pandemic, my team made COVID-19 preprints and grants accessible through Europe PMC, helping researchers track the latest developments in record time."

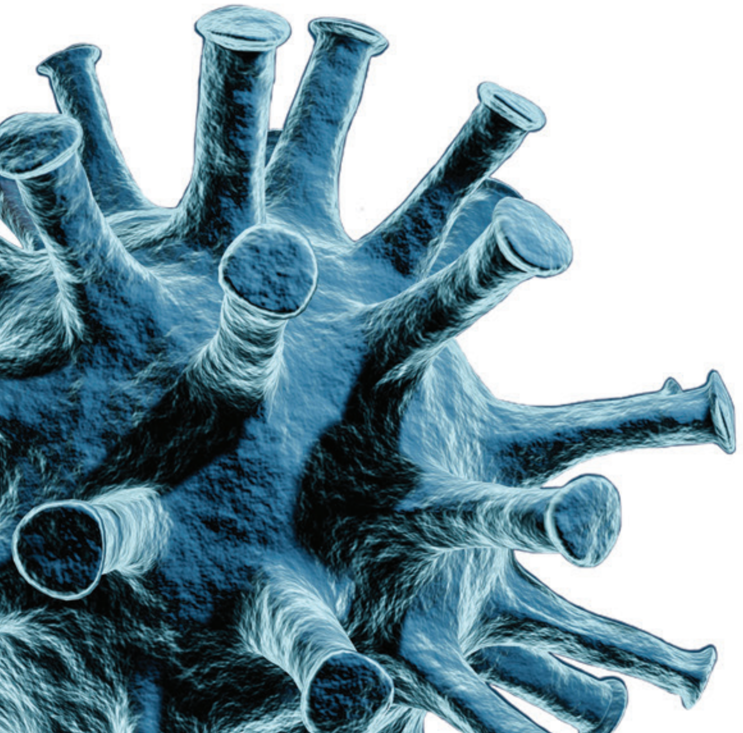
Research

EMBL-EBI's research groups focus on computational biology and making sense of vast, complex datasets. Our researchers work closely with experimental scientists worldwide, increasingly tackling problems of direct significance to medicine and the environment.

Big data analysis shines light on the pandemic

Several EMBL-EBI research groups used bioinformatics to reveal valuable insights on emerging SARS-CoV-2 lineages, differences in immune response and repurposing drugs to treat COVID-19.

The Gerstung group, in collaboration with the Wellcome Sanger Institute, used genomic data to produce the most detailed analysis[2] of how the COVID-19 pandemic developed in England. They tracked over 70 virus lineages and were among the first to sound the alarm regarding the increased transmissibility of Delta and Omicron.



Their work highlighted the value of genomic surveillance combined with big data analysis. Being able to see viral lineages side-by-side and mapped to specific locations helped to understand the spread of the virus and to inform public health measures.

The Marioni group and collaborators used single-cell sequencing to identify differences in the immune response to COVID-19 between symptomatic and asymptomatic people[3]. The research offers a view at an unprecedented resolution into how the body fights the disease, and reveals useful insights to guide drug discovery and treatment.

The Leach group and collaborators used EMBL-EBI's ChEMBL database of small molecules to identify proteins that could be good drug targets for COVID-19 therapies[4]. The work brought together experts from a range of different areas, who quickly responded to the urgent need for treatments.

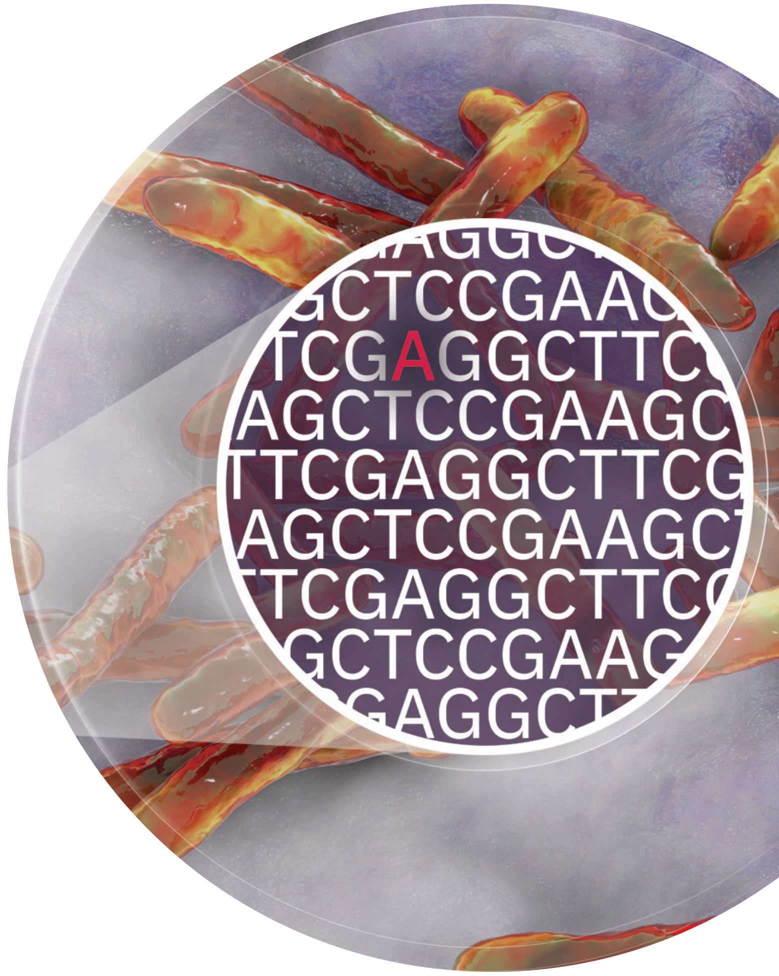
[2] Vöhringer H.S. *et al.* (2021). Genomic reconstruction of the SARS-CoV-2 epidemic in England. Nature. DOI: 10.1038/s41586-021-04069-y

[3] Stephenson E. *et al.* (2021). Single-cell multi-omics analysis of the immune response in COVID-19. Nature Medicine. DOI: 10.1038/s41591-021-01329-2

[4] Gaziano L. *et al.* (2021). Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. Nature Medicine. DOI:10.1038/s41591-021-01310-z

Targeting drug-resistant tuberculosis

As part of the international CRyPTIC consortium, the Iqbal group worked on developing the world's largest data resource for the study of drug resistance in tuberculosis (TB). The work informed the World Health Organisation's new TB resistance catalogue.



CRyPTIC aims to reveal the genetic causes of drug resistant TB to enable the development of DNA-based diagnostics. For this study[5], the consortium used a new experimental assay to measure the level of drug resistance and to scan the genome of over 15,000 *Mycobacterium tuberculosis* strains from across the globe.

The resulting dataset, hosted at EMBL-EBI and openly-available, can be used to probe the genetic causes of drug resistance in TB. The end goal is to apply these insights to DNA-based diagnostics, predicting the drug resistance of a patient's TB strain and enabling personalised treatment.

The genetics of age-related diseases

Researchers in the Thornton group used medical and genetic data from the UK Biobank to investigate links between genetics and age-related diseases. Their work produced the first data-driven classification of 116 age-related diseases, including anaemia, deep vein thrombosis and depression[6].

