EMBL-EBI

Highlights 2022

ebi.ac.uk

European Bioinformatics Institute (EMBL-EBI)

Contents

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

Cover illustration: Karen Arnott/EMBL-EBI

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

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 research, provides training and state-of-the-art research infrastructures. EMBL is an intergovernmental organisation with 28 member states. EMBL's latest scientific programme – Molecules to Ecosystems – seeks to better understand life in context.

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 life sciences are undergoing a fundamental transformation driven by the adoption of artificial intelligence and machine learning. But leveraging these technologies is dependent on the existence of vast, good-quality training data sets, such as those managed by EMBL-EBI.

As the custodians of the world's biological data, we are contributing to the AI revolution. We provide high-quality data that can be used to train AI systems, and we make AI-generated predictions available to the world, carefully integrating them into the bioinformatics ecosystem when appropriate. We also use AI to improve our data resources and explore new research avenues. Finally, we have a responsibility to train researchers on using AI-generated data, and to participate in the public dialogue regarding the ethics of AI.

As more researchers venture into this space, a society-wide discussion about the role, limitations, regulation and ethics of AI becomes crucial. Science and society must define how these innovative technologies will be used in equitable and fair ways, in the life sciences and beyond.

This report outlines some of EMBL-EBI's recent AI developments, alongside other major achievements from 2022, including advances in bioimaging, genomic medicine and biodiversity.

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. Our work would not be possible without them.

Ewan Birney and Rolf Apweiler
EMBL-EBI Directors

2022 in numbers

794 members of staff*

from **71** countries

**includes fellows, supernumeraries, trainees and visitors*

3,700 people participated in our 27 public engagement activities

107 million requests to our data resource websites on an average day

from **41** million unique IP addresses

471,000 unique IP addresses accessed our online training

22 live courses reaching 592 delegates

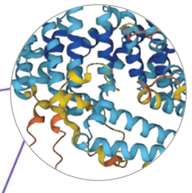
129,389 webinar views

197 active grants

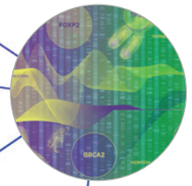
158 collaborative grants with researchers in 53 countries from 753 institutes

193 journal papers published

Highlights of the year



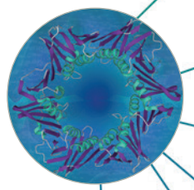
AlphaFold Database updated to over 200 million protein structure predictions (page 10)



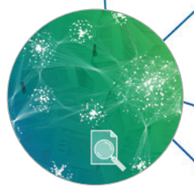
Ensembl employs deep learning to allocate gene function (page 12)

European institutes commit to data access across borders through the Federated European Genome-phenome Archive (page 15)

Supporting the African BioGenome project (page 17)

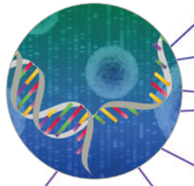


UniProt uses natural language processing to speed up protein annotation (page 11)

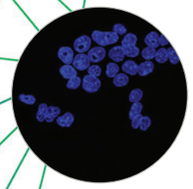


Europe PMC harnesses text mining to make knowledge more findable (page 12)

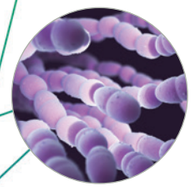
European Genomic Infrastructure project begins (page 15)



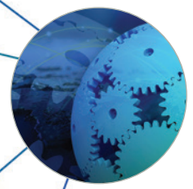
Cortes-Ciriano group uses long-read sequencing to understand cancer (page 18)



Uhlmann group exploits deep learning for bioimage analysis (page 20)



Lees group explores pathogen genome variation to inform vaccination (page 21)

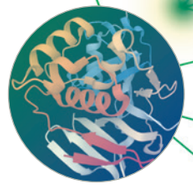


Industry programme welcomes four new companies (page 25)

Thornton Building construction starts (page 26)

EMBL-EBI website sees major update (page 28)

Petsalaki group improves the proteome-wide discovery of human linear motifs (page 21)



In-person training resumes alongside an extensive virtual programme (page 22)

Workshop explores AI opportunities with industry partners (page 25)

Hybrid working pilot (page 26)

Increased energy savings from building and technical infrastructure (page 28)

Data resources

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

Critical to life science research

The Global Biodata Coalition (GBC) recognised 16 data resources managed by EMBL-EBI and partners as being critical to life science and biomedical research worldwide. The GBC is a forum for research funders, which aims to coordinate the efficient management and growth of biodata resources, and to ensure sustainable financial support for global biodata infrastructure.

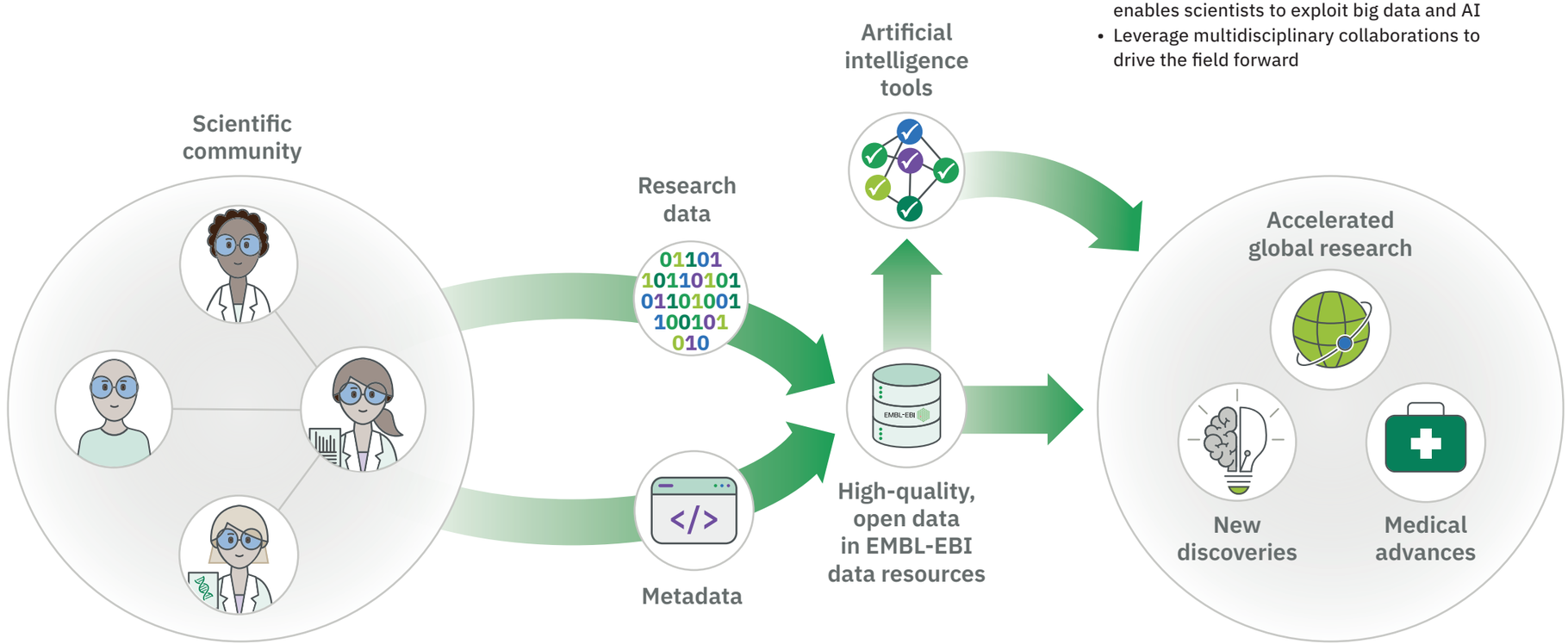
"This is a major step in defining the global life sciences data infrastructure on which biomedical research depends, and highlights the need for sustainable funding for these data resources."
Jo McEntyre, Associate Director of EMBL-EBI Services

Powering the AI revolution

Artificial intelligence (AI) is increasingly used to accelerate life science research and develop solutions for global challenges. But the quality of AI systems is only as good as the data used to train them. As a world-leading provider of high-quality biological data, EMBL-EBI is a key player in this growing field, and in 2022 made several leaps.

