

Annual Report

2019



ebi.ac.uk

European Bioinformatics Institute (EMBL-EBI)

© 2020 European Molecular Biology Laboratory

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

Cover illustration: Spencer Phillips

For more information about EMBL-EBI please contact:

comms@ebi.ac.uk

Table of contents

Who we are	4
Foreword	7
EMBL-EBI 25th anniversary	8
What we achieved in 2019	10
2019 in numbers	11
Highlights of the year	12
Progress against our strategy	14
Increasing usage, utility and application of bioinformatics	15
Extending collaboration and coordination	29
Continuous improvement, maximising efficiency	34
Building capacity and capability	36
Supporting global expansion of biomolecular resources	42
Key information	44
Financial figures	45
Organisation of EMBL-EBI leadership in 2019	46
Our governance	48
Our funders	50
List of acronyms	51

Who we are

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

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

Our vision

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

Our missions

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

Our strategic priorities

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



Foreword

Although this report reflects EMBL-EBI's achievements in 2019 it was compiled in 2020, in the midst of the COVID-19 pandemic, a crisis unprecedented in modern times. For the first time in history, scientists are a constant presence in the news and science is at the forefront of public debate. At no other time has the public's trust in science been so crucial to the common good.

This pandemic has shown us that the significant advances science has made in recent decades are only a modest fraction of what is needed in the face of a global health crisis. In order to achieve impactful and long-lasting progress we need effective international collaboration, backed by trusted expertise and an open dialogue with the public.

Our work up to date demonstrates our commitment to these principles. Our achievements in 2019 were only possible because of the cooperative, international nature of our institute. Building on its 25 years of international collaborations in bioinformatics and data brokering, EMBL-EBI is working with the international scientific community to drive a new era of world-wide scientific cooperation. We are developing new ways for researchers, clinicians and public health professionals to share data, gather new insights and build on each other's discoveries.

In response to the current crisis, in April 2020 we worked with our partners to launch the COVID-19 Data Platform, which aims to facilitate sharing and analysing data to accelerate coronavirus research. Our role in the global fight against COVID-19 has highlighted the privileged position EMBL-EBI has earned as a source of trusted expertise. We aim to use this position to broker new scientific partnerships and to coordinate global research efforts that can drive science forward.

Now more than ever, the world needs scientists to work together. Our efforts cannot end with the current pandemic: we must look to the future, using what the current crisis has taught us to forge new ways of doing science that can keep up with our ever-changing world.

We would like to thank all our collaborators, partners and funders for their support over the last 25 years, and we look forward to many more years of furthering open data and enabling scientific discovery.

Sincerely,

Rolf Apweiler, Joint Director



Ewan Birney, Joint Director



EMBL-EBI 25th anniversary

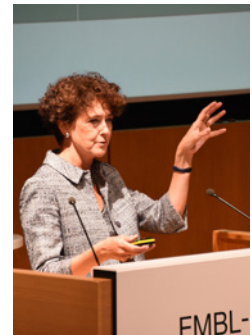
In September 2019, EMBL-EBI celebrated its 25th anniversary, marking the sustained growth that enabled it to become a global hub for bioinformatics services and research, as well as a resource for the international life science community. We would like to thank our users, collaborators and funders for their continued support without which none of this would have been possible.

In 25 years, EMBL-EBI grew from a small group of eight people to a world-leading institute, with over 800 staff and visitors from more than 70 countries. During this journey, we have strived to serve a growing number of scientists, across multiple disciplines, as well as constantly improve our technical infrastructure to support the deluge of biological data of the last quarter of a century. In line with these efforts, we have seen the discipline of bioinformatics move from scientific obscurity to the main stage in the life sciences.

EMBL-EBI will continue to be the keeper of the world's biological data well into the future, adapting its offering to further empower scientific discovery in Europe and the world.



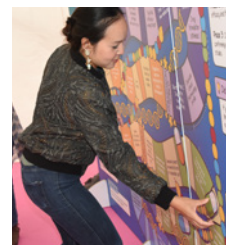
EMBL-EBI and EMBL current and past leaders. Left to right: Paul Flicek, Jo McEntyre, Rolf Apweiler, Edith Heard, Michael Ashburner, Ewan Birney, Janet Thornton, Graham Cameron.