They revealed genetic links between diseases with the same onset profile, suggesting that they may share a common cause. Their work is a prime example of the unique insights that UK Biobank data combined with bioinformatics expertise can bring to light.

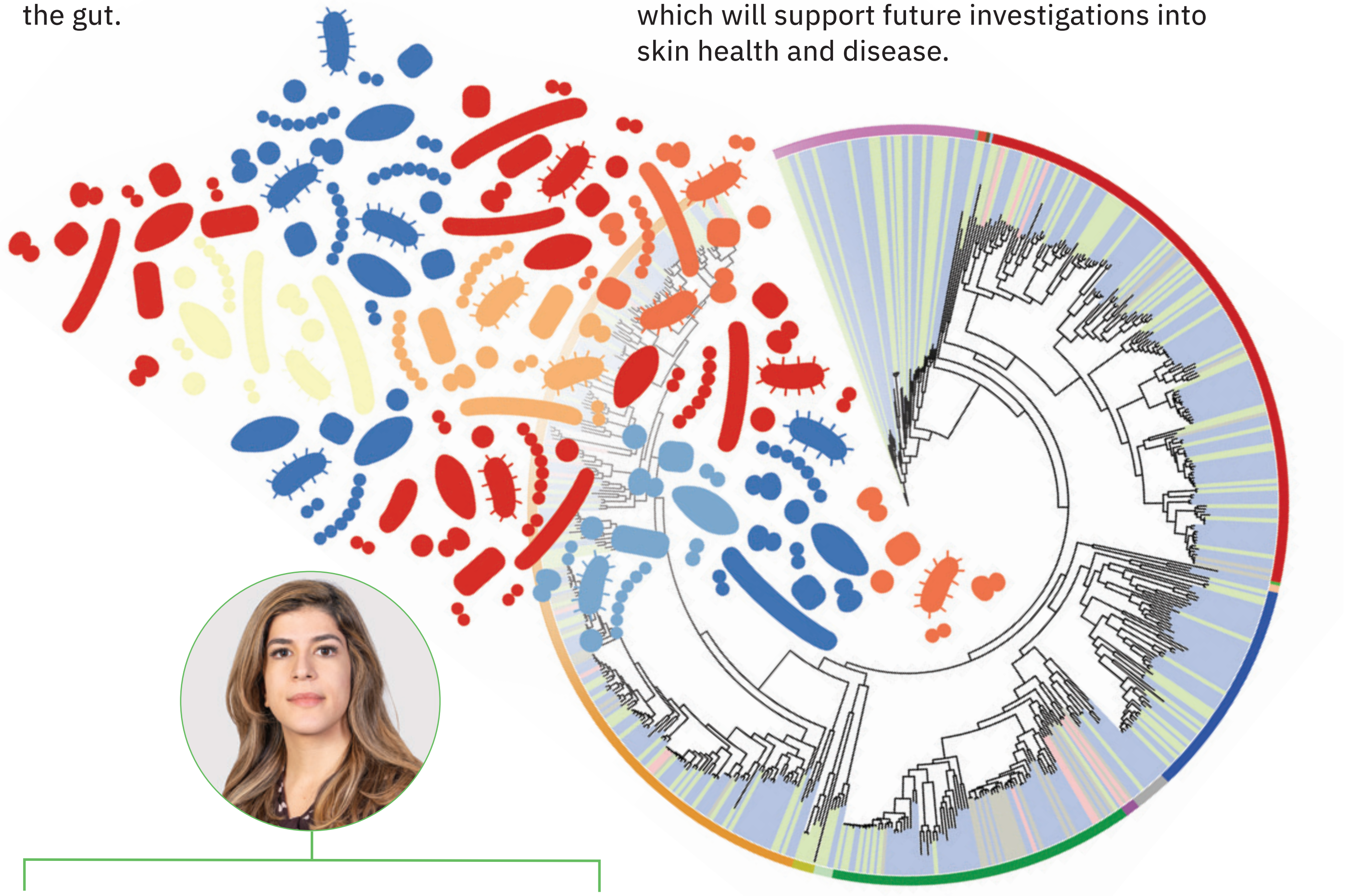
[5] Brankin A. *et al.* (2021). A data compendium of Mycobacterium tuberculosis antibiotic resistance. bioRxiv. DOI: 10.1101/2021.09.14.460274

[6] Donertas M.H. *et al.* (2021). Common genetic associations between age-related diseases. Nature Aging. DOI: 10.1038/s43587-021-00051-5

A closer look at gut viruses and skin bacteria

The Finn research group used metagenomics to reveal new insights about the human gut and skin microbiomes. One paper[7] identified more than 140,000 viruses living in the human gut, including a highly prevalent group that has never been seen before. The work opens up new research avenues to understand how human health and disease are affected by viruses living in the gut.

A second paper[8], in collaboration with the National Institutes of Health, found hundreds of previously unidentified viruses, bacterial and fungal species living on human skin. Combining traditional lab methods with an innovative metagenomic sequencing approach, the researchers developed a comprehensive collection of reference genomes for the human skin microbiome, which will support future investigations into skin health and disease.



"Our future work will aim to understand what the different microbes living on our skin are responsible for."
Sara Kashaf, Predoctoral Fellow

[7] Camarillo-Guerrero L.F. *et al.* (2021). Massive expansion of human gut bacteriophage diversity. Cell. DOI: 10.1016/j.cell.2021.01.029
[8] Kashaf S.S. *et al.* (2021). Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. Nature Microbiology. DOI: 10.1038/s41564-021-01011-w

FRANCESC MUYAS REMOLAR

Using bioinformatics to understand cancer

A liquid biopsy is a non-invasive alternative to surgical biopsies, which enables scientists to characterise tumours using a blood sample. This is just one of the methods Francesc Muyas Remolar, a Postdoctoral Fellow in the Cortes-Ciriano research group, is investigating, in the hope of maximising the potential of genomics in the clinic, with special interest in early cancer detection.

Francesc was born near Valencia and studied genetics, bioinformatics and biomedicine in Spain and Germany. He is using new technologies including Nanopore long-reads and single-cell RNA sequencing data, to characterise cancer genomes, understand how mutations occur, and why patients react differently to treatment.

"We're harnessing knowledge from many data types," said Francesc. "I would like our work to translate into clinical applications that improve cancer diagnosis and treatment.

"Joining a research group at EMBL-EBI is an amazing opportunity to engage with world-leading experts in my field and beyond, and get an overview of the latest science."

Training

EMBL-EBI delivers world-leading training in bioinformatics and scientific service provision. We 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 EMBL's International Centre for Advanced Training (EICAT).

Making training FAIR

Bioinformatics and big data analysis are becoming essential skills for life science researchers. To support scientists at all levels and build capacity, the EMBL-EBI Training team brings together a wealth of content and expertise, and makes bioinformatics training more Findable, Accessible, Interoperable and Reproducible (FAIR).

One of the biggest challenges is access to training. During the COVID-19 pandemic, the team redoubled its efforts to provide easily-accessible online training and webinars. In 2021, the uptake of webinars exceeded all expectations, showing the critical global need for increased bioinformatics capacity.