We use and support the development of AI algorithms in a number of ways:

- Annotate, classify and enrich data to create the high-quality datasets essential to train AI
- Use AI to streamline and improve our data resources
- Develop new, open source AI methods to conduct novel research and data analysis
- Work with AI experts to develop training that enables scientists to exploit big data and AI
- Leverage multidisciplinary collaborations to drive the field forward



Explore AI at EMBL-EBI

Case study: AlphaFold AI uses open data to uncover the 3D protein universe

Proteins make up all living things, but how they fold to create unique shapes is one of biology's biggest mysteries. In 2020 for the first time ever, an AI system predicted protein structures with a level of accuracy comparable to experimental methods.

DeepMind's revolutionary AI, AlphaFold, was trained using publicly-available data, such as those managed by EMBL-EBI. These included experimentally-determined protein structures from the Protein Data Bank, protein sequences and annotations from UniProt, and metagenomics data from MGnify.

DeepMind approached EMBL-EBI to co-develop a data resource that could host AlphaFold protein structure predictions. The AlphaFold Database, developed in record time, was launched in 2021 with just 20,000 predictions covering all human proteins. Within a year, in a gargantuan effort, EMBL-EBI and DeepMind released a major update, covering over 200 million protein structure predictions – virtually all proteins known to science.

This included organisms on the WHO's neglected tropical diseases and antimicrobial resistance list, which hold immense research potential.

As the data management expert, EMBL-EBI contributed the creation of data standards to ensure quality, data curation to make predictions easy to find, and importantly, data integration – cross linking AlphaFold predictions with other biomolecular data resources. This integrated the AlphaFold predictions into the global bioinformatics ecosystem, adding value for the research community.

The AlphaFold system and database have disrupted the way biology is studied, and have stimulated new research directions in the field of structural biology, such as tackling antibiotic resistance, researching disease cures and vaccines – from leishmaniasis, to malaria – improving our understanding of honey bees, and developing new tools for exploring the protein universe.

"EMBL's support in developing the AlphaFold Database was crucial; it significantly amplified the impact and reach of AlphaFold predictions across the global scientific community."
John Jumper, Senior Staff Research Scientist at DeepMind & AlphaFold Team Lead

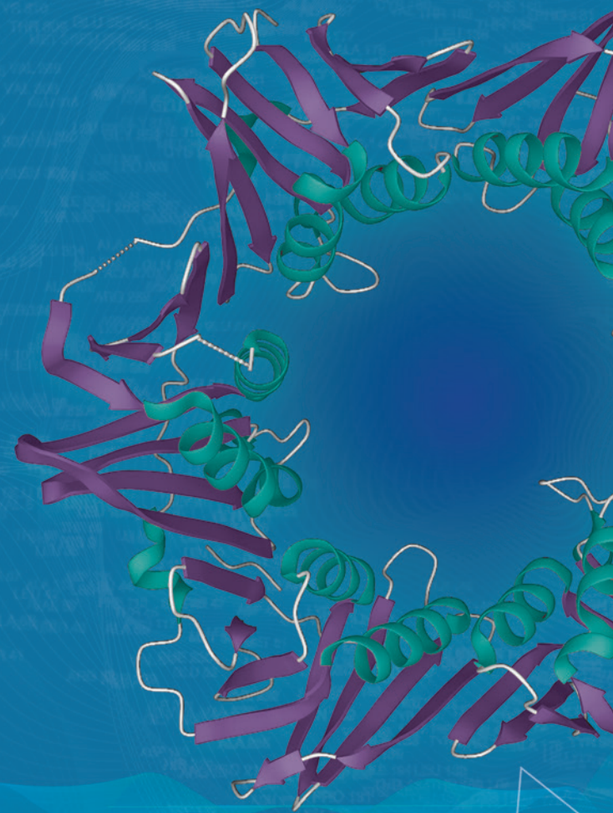
"Scientists build on the shoulders of giants. Most often, those shoulders are data. Having these millions of structure predictions will change the face of biology. This is useful for medicine, agriculture, biotech, everything – it's just fantastic."
Janet Thornton, Director Emeritus, EMBL-EBI

AI speeds up protein annotation

Over 40 million previously-uncharacterised proteins in the UniProt database have now been annotated – given a functionally-relevant name – using a language processing model developed by Google Research and evaluated with help from EMBL-EBI's data curation experts.

The model, called ProtNLM, accurately predicts descriptions of protein function directly from a protein's amino acid sequence. In the future, the team aims to use the vast gene ontology data at EMBL-EBI to train machine learning models to provide additional protein function information, alongside protein names. The aim is to enable more researchers to explore the functional significance of these proteins for the first time.

"The team at UniProt is absolutely top-notch, and we couldn't be more pleased to contribute back to these amazing resources. We're helping millions of people do their research."
Max Bileschi, Staff Research Software Engineer and Manager, Google Brain



Allocating gene function with machine learning

In a similar project to ProtNLM, the Ensembl team started using machine learning to allocate functions to genes in newly-annotated vertebrate genomes. Annotating a genome means identifying the locations and structures of genes and other genomic features. Genome annotation is incredibly useful to researchers, but assigning gene function usually requires additional work and experimental evidence.

By using machine learning to allocate gene function, the team is saving significant computing time and simplifying the entire process. This is particularly valuable in the context of large biodiversity projects, such as the Darwin Tree of Life, which aims to sequence all the eukaryotic species in Britain and Ireland.

"If we are to annotate the genomes of all species on Earth, we need to think of where we can make computational savings. It's really exciting to explore how machine learning could improve the quality of our data and drastically reduce the associated computational cost and environmental footprint."
Fergal Martin, Eukaryotic Annotation Team Leader, EMBL-EBI

Harnessing text mining to accelerate life science research

Europe PMC – our open science platform for life science publications – is using text mining in a variety of ways. Text mining enables researchers to rapidly find new and hidden information in text-based sources by analysing vast amounts of material to extract meaningful concepts, relationships, and trends.

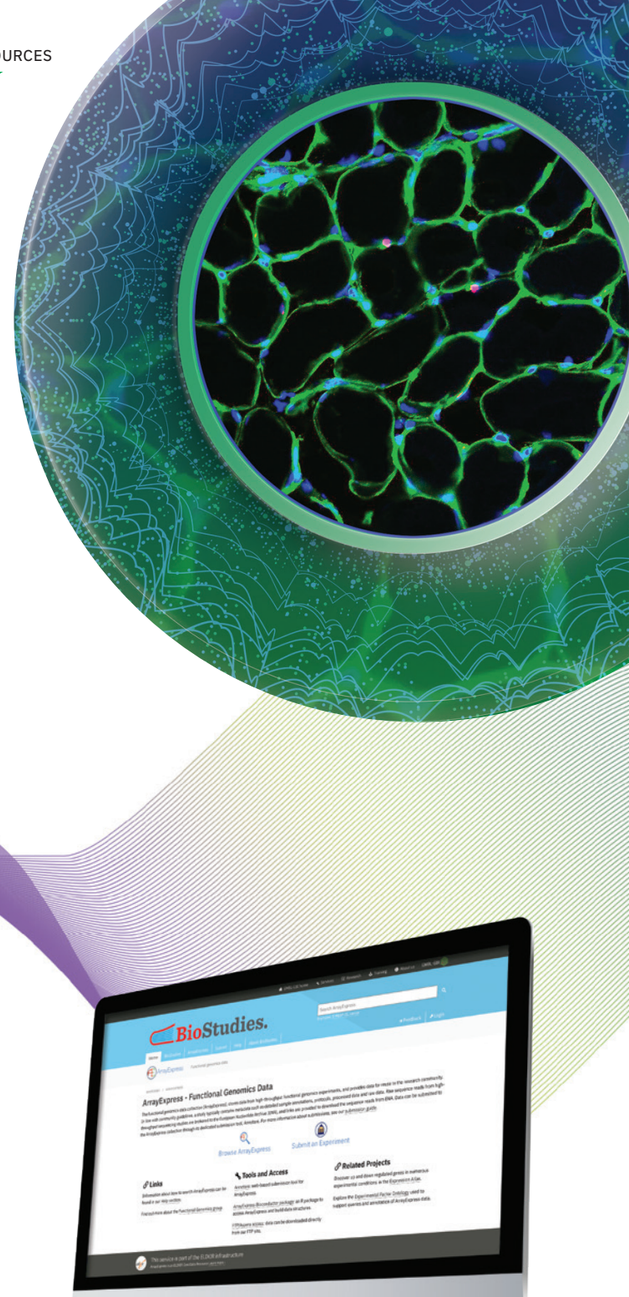
In collaboration with Open Targets, Europe PMC also uses text mining to identify genes, proteins and chemicals that could be useful in drug target identification and prioritisation. This approach helps to speed up this lengthy and costly process.