The anniversary began with a scientific symposium about the evolution and applications of bioinformatics.



All photos: Phil Mynott



The celebration was the perfect opportunity to test a range of interactive activities that will be repurposed for public engagement.

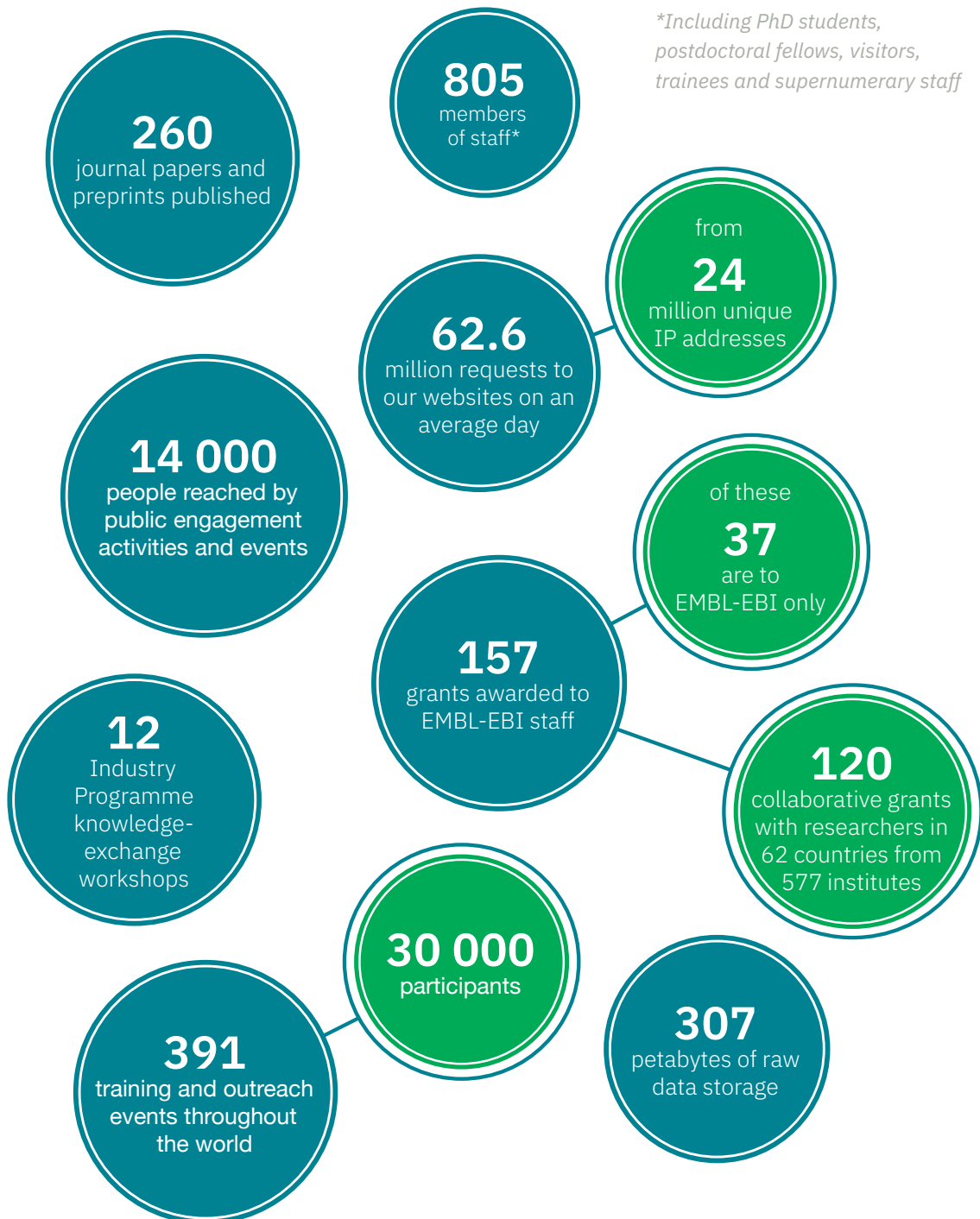


What we achieved in 2019

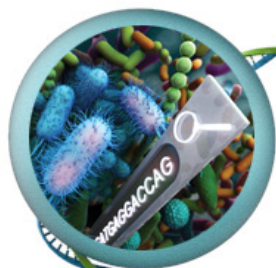
Our open data resources are used increasingly across the globe. In 2019 we focused on improving our technical infrastructure, strengthening our collaborations and enhancing our training programme.