A second challenge is understanding what competencies are required for certain career paths. To guide users in their career and training choices, our Training team is developing a Competency Hub, which lists in a clear and accessible way the competencies individuals might need in their career.

Finally, keeping track of training and being able to refer back to it is essential for learners. The new EMBL-EBI training platform empowers researchers to design their own training journey. It brings together curated collections from hundreds of online tutorials, webinars and courses, while also proposing new formats for handbooks, making it easier to keep track of notes once courses finish.

In close consultation with the global user community, EMBL-EBI continues to grow its training offering, and develop its competency hub with the aim of opening up bioinformatics to a wider, more diverse group of people.

"I really enjoyed the course and I will benefit a lot in my future work! By far the best course I've ever attended."
Trainee, Germany

"The expertise of the trainers on their topics was one of the best things, compared to other courses I've taken. The state-of-the-art information provided will be most useful in the future."
Trainee, Spain

Building capacity in Latin America

The CABANA project aims to strengthen capacity in data-driven biology in Latin America. Led by EMBL-EBI and nine research organisations from Latin America, CABANA helped researchers in the region participate in large global consortia equitably, and contribute to solving global challenges, specifically biodiversity, food security and communicable diseases.

The project, which ran between October 2017 and May 2022, was funded through the UK Research and Innovation's Global Challenges Research Fund. CABANA enabled the delivery of bioinformatics workshops in Latin America, the creation of bespoke e-learning courses and train-the-trainer activities to increase capacity. At the heart of the project were secondments that enabled Latin American scientists to visit other research institutes and embed themselves in another lab. The project also supported seven collaborative research projects in the region.

Despite the challenges that the pandemic raised for this project, which was based on moving people from one continent to another, all the objectives were achieved and surpassed, resulting in a strong network that future collaborations can build on.

CABANA ACHIEVEMENTS

39
secondments

28
workshops

9
e-learning courses

11
train-the trainer courses (120 people trained)

800
scientists trained in bioinformatics

9
research innovation awards*

*of which the biggest one sequenced SARS-CoV-2 genomes and identified strains in Latin America

"Projects like CABANA allow people in Latin America to build further bonds between bioinformatics groups, to be a part of this community and carry out research using bioinformatics."
Guillermo Rangel-Pineros, Secondee from University of Los Andes, Columbia

"For teaching, you need different views and different expertise and we get all this from CABANA."
Alfredo Herrera-Estrella, CABANA Co-investigator, National Laboratory of Genomics for Biodiversity, Mexico

"There's no doubt now that the CABANA network will continue and is committed to delivering training and pursuing its challenge-led research goals. We've helped create a community of data-driven biologists working across an entire continent, and that's a really powerful thing."
Cath Brooksbank, EMBL-EBI Head of Training

SARAH MORGAN

Not even a pandemic can stop training

"Bioinformatics is essential for anyone doing a PhD in the life sciences – it unlocks access to so much data," explained Sarah Morgan, Training Programme Manager at EMBL-EBI.

Born in Wales, Sarah studied biomedical science at Cranfield University, where she also worked as a lecturer before joining EMBL-EBI. Sarah manages the day-to-day running of the EMBL-EBI training programme, and has overseen its development into one of the most extensive bioinformatics training offerings in the world.

"The pandemic was a big incentive to create a framework for running our courses virtually. The team was fantastic in coming up with ideas and making things happen. Interest in our courses spiked during the pandemic, and we developed a range of tools that we'll continue to use in the future. We also launched a course on single-cell RNA sequencing, which has proven very popular."

EMBL-EBI TRAINING IN 2021

17 live training activities reaching **486** delegates

Eight new online tutorials, bringing the total to **83**

110,000 people watched our webinars

497,000 unique IP addresses accessed EMBL-EBI training website

Strategic partnerships

As biology becomes more data-driven, EMBL-EBI is developing strategic partnerships with private and public sector organisations to help them maximise the potential of bioinformatics.

Building strong ties with industry

EMBL-EBI's Industry Programme is a subscription-based initiative, aimed at global companies that make significant use of EMBL-EBI data resources as part of their research and development activities.

"The EMBL-EBI Industry Programme has been by far the strongest public-private enterprise to shape the relationships between academia and pharma."
Bertram Weiss, Head of Research Product Platform, Bayer AG

In 2021, EMBL-EBI also consolidated its ties with small and medium companies (SMEs) through the inaugural edition of the Bioinformatics for BioBusiness event, organised in collaboration with Medicines Discovery Catapult. This was the perfect opportunity for SMEs to get up to speed with the latest developments in bioinformatics and get bespoke specialist advice from our experts in genome annotation, single-cell sequencing, metagenomics and drug discovery.

"We use ChEMBL all the time to access public domain chemical structures. Recreating that would not be possible for a single company."
Impact survey respondent, UK

Accessing biomedical data for data analysis

SME forum

Wednesday 6 July 2022
University of Manchester
Barnes Wallis Building, M1 7AJ
View on a map

Registration opening
12 May 2022

Jointly hosted by
EMBL's European Bioinformatics Institute (EMBL-EBI) and Medicines Discovery Catapult (MDC)

EMBL-EBI INDUSTRY PROGRAMME

Founded in **1996**

26 member companies from pharma, agri-food and consumer goods

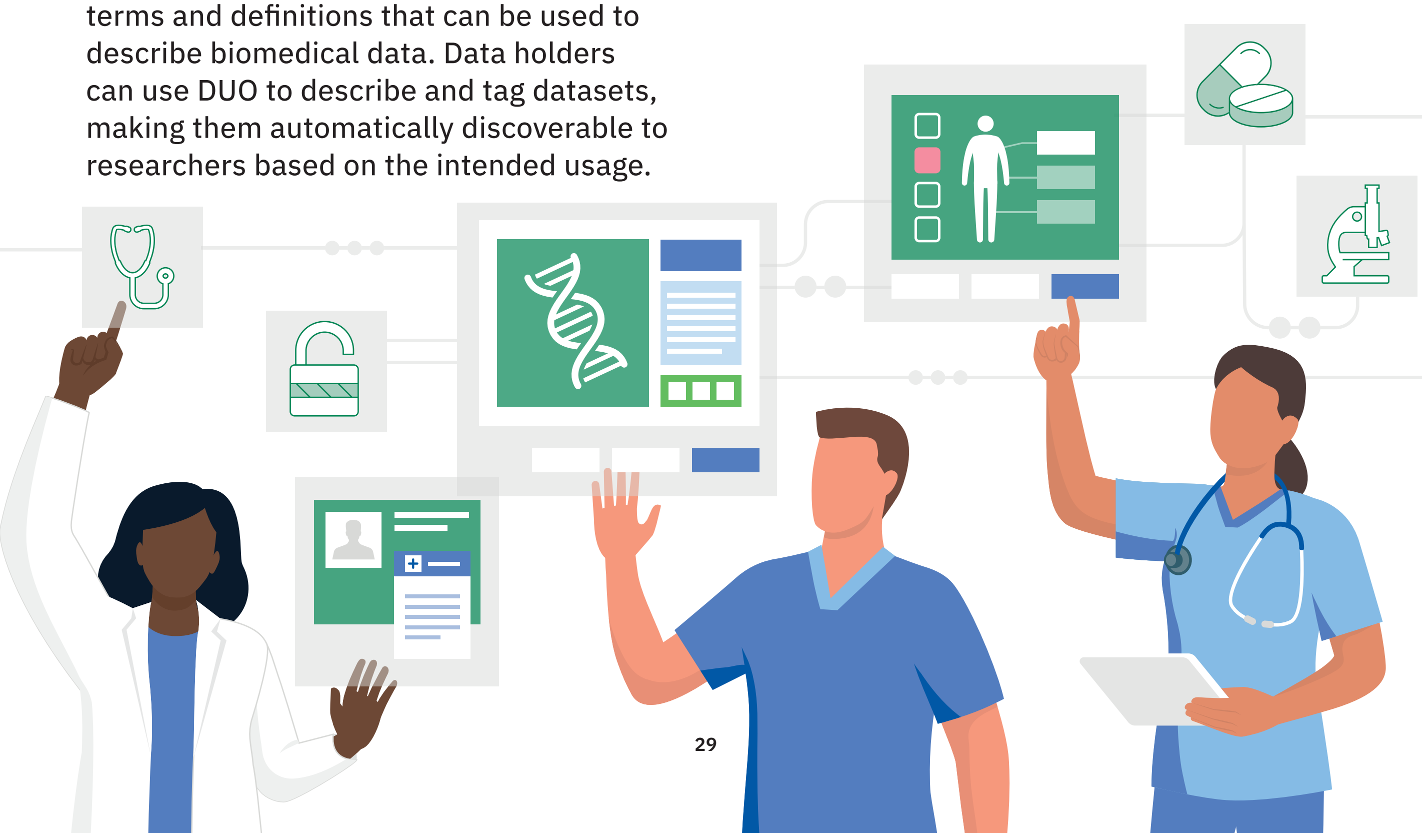
Six workshops with **900** participants in 2021