Data types to watch

Making bioimaging data FAIR

The BioImage Archive is EMBL-EBI's freely-available online resource for storing and accessing biological images. The archive continued to grow in 2022, releasing twice as many datasets as during the previous year. These cover a wide range of biological imaging techniques, organisms, model systems, experimental designs and research areas.

Following repeated testing and community consultations, the archive implemented the Recommended Metadata for Biological Images (REMBI) guidelines as its new metadata model. REMBI helps to make biological image data more Findable, Accessible, Interoperable and Reusable (FAIR), enabling users to understand how and why images were created and how best to work with them.



Retirement of ArrayExpress

When an EMBL-EBI data resource is no longer used to the same extent due to technology or science moving on, it is gradually retired, ideally in a way that still allows users to access the data if necessary.

After more than 20 years in action, the ArrayExpress interface was retired in late 2022. All datasets have now been migrated into the BioStudies ArrayExpress collection, and continue to be accessible freely and openly by researchers around the world.

Similarly, the Pfam website was retired, but the functionality and data are still available in EMBL-EBI's other protein families database, InterPro.

Case study: Metagenomics helps identify plastic degrading enzymes

EMBL-EBI's metagenomic database MGnify is a treasure trove for the discovery of novel enzymes with exciting properties, such as the ability to degrade plastic. But due to their large number, enzymes are often poorly characterised, which means identifying enzymes with specific properties, such as plastic degradation, can be very challenging.

A new collaboration between four investigators based at EMBL-EBI, University College London, and the University of Cambridge, and funded by UK Research and Innovation (UKRI), aims to address these challenges by employing a combination of computational and protein engineering approaches to discover new plastizymes and improve their catalytic ability.

"MGnify is probably the most comprehensive [metagenomics] database. It's easy to use and it keeps on growing."
Florian Hollfelder, University of Cambridge

VARSHA KALE
BIOINFORMATICIAN

Microbes are the new 'cool'

After working in the UK's National Health System, and at Public Health England's Microbiology surveillance laboratories, Varsha pursued a Masters in bioinformatics. She now works in the MGnify team, the world's largest metagenomic data resource.

Varsha uses metagenomics and metatranscriptomics to characterise the chicken and salmon gut microbiomes. This kind of work is essential for developing new, sustainable feeds for farmed animals, and improving food security worldwide. Varsha also contributes to the MGnify analysis pipelines and is one of the trainers on EMBL's Metagenomics Bioinformatics course, supporting others to access and analyse metagenomic data.

Asked what she would have become if she wasn't a scientist, Varsha replied: *"I went home one day from school and startled my parents with the news that microbes are the new "cool" – so I'm not sure there is any other career for me."*

Integrating genomics into the clinic

Many countries have emerging personalised medicine programmes and research initiatives which generate useful data for understanding human health and disease. These datasets reveal deeper insights when compared to data from other countries. For this reason, safe and secure data sharing across national borders is essential to help researchers understand the causes of diseases, such as cancer or rare diseases, and to develop new medicines and treatments.

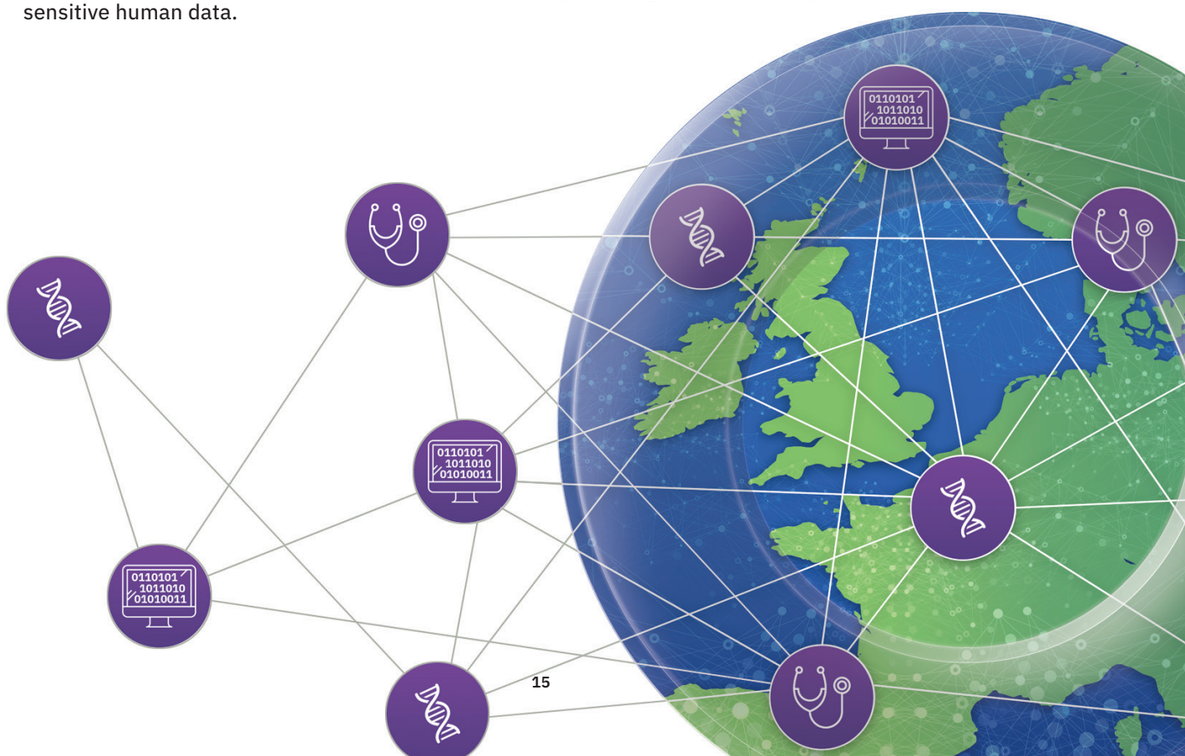
In 2022, EMBL-EBI initiated two major initiatives to improve controlled-access to genomic data across borders.

The Federated European Genome-phenome Archive (FEGA), a data resource that is jointly run by EMBL-EBI and the Centre for Genomic Regulation in Spain, announced its first five nodes in 2022, with Finland, Germany, Norway, Spain and Sweden joining efforts to create one of the world's largest networks for the discovery and access of sensitive human data.

Building on FEGA, EMBL-EBI is also supporting the European Commission-funded Genomic Data Infrastructure (GDI) initiative. This ambitious project aims to provide cross-borders access to at least one million genomes and associated clinical data, to enable European countries to reap the benefits of genomic medicine.

EMBL-EBI supports the GDI project in the development and implementation of common IT standards, and operating procedures across its vast international network.