2019 in numbers



Highlights of the year

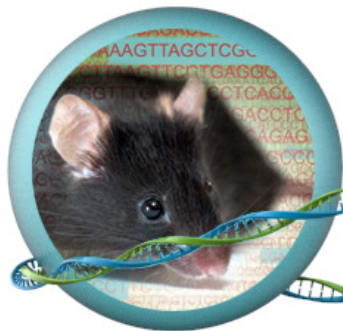


FEBRUARY

- ◎ BIGSI, the DNA search engine for microbial data, released (page 26)



- ◎ Researchers use computational methods to identify 2000 novel human gut bacteria (page 23)



MARCH

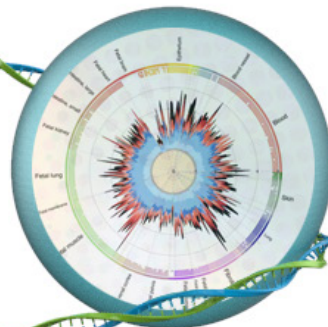
- ◎ UK Research and Innovation awards £45 million to EMBL-EBI to enhance technical infrastructure (page 45)
- ◎ A new knowledge base, PDBe-KB launches, enabling researchers to visualise all publicly available structure data for proteins of interest in an integrated manner (page 16)

APRIL

- ◎ Open Targets uses CRISPR to identify thousands of genes that are essential for cancer survival, making them good candidates for cancer therapeutics (page 23)
- ◎ Ensembl releases first complete version of manually-annotated mouse reference genome

MAY

- ◎ The European Genome-phenome Archive is awarded funding to store and share the first 50 000 genomes from UK Biobank (page 29)



JUNE

- ⦿ Reactome enables researchers who produce or review biological pathways to link their contributions to their ORCID (page 21)
- ⦿ European Variation Archive makes its first public data release of permanent identifiers for 750 million variants in 203 non-human species (page 18)

JULY

- ⦿ BioImage Archive, EMBL-EBI's dedicated data resource for reference biological images, launches (page 16)

AUGUST

- ⦿ Paul Flicek and Jo McEntyre start as Associate Directors of EMBL-EBI Services (page 38)
- ⦿ OmicsDI develops Omics score, a set of metrics that can help quantify the reuse and impact of biomedical datasets (page 19)

SEPTEMBER

- ⦿ EMBL-EBI celebrates its 25th anniversary (page 8)
- ⦿ Researchers use epigenetic clock to explore the molecular mechanisms that drive human ageing (page 23)

OCTOBER

- ⦿ EMBL-EBI's receives funding for the ARGENT project to pilot a global genomics-based tuberculosis monitoring service to help track the disease (page 26)



NOVEMBER

- ⦿ EMBL-EBI awarded funding for the Darwin Tree of Life project, set to sequence thousands of species in Britain and Ireland (page 27)
- ⦿ GlyGen, a new informatics portal for glycoscience, launches (page 31)
- ⦿ EMBL-EBI marks the 30th anniversary of the Convention on the Rights of the Child by working with UNICEF to store the text in synthetic DNA (page 41)

DECEMBER

- ⦿ Researchers apply machine learning to create the largest reference phosphoproteome resource to date of almost 120 000 human phosphosites (page 24)
- ⦿ Scientists create a single-cell resolution multiomics map of the mouse embryo during very early development (page 23)

Progress against our strategy

Increasing usage, utility and application of bioinformatics

Our data resources are central to international life science research and we serve a growing user base with very diverse needs. Our priority is to deliver the best possible quality of data resources to this community.

Data usage and data submission

A hallmark of 2019 was the high demand for our data services. On an average day, we saw more than 62.6 million requests to our websites, from almost 24 million unique IP addresses.

The heaviest usage, measured by IP addresses, comes from the USA (28%), China (16%), the UK (9%), Germany (7%) and France (4%).

We have also seen increasing amounts of data being submitted to our resources. To manage this growth, we are continually expanding our computational infrastructure. At the end of 2019, EMBL-EBI had 307 petabytes (PB) of raw storage, up from 273 PB in 2018.