Streamlining access to biomedical data

As part of the Global Alliance for Genomics and Health (GA4GH), EMBL-EBI researchers develop genomic data standards and tools to enable responsible genomic data sharing within a human rights framework.

In 2021, with colleagues from the Broad Institute, we developed the Data Use Ontology (DUO), which provides standard terms and definitions that can be used to describe biomedical data. Data holders can use DUO to describe and tag datasets, making them automatically discoverable to researchers based on the intended usage.

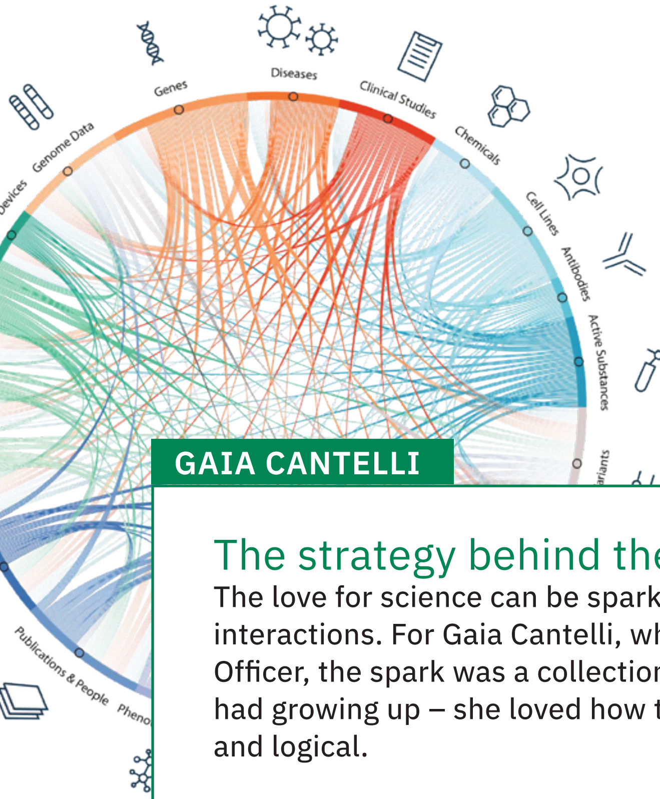
DUO is set to streamline researcher access to biomedical datasets, making it easier to navigate the complex landscape of sensitive biomedical data. Better and faster access to biomedical data within a safe and secure system will speed up genomic research, and the use of genomics in a clinical setting.



Coordinating bioinformatics in Europe

ELIXIR is an intergovernmental organisation that coordinates, integrates and sustains bioinformatics resources from different European providers – including EMBL-EBI – and enables users in academia and industry to access services that are vital for their research.

As an ELIXIR Node, EMBL-EBI supports the coordination of biological data provision throughout Europe. Many EMBL-EBI data resources are included in ELIXIR's Core Data Resources list – a repository of databases deemed to be fundamental to the life sciences. EMBL-EBI provides computing power, training, and standardisation activities to the ELIXIR community. We are also collaborating with ELIXIR on the BY-COVID project, which provides comprehensive open data on SARS-CoV-2 and other infectious diseases across scientific, medical, public health and policy domains.



GAIA CANTELLI

The strategy behind the science

The love for science can be sparked by the smallest interactions. For Gaia Cantelli, who is an EMBL Strategy Officer, the spark was a collection of illustrated books she had growing up – she loved how they were both beautiful and logical.

Gaia was born in Bologna, Italy, and studied cell and molecular biology in Cambridge and London before doing a postdoc and teaching scientific writing at Duke University in the USA.

"My current role involves strategic planning, implementation and reporting," explained Gaia. "My team works with all the parts of EMBL to align everyone to a common direction of travel. We're currently focusing on the implementation of EMBL's new and ambitious programme – Molecules to Ecosystems – which aims to study life in context. In 2022 we're going to really see the programme come to life, as work on new initiatives ramp up."

Culture and infrastructure

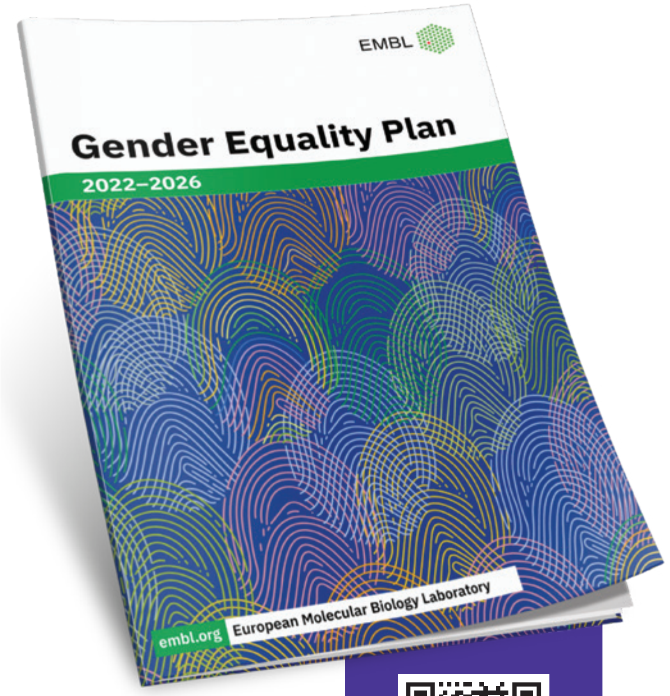
The success of EMBL-EBI relies on its collaborative culture, diverse workforce, and robust technical infrastructure. In 2021, some of the key focus areas were equality, diversity and inclusion, environmental sustainability, and consolidating existing technical infrastructure.