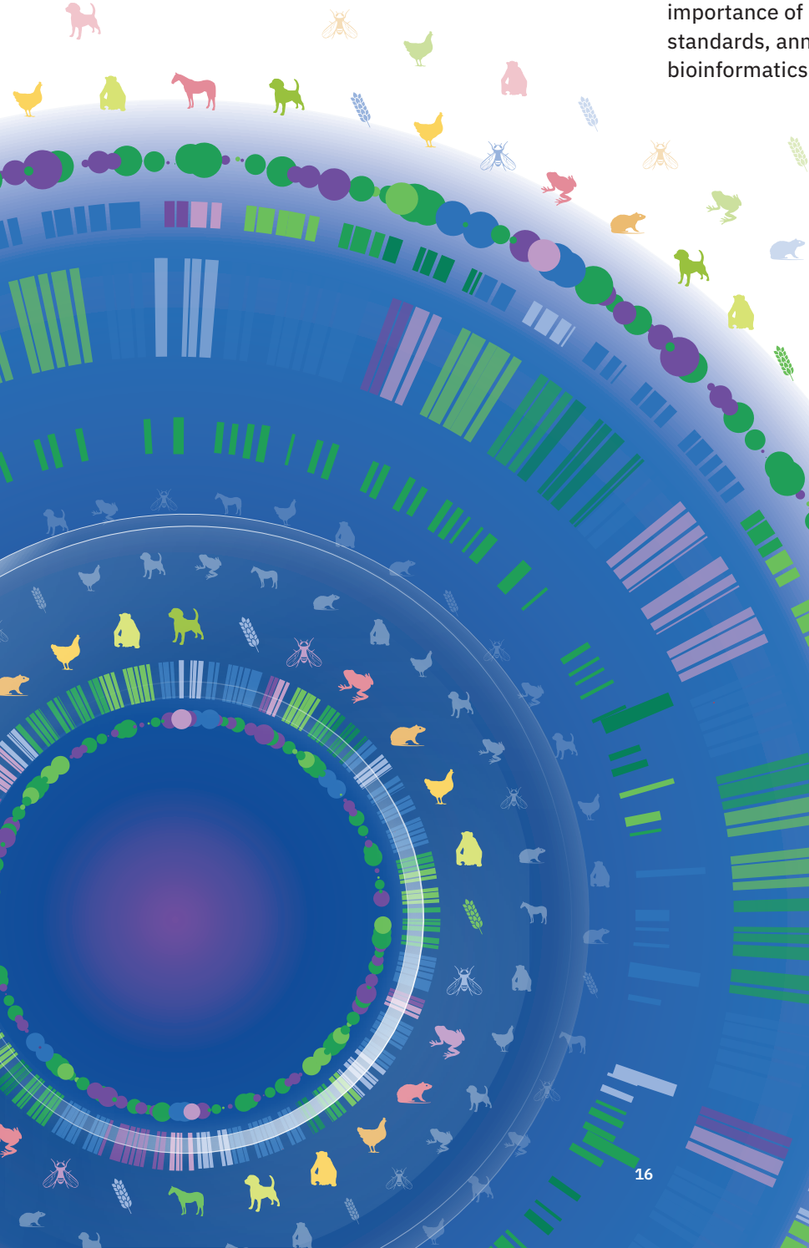
"It is fascinating to see many years of hard development work finally turn into a real service. I am positive this will be a major step forward for the biomedical research community in Finland and elsewhere, promoting the FAIR principles for sensitive research datasets."
Jaakko Leinonen, Senior System Specialist at CSC – IT Center for Science



Genomic sequencing in aid of biodiversity

Major genomic sequencing initiatives produced increasing volumes of data in 2022, which are rapidly becoming available in EMBL-EBI's data resources.

The two projects overleaf showcase how the EMBL-EBI data infrastructure can scale up to accommodate major biodiversity genome sequencing initiatives, and highlight the importance of knowledge exchange for data standards, annotation and training to support bioinformatics capacity building worldwide.



Darwin Tree of Life

The Darwin Tree of Life (DToL) initiative, which aims to sequence all the eukaryotic species in Britain and Ireland, has released its first 500 genome assemblies. By the end of 2022, 296 different species had already been annotated and made accessible through Ensembl Rapid Release and the DToL Data Portal, both developed by EMBL-EBI.

The DToL Data Portal is an open access platform which brings together all DToL data. Newly-implemented features allow users to track the sequencing progress of their species of interest, and explore an interactive map of where samples have been collected.

African BioGenome Project

The African BioGenome Project (AfricaBP) is an African-led effort to sequence the genomes of all species indigenous to the continent of Africa. It also aims to help build bioinformatics capacity for the researchers working there.

In 2022, EMBL and AfricaBP signed a memorandum of understanding which formalises the support that AfricaBP receives from the genome annotation and scientific training specialists in the Ensembl teams at EMBL-EBI.

"The partnership with Ensembl at EMBL-EBI provides knowledge sharing in the area of automated annotation, training, and workflow development. Giving access to the technical expertise and infrastructural resources of Ensembl will certainly advance the goals of the AfricaBP while developing and training local capacity."
Denye Ogeh, Bioinformatician, EMBL-EBI

Research

EMBL-EBI's research groups use computational methods and machine learning to make sense of vast, complex datasets. Our researchers work closely with experimental scientists worldwide, increasingly tackling problems of direct significance to medicine and the environment.

Leveraging long-read whole-genome sequencing for cancer research

Cancer is a genetic disease caused by changes – or mutations – in a person's DNA. Genomic sequencing is an essential tool for cancer diagnosis and treatment because it offers a snapshot of these mutations.

The Cortes-Ciriano group have initiated several projects that use long-read genome sequencing to better understand cancer and support healthcare professionals. Using nanopore sequencing, they are exploring whether a simple blood test could help to anticipate the emergence of cancer relapse in children. The project is a collaboration with the Institute of Cancer Research and Great Ormond Street Hospital.

Another project, this time in collaboration with the University of Birmingham, is following 300 individuals who have Lynch Syndrome, a rare hereditary condition associated with an increased risk of developing cancer. The researchers want to find out whether invasive tests, such as colonoscopies, could be replaced with blood tests, and whether these can be used to detect cancer earlier than other methods, such as imaging.

In parallel, the team is also developing software tools to address some of the main challenges of long-read sequencing technology: dealing with longer reads and the fact that the signal from the sequencer is noisier.

"Long-read sequencing has the potential to tell you where cancer is in the body in a quick and relatively cheap way through the analysis of methylation patterns in cell-free DNA. This could be a real game changer in cancer identification and management."
Isidro Cortes-Ciriano, Research Group Leader, EMBL-EBI

HILLARY ELRICK
PREDOCTORAL FELLOW

Transformative tools for cancer research

Despite long-read genome sequencing being increasingly used in cancer research, data analysis tools are lagging behind. Hillary Elrick, a Predoctoral Fellow in the Cortes-Ciriano group, has developed a transformative tool called SAVANA, which can detect somatic structural variants – or large changes in DNA – in cancer tumours. SAVANA unlocks new insights from long-read sequencing data, enabling researchers to understand cancer mutations better.

Before EMBL-EBI, Hillary studied biomedical computing at Queen's University in Canada. The internationality of the EMBL PhD programme and the proximity to Cambridge were two of the things that attracted her to EMBL-EBI.

"EMBL-EBI is one of the bioinformatics centres of the world, and attracts people from different countries. It's a unique place and a great opportunity to explore new ideas and gain useful skills."

Deep learning driving bioimage analysis

Deep learning has significant potential for improving bioimage analysis, from automation to better reproducibility and data integration. However, there is still a large gap between the life scientists producing and studying imaging data, and the developers designing novel deep learning methods.

Improvements are required on both sides: bioimage datasets need to be made available upon publication together with associated metadata, while the applicability of existing deep learning algorithms need to be clearer.

Virginie Uhlmann and collaborators have started a conversation about what is needed to make supervised deep learning methods easier to reuse, and what scientists who are not deep learning experts should consider when choosing a method for their study. The aim is to bring the two communities closer together.

On the data side, the new AI4Life initiative, which EMBL-EBI is involved in, is working with the community to develop annotation standards, and to improve how EMBL-EBI's BioImage Archive manages data annotations.

"There's an obvious need and appetite for reusing, fine-tuning, and improving deep learning models in bioimage analysis, and the opportunities are huge."
Virginie Uhlmann, Deputy Head of Research, EMBL-EBI

CRAIG RUSSELL

BIOIMAGE AI DEVOPS ENGINEER AND RESEARCH ASSOCIATE

At the junction of research and data services

After studying physics and engineering at the University of Cambridge, Craig Russell went from building microscopes to using machine learning to analyse the data they capture.

Craig currently sits at the junction of research and engineering. As a Research Associate in the Uhlmann group, he uses deep learning and data science to analyse microscopy data. This complements his role as an AI Devops Engineer in the BioImage Archive team, where he is developing pipelines that others can use to apply deep learning to their imaging data.

"As science and engineering become more intertwined, there is a need for people who straddle the two disciplines. This helps create accessible tools and pipelines that leverage new technologies to extract more value from data, quicker than ever before, often using artificial intelligence."

Spotlight on human and viral protein motifs

Short linear motifs (SLiMs) are stretches of 3-10 amino acids within a protein, typically with specific sequence patterns, that enable protein-protein interactions. SLiMs are important in promoting normal cell function. When mutations arise within SLiMs, it often results in disease and viruses often mimic SLiM patterns to interfere with host protein interactions.