Equality, diversity and inclusion

Equality, diversity and inclusion (EDI) became embedded across all EMBL sites, including EMBL-EBI, with the establishment of the EMBL EDI Office, a cross-EMBL volunteer forum and the EDI governing body.

The results of an all-staff EMBL survey, conducted in March 2021, about perceptions of EDI across the organisation informed EMBL's first Equality, diversity and inclusion strategy, now in its early implementation phase.

The four axes of the strategy focus on attracting and promoting underrepresented groups at EMBL, ensuring that EDI principles are built into EMBL's systems and processes, developing a model of inclusive leadership in science, and engaging on EDI with EMBL's many collaborators.

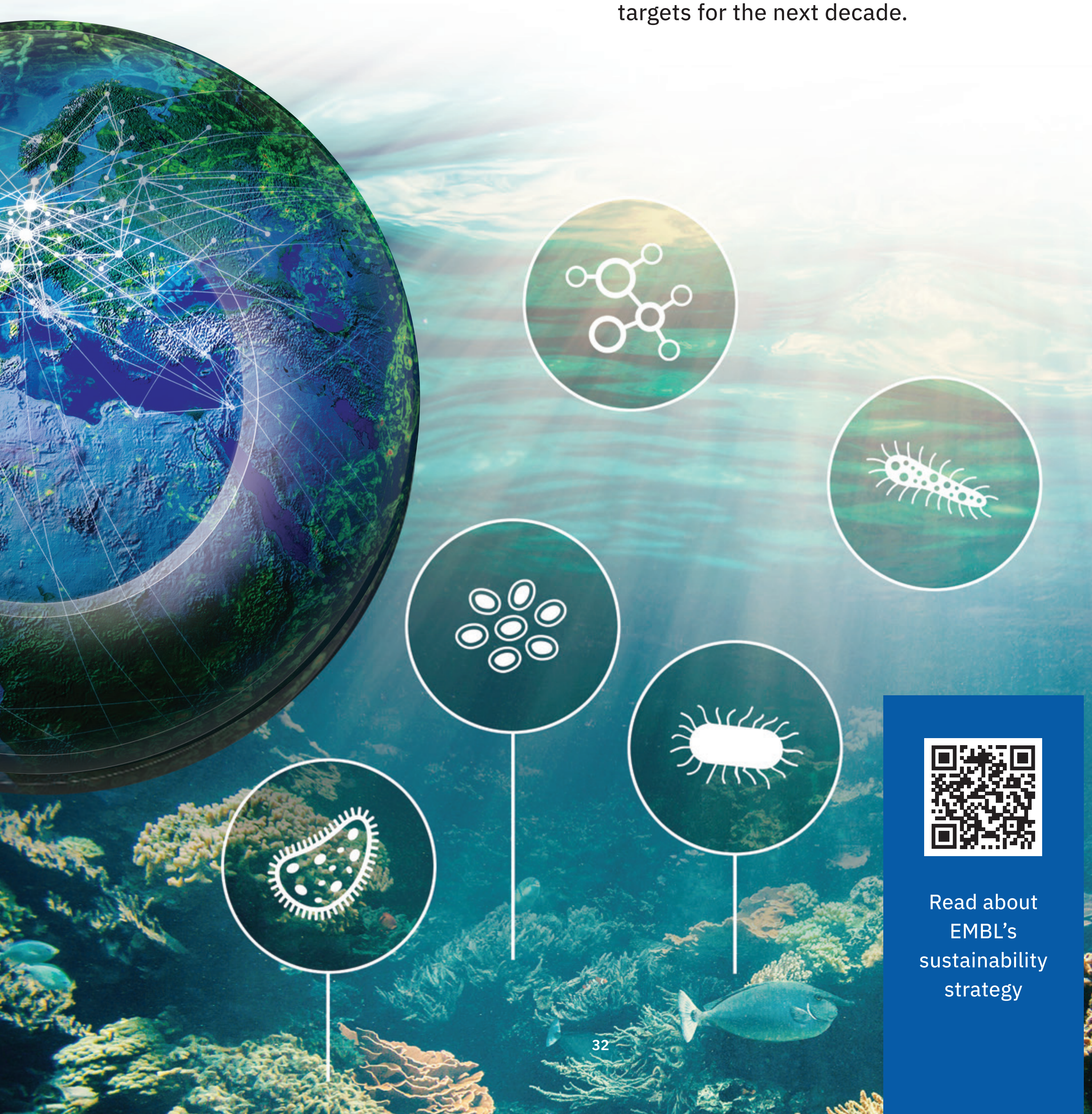


Read more about EMBL's Equality, Diversity, and Inclusion strategy

Environmental sustainability

To show EMBL's commitment to sustainable research, and help reduce its impact on the environment, the organisation commissioned a materiality assessment, which invited input from staff, leadership and key external stakeholders.

The assessment identified three pillars of sustainability at EMBL: environmentally-responsible research, environmentally-relevant research and promoting sustainable science. The organisation has now also set ambitious environmental targets for the next decade.



Read about EMBL's sustainability strategy

Public engagement

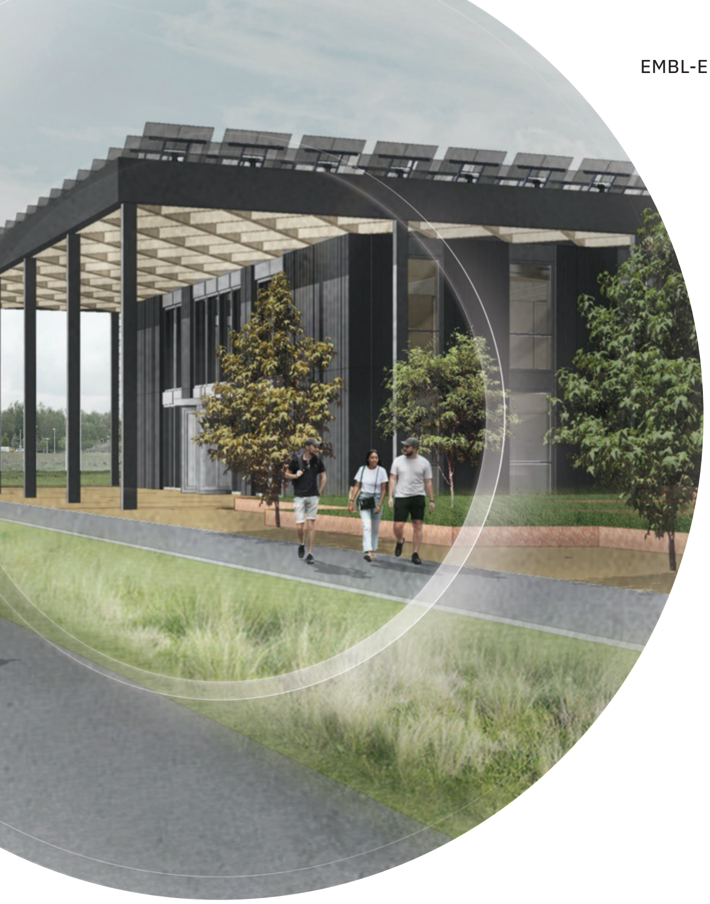
A key part of EMBL-EBI's public engagement strategy is to connect with audiences to understand their needs and to develop mutually-beneficial relationships. The audience scoping project, in collaboration with Graphic Science, was the first step on this journey. We connected with five community intermediaries and conducted a teacher focus group. One project has already been delivered, and several new activities are in development.

Find out more about our public engagement activities

Supporting staff during the pandemic

As the pandemic continued into the second year, EMBL-EBI's Business Continuity Planning team supported the institute to navigate the uncertainties of returning to the office when possible. EMBL laid the groundwork for launching a hybrid working pilot, which will enable staff to adopt more flexible working patterns, while continuing to tap into the creative and collaborative culture of the campus.

EMBL-EBI recruited and onboarded over 200 new members of staff in 2021, and continued to support staff with the uncertainties and challenges of working during the COVID-19 pandemic.



A new building

To increase office space, we built the Modular Accommodation, which houses 100 staff. As part of the Wellcome Genome Campus expansion, EMBL-EBI is set to construct a third permanent building, which will accommodate 250 members of staff. The new building is funded by an award from UK Research and Innovation and Wellcome. The planning application for the building has been accepted, and significant progress has been made through staff consultation and outlining the design concept.

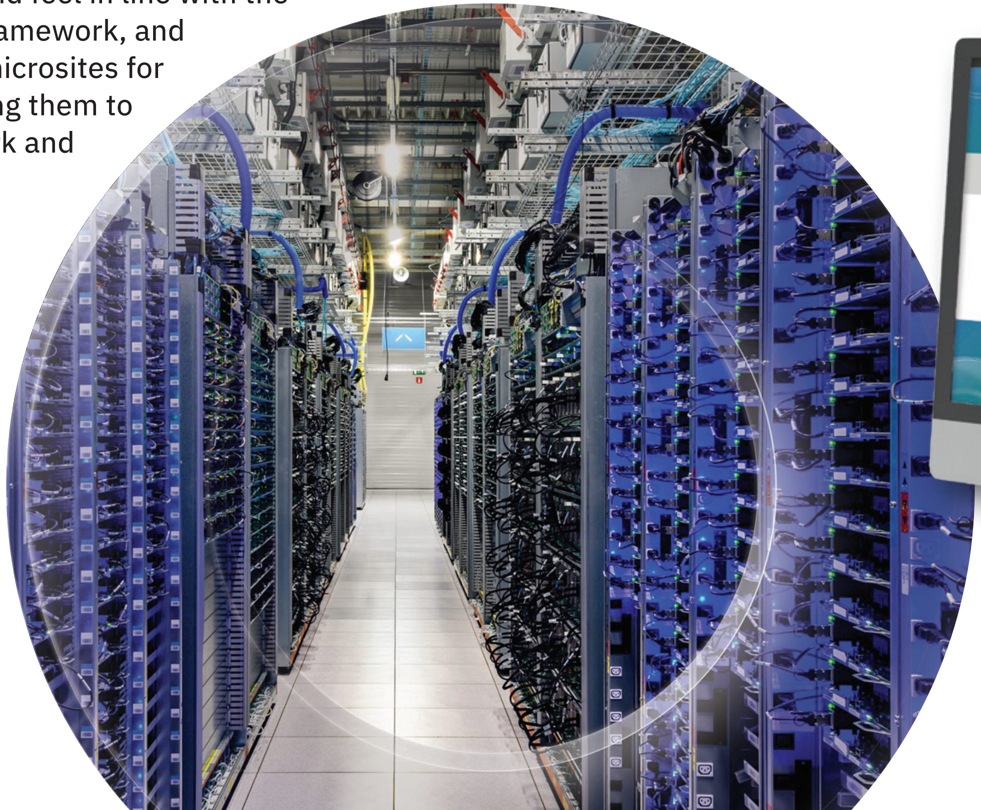
Construction is planned to begin in late 2022, with the building set to open in 2024.

Improved technical infrastructure

EMBL-EBI's total data footprint has been growing year on year, and is now approaching 400 petabytes of raw storage. We're currently focusing our efforts on consolidating the storage infrastructure by removing older temporary data that no longer has value, or that is duplicated elsewhere, in preparation for new data types and future growth.

Alongside this, our teams have also been tapping into the potential of cloud storage and computing, with two new strategic partnerships with Google Cloud and Amazon Web Services announced in 2021. This is the next step in EMBL's hybrid multi-cloud strategy, and was made possible through funding from UK Research and Innovation.

The institute has also updated its website, bringing the look and feel in line with the new EMBL visual framework, and creating bespoke microsites for every team, enabling them to showcase their work and collaborations.



ANIA NIEWIELSKA

Developing creative tech solutions

After studying computer science in the city of Gliwice, Poland, Ania Niewielska spent over ten years developing software solutions for the banking sector.

Following a life-long interest in the life sciences, she joined EMBL-EBI and currently works as a Lead Software Engineer. After working on several EMBL-EBI projects, Ania has recently taken over management of the institute's Resource Usage Portal, which helps internal teams get the technical resources they need for their projects. This covers virtual machines, databases and the cloud.

"I like working on real use cases, interacting with users and stakeholders, asking them questions and drawing out what problems they're really trying to solve, then finding the best tech solutions for them," said Ania.

"The thing I enjoy most about EMBL-EBI is the campus – it's a safe space and a great environment to exchange ideas."

EMBL-EBI

Unleashing the potential of big data in biology

Deep learning models help predict protein function

Genomic data for biodiversity – a global challenge

The Earth Biogenome Project: on track and ready to ramp up

Improving metadata for genomic sequences

The value of open data

Independent economic report finds EMBL-EBI data resources are critical for global research and development.

New leadership

Meet the group and team leaders who joined EMBL-EBI in 2021

Mary Barlow
Mary was promoted to Head of Campus Operations and Capital Projects. Her team manages a range of crucial functions, including facilities, new building development, process improvements, and oversight of operational and capital projects.

Sihem Bennour
Sihem is EMBL-EBI's new Head of Human Resources. Her team supports EMBL-EBI staff during every stage of employment, and promotes an enabling, inclusive, international working environment for all.

Sarah Dyer
Sarah is the new Non-vertebrate Genomics Team Leader. She brings a wealth of experience working with plant and crop genomes, and leads three Ensembl teams – Plants, Metazoa and Outreach.

Tudor Groza
Tudor joined EMBL-EBI to lead the newly-formed Phenomics team, which focuses on developing data acquisition and integration services to support advances in computational phenotyping for rare and complex disorders.

Melissa Harrison
Melissa joined EMBL-EBI as a Team Leader for Literature Services. Her team runs Europe PMC, an open science platform that enables access to life science publications and preprints.

Peter Harrison
Peter is the new Team Leader for Genome Analysis. His team develops the regulatory components, core infrastructure and web applications for the Ensembl project, a world-leading browser for vertebrate genomes.

Matthew Hartley
Matthew leads the BioImage Archive team, which develops and operates EMBL-EBI's broad scope bioimage data resource. His team's experience crosses biological imaging, data management and software development.