Due to their small size, it is difficult to identify functional SLiMs. The Petsalaki group systematically identified 7,300 human SLiMs that are also present in viral proteins that interact with common human proteins, and integrated domain, structural and disease variant information.

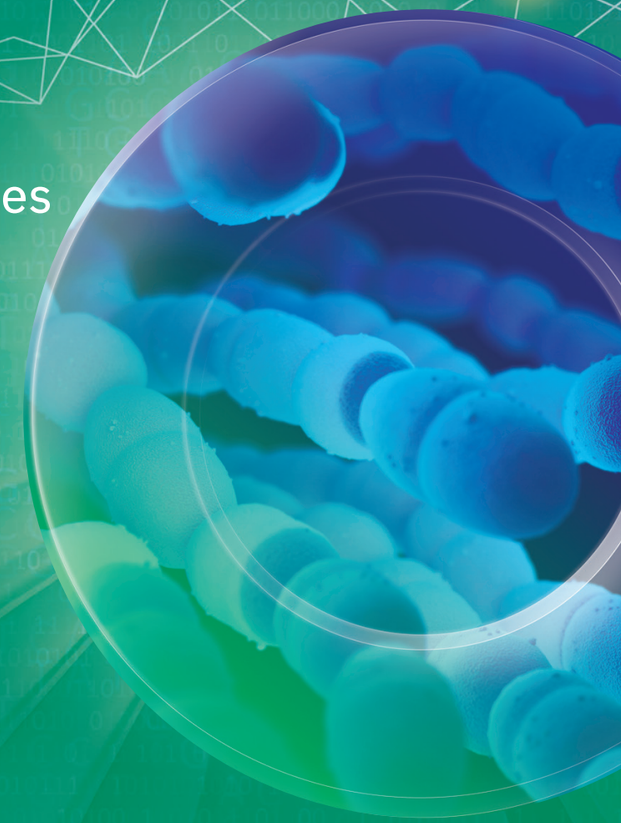
The group has developed an openly-available analysis pipeline that improves the identification of functional linear motifs shedding light on human SLiM functions and the associated mechanisms of viral interference.

Guiding vaccine strategies

The bacterium *Streptococcus pneumoniae* can cause a range of infections from mild sinusitis to dangerous pneumonia and meningitis, the latter being a significant cause of mortality in children and the elderly.

Infection biology rarely brings genomic evidence into vaccine design, but recent vaccine modelling studies have suggested that to more effectively immunise against invasive pneumococcal disease, genomics of the pathogen should be used. Specifically, the population genetic dynamics of the pathogen, and its distribution in adults and children, should be accounted for.

Based on two genomic datasets, findings from the Lees group and collaborators support proposals for future vaccination strategies which are primarily targeted at dominant circulating serotypes in specific pathogen populations, rather than the one-size-fits-all approach currently used.

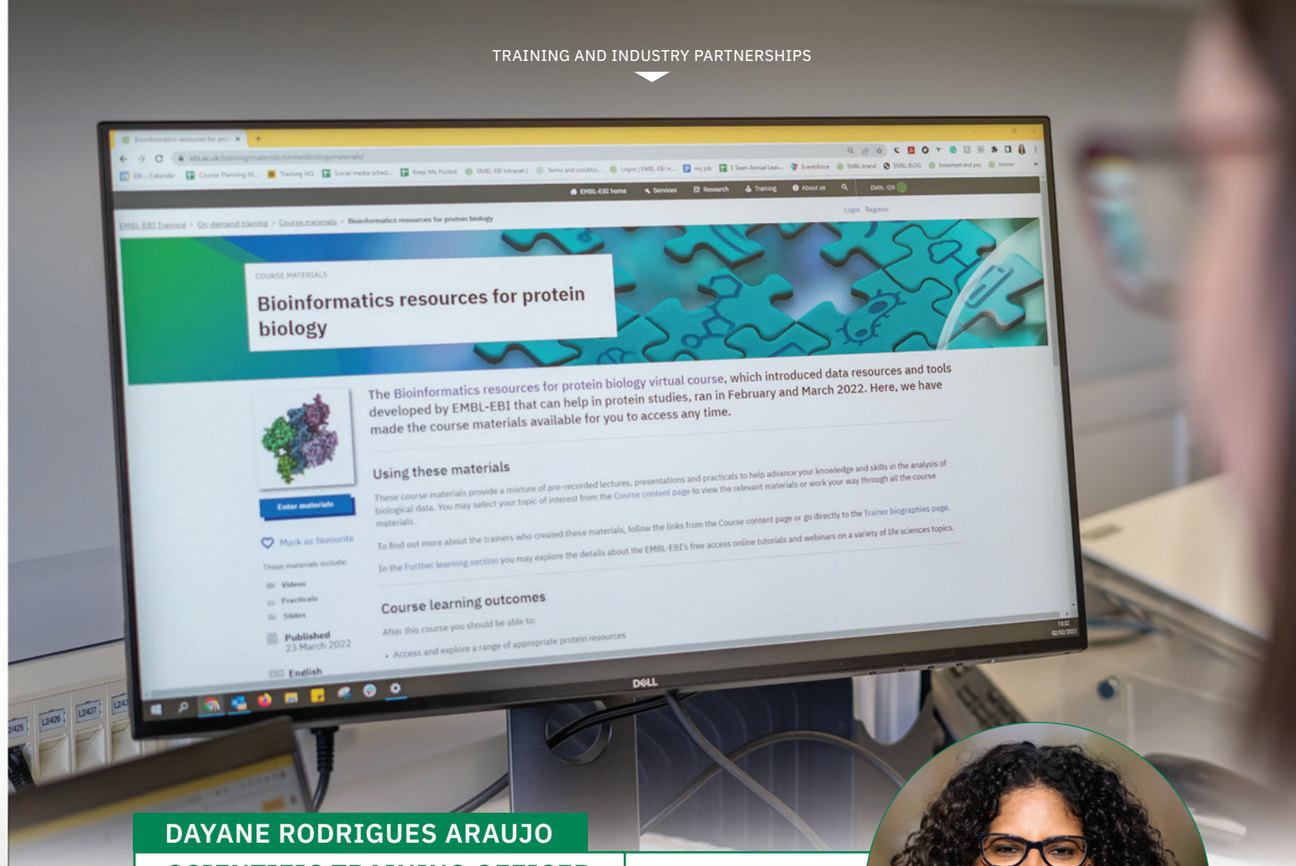


Training and industry partnerships

EMBL-EBI delivers world-leading training in biodata science. 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).

In 2022, EMBL-EBI Training resumed its lively programme of face-to-face courses, while maintaining a strong portfolio of live virtual courses, and on-demand digital learning.

This resulted in an increase in trainee numbers, also facilitated by the fact that users can now track their progress on the website.



DAYANE RODRIGUES ARAUJO

SCIENTIFIC TRAINING OFFICER

A passion for training and public engagement

As a child, Dayane Rodrigues Araujo was fascinated by how life worked, often performing impromptu insect dissections. She studied biology in Brazil, after which she completed a PhD in Microbiology in Germany.

Dayane joined EMBL-EBI as a Technical Outreach Officer for Europe PMC, the database for life science publications. There, she engaged with users, researchers, developers, policymakers and funders. She discovered a passion for training, which she was able to pursue by moving to EMBL-EBI's Training team.

Dayane now develops and delivers bioinformatics courses and is particularly pleased that, post-pandemic, the institute continued its virtual courses, which are more accessible and attract more diverse attendees.

"Science belongs to everyone. I'm part of the public engagement working group at EMBL-EBI, and love sharing the excitement of science with others. I think all scientific institutes that are publicly funded should do public engagement. It's a way to give back to society."

Hands-on bioinformatics internships

In collaboration with the French Embassy in London, EMBL-EBI runs a unique internship programme for students registered with French universities. The aim is to give students hands-on experience working on real-world projects, and to help them build their professional network in an international and cutting-edge environment.

Students don't require a biology background to apply, but many participants go on to study or work in the life sciences. This highlights the wide-range of careers available in the life sciences.

The programme has been running since 2017, and in 2022, funding was renewed for a further three years. The initiative is the first of its kind at EMBL-EBI, and could serve as a model for future collaborations.