Facts and figures

Staff numbers

Personnel categories in 2021

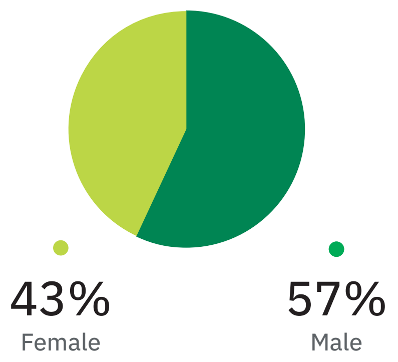
708

621
Staff members

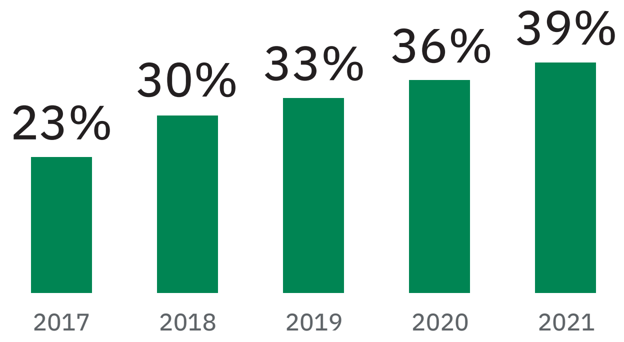
47
Postdoctoral fellows

40
Predoctoral fellows

Gender distribution of staff in 2021



Senior roles held by women

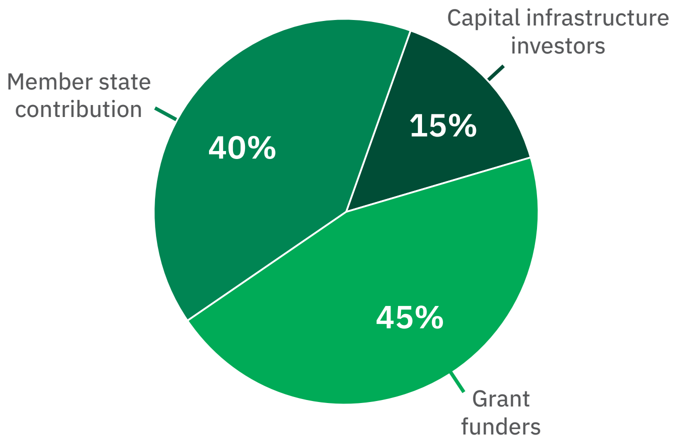


Financial figures

We are grateful for the continued support of our member states and other funding bodies. In 2021 they helped us maintain our data resources, conduct vital research and training, as well as respond to the requirements of the international scientific community.

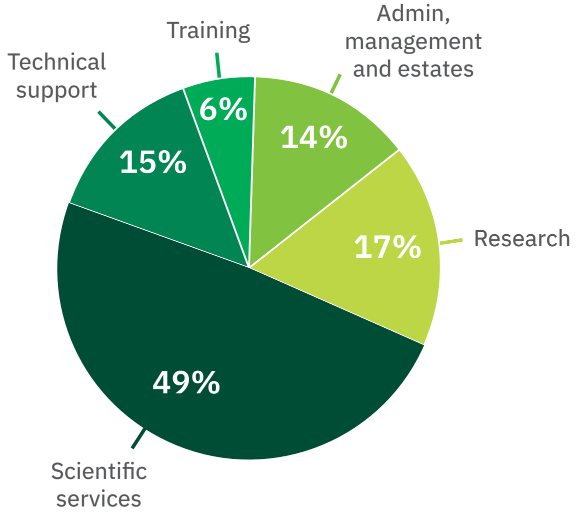
Funding sources

EMBL-EBI receives its funding through three main funding channels. The operating expenditure is funded by either EMBL member state contributions (40%) or from external funding bodies via grant awards (45%). 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 2021 was €92.6m from all funding sources. Scientific services accounted for 49% of the expenditure in support of EMBL-EBI's data resources, with a further 17% allocated to research and 6% to external training. Approximately 14% of costs were on technical support which maintains and develops the technical/IT infrastructure, and 14% on management, administration, and estate costs.



Capital investment

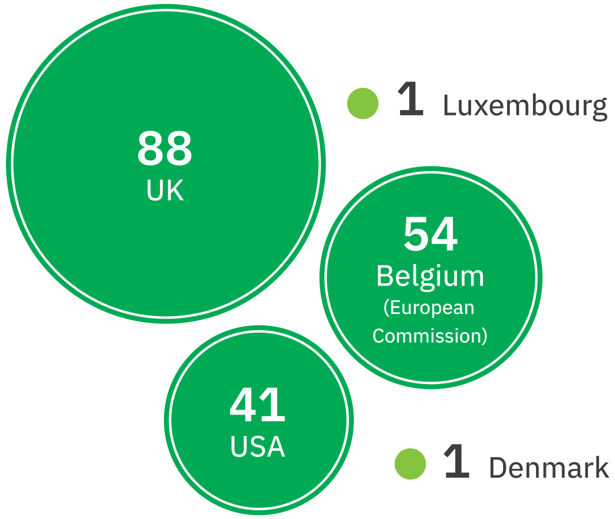
EMBL-EBI's Data Infrastructure for Life Sciences programme began in 2019 and runs to 2024. The multi-year programme oversees £70 million of capital investment, of which £45 million comes from the UK government's Strategic Priorities Fund via UK Research and Innovation (UKRI). This enables the transformation of storage, compute and networking facilities as the demand for our data resources continues to grow, and helps to establish the BioImage Archive, a central, open data resource for biological images (page 16).

The second part of the programme is to increase the office space available to EMBL-EBI. The Modular Accommodation was completed in 2021 providing accommodation for 100 EMBL-EBI staff. This was funded with contributions from the UKRI and EMBL. The UKRI and Wellcome Trust have funded a longer-term office space solution – a third permanent building called the Thornton Building (page 34).

Expenditure relating to these capital investments amounted to €14.5m during the 2021 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 2021 we received €39.5m in research and service grant funding from 23 funders (page 41) providing support for 186 projects.