"My internship went really well and afterwards I was hired at EMBL-EBI. The best part of my job is knowing that what I do is helpful for the world."
Eliot Ragueneau, former Intern currently Software Engineer, EMBL-EBI

"It's important to show that science has no borders and to invest in the young generation. So having students coming to EMBL-EBI to be trained in research, and bringing the knowledge and connections back to France is what we wish to do."
Minh-Hà Pham, Counsellor for Science and Technology, French Embassy in the United Kingdom

Consolidating industry collaborations

EMBL-EBI is increasingly working with the private sector on knowledge transfer projects. A major vector for collaboration is the institute's Industry Programme, a subscription-based initiative aimed at global companies that make significant use of EMBL-EBI's data resources for research and development.

In 2022, the Industry Programme membership grew to 30, with four companies joining: Exscientia, BenevolentAI, Moderna and Pierre Fabre.

The programme held nine workshops on cutting-edge topics, including one on artificial intelligence and the applications of AlphaFold for industry, which had over 250 attendees.

To support the many SMEs using EMBL-EBI open data resources, the institute also organised the second edition of the Bioinformatics for BioBusiness event, in collaboration with Medicines Discovery Catapult and UK Research and Innovation. The event was held at the University of Manchester, with a focus on accessing biomedical data for analysis.

Culture and infrastructure

To fulfil its missions, EMBL-EBI relies on a deeply-collaborative culture, a diverse workforce and robust technical infrastructure.

Thornton building construction begins

Construction has begun on EMBL-EBI's third building named after the institute's Director Emeritus, Professor Dame Janet Thornton. The team broke ground in September 2022, and construction is set to finish in 2024.

The building will serve as a space to expand the academic and pre-commercial collaborations needed to translate the institute's data management expertise into practical solutions for global challenges. Sustainability and cooperation are central principles for the building design.

Hybrid working pilot

To support staff wellbeing post-pandemic EMBL-EBI piloted more flexible working arrangements. This enables staff members and line managers to decide what proportion of their working time is spent at home and in the office. EMBL-EBI values working together in person and the opportunities for serendipitous encounters that this offers, but there is also a desire to empower staff to have more flexible working arrangements.

A staff survey performed at the end of 2022 showed that 86% of staff considered hybrid working a success and identified areas for improvement. The pilot continues in 2023.

Sustainability

Buildings and facilities

Against the backdrop of the energy crisis, EMBL-EBI implemented a range of measures to save energy and costs. This included reducing the temperature in the buildings, significant reduction in work-related travel and a communications campaign encouraging staff to be more mindful of their energy usage. Major changes were also made to the institute's technical infrastructure to deliver energy efficiencies.

Read about EMBL's sustainability strategy

MATTHIAS BLUM

DEVELOPMENT PROJECT LEADER

Measuring the carbon footprint of computing

Originally from France, Matthias Blum studied biology and became increasingly interested in bioinformatics during his Masters degree. After joining EMBL-EBI, he fully transitioned to computer science, software development and service provision.

Six years later, Matthias is a project leader for Pfam and InterPro, two essential data resources for protein families. He likes that he can still code, alongside managing a small team.

Matthias has an interest in sustainable computing, and has developed a tool to help EMBL-EBI staff measure the carbon footprint of their computing work.

"Just like we're making an effort to reduce travel or electricity consumption, we should be mindful of the carbon footprint of the computing jobs we run. As a test, we designed a tool that helps colleagues get a quick estimate."

Technical infrastructure updates and efficiencies

EMBL-EBI operates a vast technical infrastructure in order to facilitate the complex storage and compute requirements of its open data resources, research and training. A multi-year project began in 2020 following the contracting of a new, state of the art data centre. The aim is to migrate much of this infrastructure to newer, more performant and efficient hardware. A critical part of this project took place in 2022: the migration of essential production environments hosting our public-facing services and platforms.

This migration has allowed EMBL-EBI to perform an in-depth analysis of its existing infrastructure, resulting in the consolidation or retirement of numerous legacy storage systems and high performance computing clusters. This significantly reduces the institute's energy consumption and physical data centre space requirements, as well as streamlining systems to improve capacity and efficiency of data transfer.

EMBL-EBI is now able to organise data based on performance requirements, ensuring data that are heavily requested are accessible via high performance and highly available hardware. Data are now better separated into differing stages of lifecycle and at differing levels of security based on storage location.

These changes in architecture have meant that, despite increasing storage and computational capacity, EMBL-EBI has considerably reduced its overall power consumption, as well as ensuring its structures are robust and better capable of supporting inevitable future growth and expansion.

The EMBL-EBI website also underwent a major update, boasting a new content management system and visual framework, as well as increased accessibility and improved user journeys.

MARK HEAD

SENIOR DATABASE ADMINISTRATOR

From science to IT and back again

Mark Head's dream of becoming a pharmacologist was cut short by a severe allergy to animals. Following a brief detour as a trainee actuary, he went on to become a database administrator, working in insurance, retail and finance, before returning to science.

He and colleagues in the Systems Applications Database team organise, manage and archive data, and develop databases. The importance of being a Database Administrator is to ensure a huge amount of data is organised, optimised, performant, and available to users.

Mark also helped to coordinate the database side of EMBL-EBI's major data centre migration, ensuring the institute's data resources continued to be accessible to the world, with minimal downtime.

"I love the diverse community, the friendly atmosphere and the challenges of working with such huge volumes of data. It feels like we're working towards the greater good, and that's a breath of fresh air."

Lilla Takacs

HR-FINANCE ADMINISTRATOR

Seizing the opportunities

After studying Economics and HR Management, Lilla worked as a housekeeping supervisor, while also volunteering alongside an HR manager to get relevant work experience. She then took an opportunity at Ford, helping to set up the company's HR department in Hungary.

Just one week after she started as a HR maternity cover at EMBL-EBI, the UK went into lockdown, so she had to quickly adapt to a new role and a new way of working. Soon after, she moved to a split role between HR and Grants, helping to optimise communication between the teams and joint processes.

"I feel at home at EMBL-EBI. It's international and everyone is incredibly friendly and supportive. There are plenty of development opportunities and exciting projects to get stuck in, you just have to keep your eyes open."

Public engagement

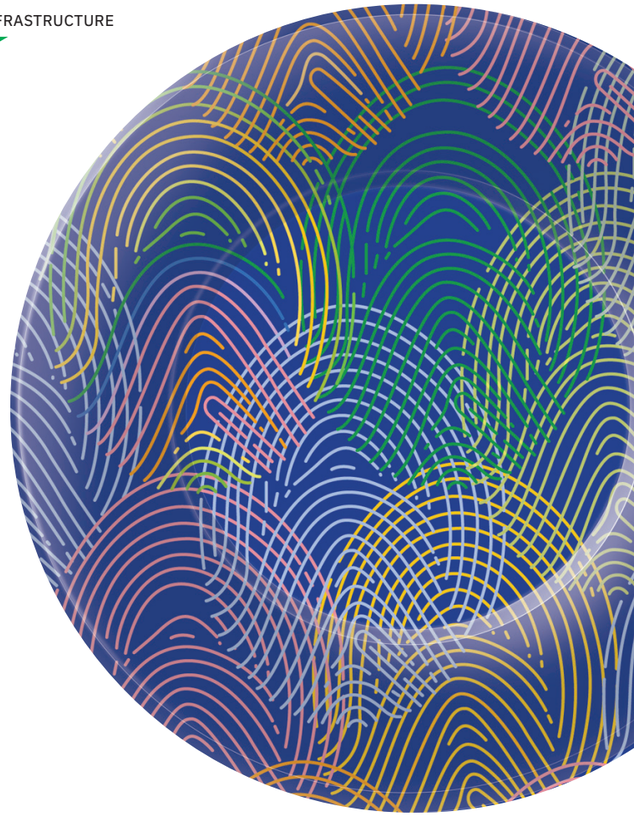
Building on our audience scoping project, EMBL-EBI has been working with the Ipswich Museum to connect audiences in Ipswich and Suffolk with science through activities from data resources such as PBDe and Ensembl Metazoa.

Working with ACCESS, a charity supporting migrants in East Anglia, EMBL-EBI has co-developed a local geocaching activity to engage migrant communities with the Darwin Tree of Life.

EMBL-EBI held its first public engagement fair in July, which brought together colleagues from across the institute with audience partners to explore our new public engagement activities.