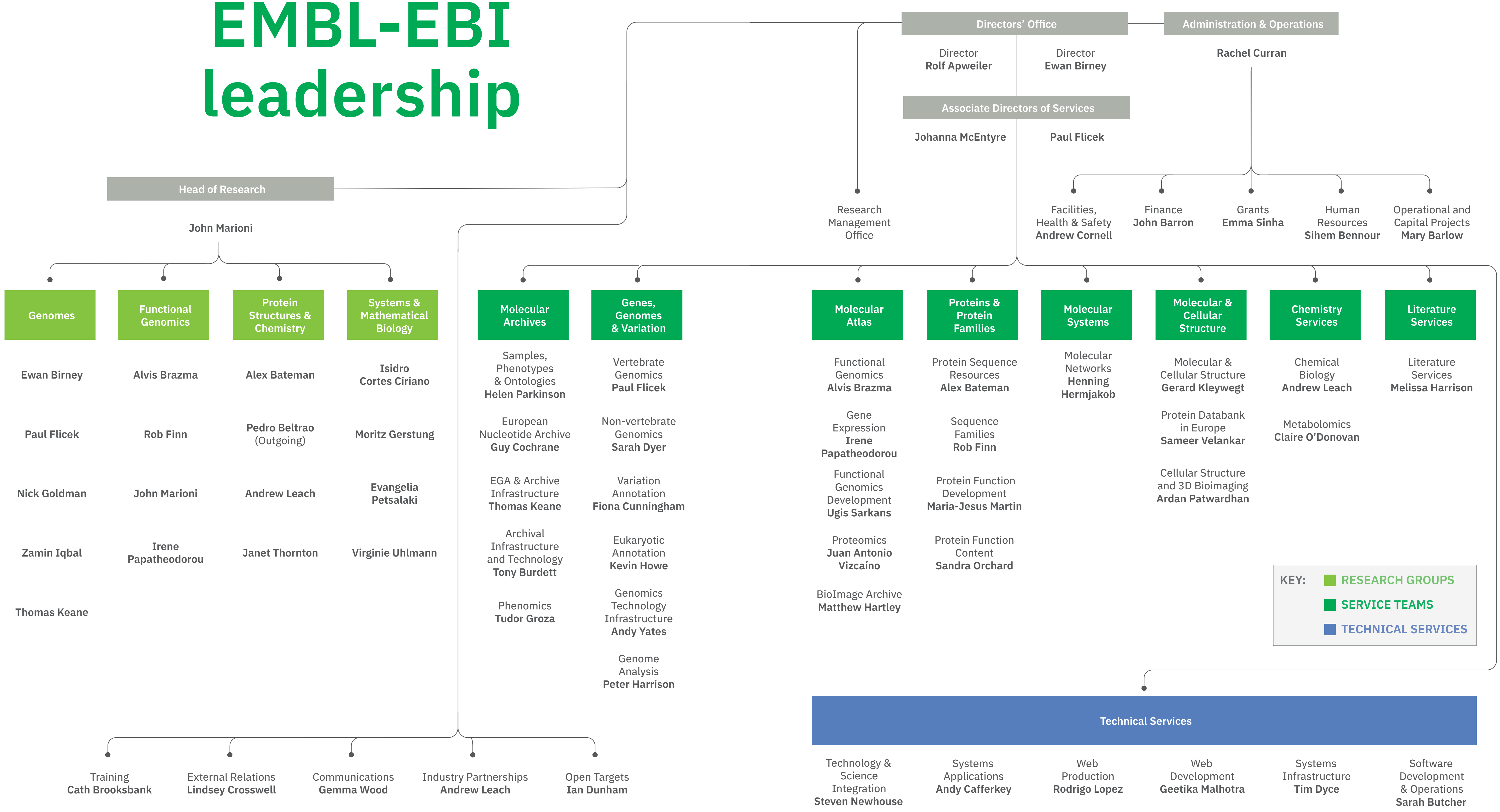
Number of grants received by EMBL-EBI in 2021 by funding body location.

Our funders

We would like to thank our funders for their continued support. We thank the EMBL member states as well as our other funders listed below.

- Alzheimer's Research UK
- Biotechnology and Biological Sciences Research Council
- Bill & Melinda Gates Foundation
- British Council
- Broad Institute of MIT and Harvard
- Cancer Research UK
- Connective Tissue Oncology Society
- Chan Zuckerberg Initiative
- European Commission
- Engineering and Physical Sciences Research Council
- Economic and Social Research Council
- Fonds National de la recherche Luxembourg
- Global Challenges Research Fund
- Gordon and Betty Moore Foundation
- Medical Research Council
- NF Research Initiative
- National Cancer Institute
- National Institutes of Health
- National Institute for Health Research
- Novo Nordisk
- National Science Foundation
- Sarcoma Foundation of America
- Wellcome

EMBL-EBI leadership



Directors' Office

Director
Rolf Apweiler

Director
Ewan Birney

Associate Directors of Services

Johanna McEntyre

Paul Flicek

Administration & Operations

Rachel Curran

Facilities, Health & Safety
Andrew Cornell

Finance
John Barron

Grants
Emma Sinha

Human Resources
Sihem Bennour

Operational and Capital Projects
Mary Barlow

Research Management Office

Head of Research

John Marioni

Genomes

Ewan Birney

Paul Flicek

Nick Goldman

Zamin Iqbal

Thomas Keane

Functional Genomics

Alvis Brazma

Rob Finn

John Marioni

Irene Papatheodorou

Protein Structures & Chemistry

Alex Bateman

Pedro Beltrao (Outgoing)

Andrew Leach

Janet Thornton

Systems & Mathematical Biology

Isidro Cortes Ciriano

Moritz Gerstung

Evangelia Petsalaki

Virginie Uhlmann

Molecular Archives

Samples, Phenotypes & Ontologies
Helen Parkinson

European Nucleotide Archive
Guy Cochrane

EGA & Archive Infrastructure
Thomas Keane

Archival Infrastructure and Technology
Tony Burdett

Phenomics
Tudor Groza

Genes, Genomes & Variation

Vertebrate Genomics
Paul Flicek

Non-vertebrate Genomics
Sarah Dyer

Variation Annotation
Fiona Cunningham

Eukaryotic Annotation
Kevin Howe

Genomics Technology Infrastructure
Andy Yates

Genome Analysis
Peter Harrison

Molecular Atlas

Functional Genomics
Alvis Brazma

Gene Expression
Irene Papatheodorou

Functional Genomics Development
Ugis Sarkans

Proteomics
Juan Antonio Vizcaíno

BioImage Archive
Matthew Hartley

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
Melissa Harrison

Technical Services

Technology & Science Integration
Steven Newhouse

Systems Applications
Andy Cafferkey

Web Production
Rodrigo Lopez

Web Development
Geetika Malhotra

Systems Infrastructure
Tim Dyce

Software Development & Operations
Sarah Butcher

Training
Cath Brooksbank

External Relations
Lindsey Crosswell

Communications
Gemma Wood

Industry Partnerships
Andrew Leach

Open Targets
Ian Dunham

KEY:
🟩 RESEARCH GROUPS
🟩 SERVICE TEAMS
🟦 TECHNICAL SERVICES

Our governance

EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an intergovernmental organisation with 27 member states, one associate member state 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 2021, 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.

Imprint

Publisher
EMBL-EBI

Portrait image credits
p. 35 (Ania Niewielska portrait): Maciej Grzywocz

All portraits not listed here are from personal collections or taken on behalf of EMBL by Jeff Dowling and Mary Todd Bergman.

Other image credits
p. 8 (AlphaFold protein) and p. 12: DeepMind
p. 8 (virus and PGS Catalog): Spencer Philips/EMBL
p. 8 (CRyPTIC project): Karen Arnott/EMBL
p. 9 (bacteria): Spencer Philips/EMBL
p. 9 and p. 16 (REMBI): BioImage Archive
p. 9 and p. 25 (CABANA project): Jeff Dowling/EMBL
p. 9 (training participants): Phil Mynott
p. 9 and 34 (servers): Google
p. 32 (composition for COP26): Kristina Arató/EMBL

Printer
Kingfisher Press. www.kingfisher-press.com

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, Australia

Prospect member states: Estonia, Latvia