Find out more about public engagement at EMBL-EBI

Equality, diversity and inclusion



Coordinated across the six sites of EMBL, equality, diversity and inclusion (EDI) is a priority at EMBL-EBI. In 2022 the Equality, Diversity and Inclusion Office continued the implementation of the EMBL EDI Strategy. The EDI Office is supported by the EDI Governing Body, the EDI Forum, and various teams across EMBL's six sites.

A newly-formed working group aims to strengthen EDI across the Ensembl project and the Genome Assembly and Annotation teams, focusing on interactions between staff, inclusive recruitment, staff development, inclusive leadership, outreach, and public engagement.

2022 also saw the wrap-up of the first cohort of the Leadership and Excellence for Aspiring Postdocs (LEAP) programme, a successful initiative to support women postdocs with progressing to leadership positions in scientific research. A second round of LEAP also commenced in 2022 for a further 17 female postdocs.

SARAH DYER

NON-VERTEBRATE GENOMICS TEAM LEADER

Leading a team part time

Sarah Dyer leads three areas: Ensembl Plants, Metazoa and Outreach. Her background in plant and crop genomics perfectly complements her interest in pests and pathogens.

Sarah's team are the custodians of hundreds of genomes that are essential for agriculture and food security, accessed by researchers around the world. They also provide training and outreach, with a focus on low- and middle-income countries.

"I currently work four days a week in my role at EMBL-EBI. It's great that EMBL-EBI gives staff the opportunity to do this. Working part time can be feasible even when you lead a team.

New leadership

Meet the group and team leaders who joined EMBL-EBI in 2022.

Mary Barlow

Mary Barlow was promoted to Head of Campus Operations and Capital Projects. Her team leads major capital investments projects, including Thornton building development, operational change projects such as pan-EMBL process improvements, the facilities and health and safety teams and coordination with the existing campus and wider development.

John Lees

John Lees joined as a Research Group Leader, working on pathogen informatics, with a focus on genome evolution, and the effects of vaccines and antimicrobial resistance in bacterial populations.

Andy Cafferkey

Andy Cafferkey was promoted to Head of Technical Services in charge of the institute's vast technical infrastructure spanning data centres, storage and internal IT requirements.

Fergal Martin

Fergal Martin was promoted to Team Leader for Eukaryotic Annotation. His team supports major biodiversity projects including Darwin Tree of Life and Earth BioGenome.

Ruth Sandland

Ruth Sandland joined as Head of Grants, supporting the institute in securing, negotiating and managing grant funding.

Virginie Uhlmann

Virginie Uhlmann was promoted to Deputy Head of Research, and will help to embed EMBL-EBI research within EMBL's new scientific programme. Virginie continues as a Research Group Leader.

Facts and figures

Staff numbers

Personnel categories in 2022*

672

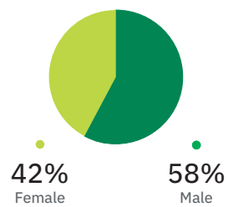
601
Staff members

36
Postdoctoral fellows

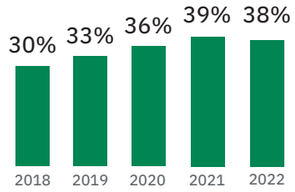
35
Predoctoral fellows

*Full-time equivalent (FTE)

Gender distribution of staff in 2022



Senior roles held by women

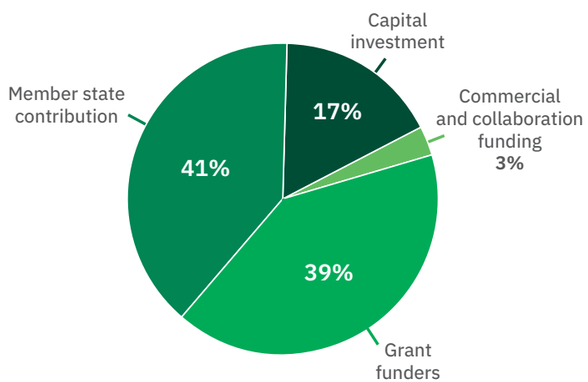


Financial figures

We are grateful to EMBL member states and other funding bodies for their continued support. They enable us to maintain and grow our data resources, conduct vital research and training, and respond to the changing requirements of the international scientific community.

Funding sources

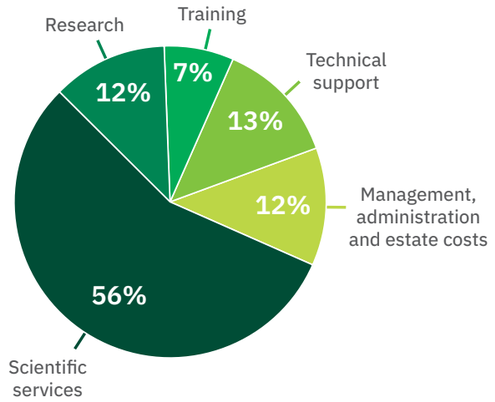
EMBL-EBI receives its funding through four main funding channels. The operating expenditure is funded by either EMBL member state contributions (41%), from commercial and collaboration funding (3%) or from external funding bodies via grant awards (39%). EMBL-EBI also receives significant capital awards for investment in buildings and data infrastructure (17%).



Operating expenditure

The total operating expenditure of EMBL-EBI in 2022 was €96.5m from all funding sources. Scientific services accounted for 56% of the expenditure in support of EMBL-EBI's data resources, with a further 12% allocated to research and 7% to external training.

Approximately 13% of costs were on technical support which maintains and develops the technical/IT infrastructure, and 12% on management, administration, and estate costs.



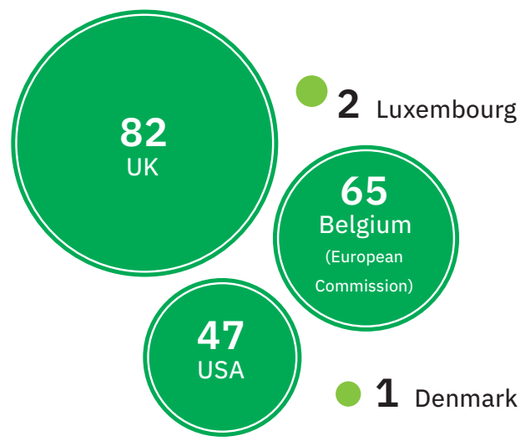
Capital investment

EMBL-EBI's Data Infrastructure for Life Sciences programme (2019 - 2024) oversees over £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.

In addition, EMBL-EBI's Thornton Building is under construction in 2023 with the project set to open in summer 2024. The building, funded by UK Research and Innovation (UKRI), the Biotechnology and Biological Sciences Research Council (BBSRC), and Wellcome, will be named after the institute's Director Emeritus and esteemed scientist Professor Dame Janet Thornton.

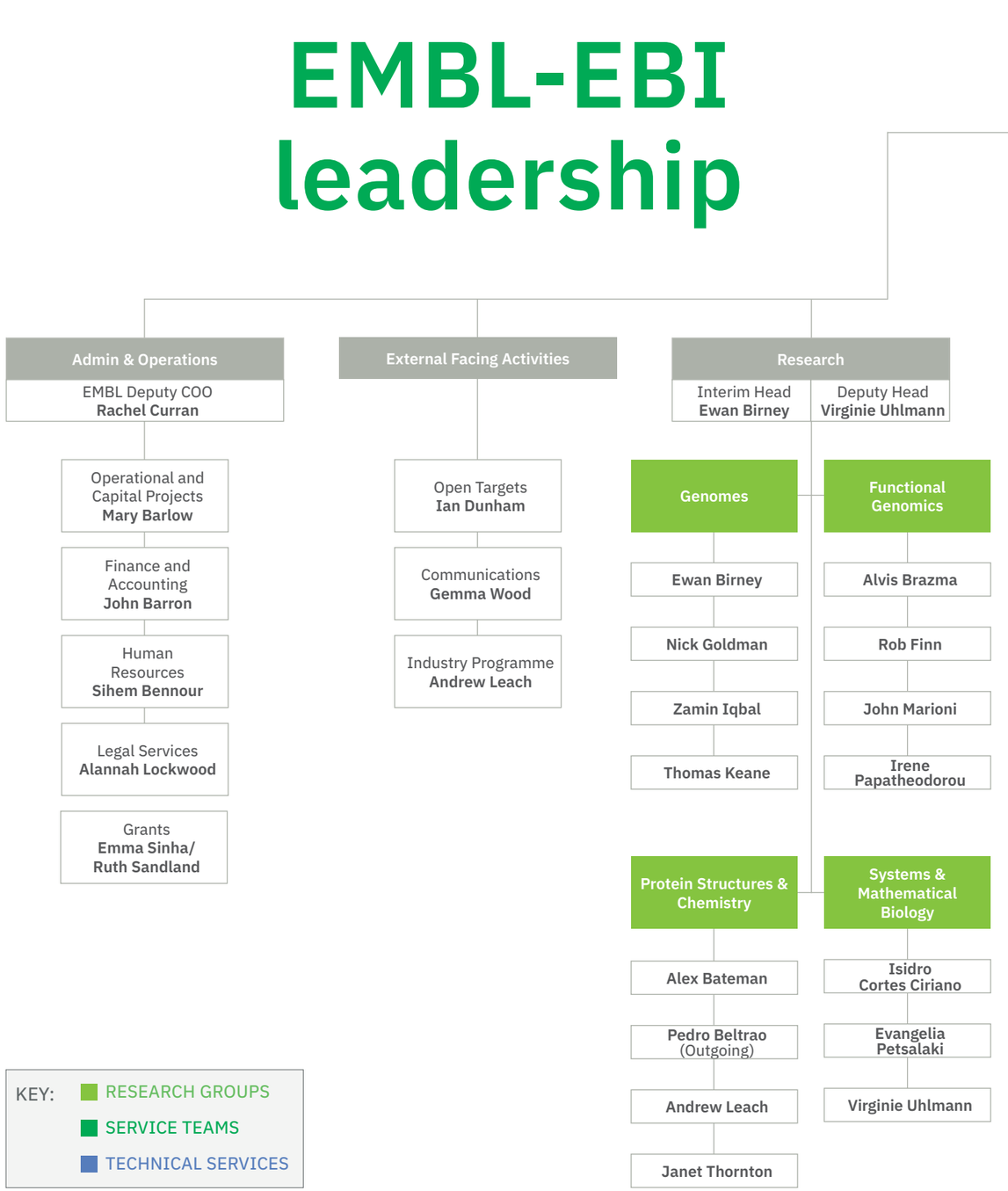
Strategic partners and funders

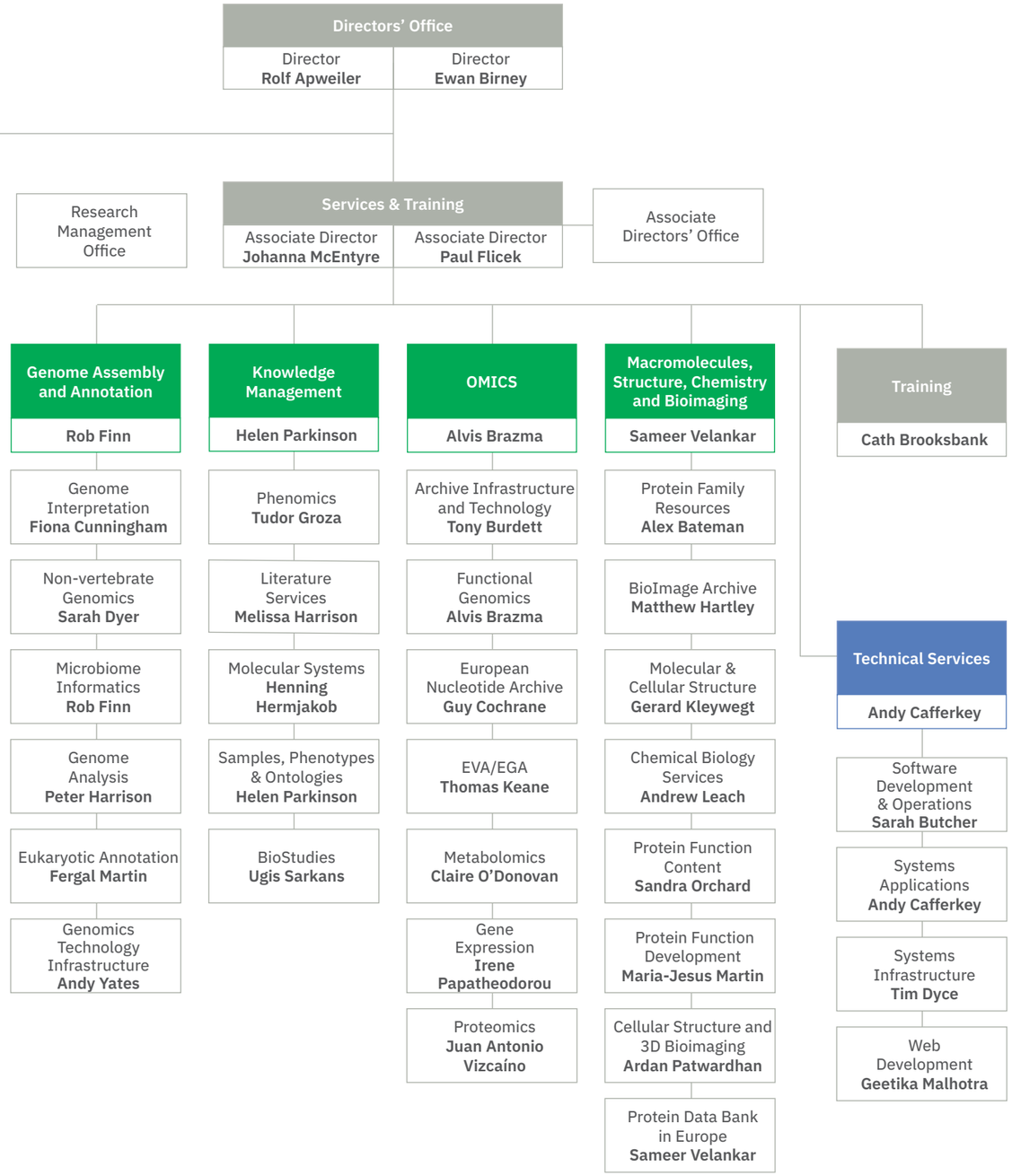
EMBL-EBI works closely with strategic partners and funders from across the globe. Alongside funding from EMBL member states, in 2022 we received €41m in research and service grant funding from 20 funders providing support for 197 projects.



Number of grants received by EMBL-EBI in 2022 by funding body location.

EMBL-EBI leadership





EMBL-EBI leadership during 2022

Data resource portfolio

Chemicals, molecules and drug discovery

- ChEBI
- ChEMBL
- MetaboLights
- Open Targets
- SureChEMBL

Genes, genomes and RNA

- Ensembl
- European Nucleotide Archive
- Expression Atlas
- HGNC
- MGnify
- Rfam
- RNAcentral
- VEuPathDB
- WormBase

Proteins

- AlphaFold DB
- Enzyme Portal
- InterPro
- PDBe
- PDBe-KB
- Pfam
- PRIDE
- 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
- DECIPHER
- 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

Governance, management and funders

EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an intergovernmental organisation with 28 member states, one associate member state and one prospect member state. EMBL is led by 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.

EMBL-EBI is led by joint Directors Rolf Apweiler and Ewan Birney, with the support of Associate Director of Services, Johanna McEntyre and the Head of Administration and Operations, Rachel Curran. In 2022, two members of the Senior Management Team moved on: Head of Research, John Marioni, and Associate Director of Services, Paul Flicek. We are grateful to both for their contributions and leadership and wish them well in their new roles.

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
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
UK Biotechnology and Biological Sciences Research Council
UK Research and Innovation
Wellcome

Imprint

Publisher
EMBL-EBI

Portrait image credits
p. 10 (John Jumper): DeepMind
p. 11 (Max Bileschi): Max Bileschi
p. 23 (Dayane Rodrigues Araujo): Mateusz Kochanowicz

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

Other image credits
p. 7 (Hybrid working): Kinga Lubowiecka/EMBL
p. 30 (Public engagement): Phil Mynott
p. 30 (Public engagement - circle): Eleanor Root

Printer
Healeys. healeys-printers.co.uk

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, Estonia, 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: Latvia