Why do IP addresses matter?

The number of IP addresses is an indication of the number of users, but not an exact count. Sometimes an entire organisation uses a single IP address, while almost all users have multiple IP addresses. Because our data resources are open, meaning users don't need to sign in, it is incredibly difficult to estimate the exact number of users, but the number of IPs is a useful indicator.

What is a request?

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

Data resources update

Launching new data resources

In 2019 we released several new data resources to help the global scientific community share and analyse data in a collaborative way.

BioImage Archive

To address the growing needs of the bioimaging community, EMBL-EBI launched the BioImage Archive, a new data resource that archives image data from the molecular to organism scale. This builds on existing EMBL-EBI resources for image data, including BioStudies and EMPIAR. The BioImage Archive hosts reference datasets that can be reused and re-analysed, enabling researchers to gain new scientific insights from complex imaging studies.

www.ebi.ac.uk/bioimage-archive

Protein Data Bank in Europe Knowledge Base

The Protein Data Bank in Europe (PDBe) launched a community-driven resource for structural and functional annotation, called PDBe Knowledge Base (PDBe-KB). The new resource offers aggregated views for protein structures. It provides mechanisms for exploiting data and functional annotations to support the broader scientific community.

www.pdbe-kb.org

In the pipeline

We have also worked on developing new resources that launched in 2020:

- ⦿ eQTL Catalogue provides uniformly processed gene expression and splicing quantitative trait loci (QTL) from a wide range of studies on humans with a view to cover all available studies in a few years. <https://www.ebi.ac.uk/eqtl/>
- ⦿ To address the global emergency caused by the COVID-19 pandemic, EMBL-EBI launched the COVID-19 Data Platform, which enables researchers to upload, access and analyse COVID-19 datasets. The aim is to facilitate data sharing and analysis, and to accelerate coronavirus research. The COVID-19 Data Platform leveraged the existing data infrastructure of EMBL-EBI's Pathogen Portal for food-borne pathogens. www.covid19dataportal.org

Improvements to our data resources

ChEMBL

ChEMBL, the manually curated database of bioactive molecules with drug-like properties, has a new interface. Key features include enhanced user experience, interactive filtering, new visualisations, and a method for structure similarity search able to operate over tens of millions of structures. Users can also generate a URL for sharing visualisations and data.

www.ebi.ac.uk/chembl

Data mining to make sense of microscopy



Andrei Istrate is a Scientific Programmer in the Cellular Structure and 3D Bioimaging team. He is from the Republic of Moldova, and he trained as a structural biologist in Moscow, after which he worked in Switzerland on nuclear magnetic

resonance (NMR) spectroscopy.

At EMBL-EBI, Andrei is part of the team that runs the Electron Microscopy Data Bank, a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.

“Researchers are using electron microscopy for lots of things, but mainly to understand what certain molecules look like,” explains Andrei. “Microscopy images are not perfect; there is often a lot of ‘noise’ so it’s hard to know exactly what we’re looking at. To help researchers better understand what they are seeing, we used data mining to analyse cryo-electron microscopy (cryo-EM) images. Essentially, we developed a pipeline that takes existing data into account and identifies potential issues with new data that comes in.”

The pipeline Andrei developed is now used to validate new EMDB data and to spot errors. This improves the overall data quality, which is beneficial for EMDB users. The pipeline also has potential for other applications, including model building.

Electron Microscopy Data Bank (EMDB)

EMDB, the public repository for electron microscopy data, implemented map and map/model validation for new depositions, to improve data quality. The growing importance of EMDB as a resource for structural biology drives home the urgent need for validation of deposited structures, and the team is working with collaborators and journals to make this happen.

www.ebi.ac.uk/pdbe/emdb/

Electron Microscopy Public Image Archive (EMPIAR)

To make imaging data easier to find and analyse across EMBL-EBI resources, EMPIAR now links to light microscopy datasets deposited in BioStudies. Additional improvements enabled scalability to support the growing number of submissions and the launch of the BioImage Archive.

www.ebi.ac.uk/pdbe/emdb/empiar/

Ensembl

The alpha version of the new Ensembl website was launched in 2019. It allows users exploring genome data to transition from the chromosomal scale to base-pairs within seconds and uses a suite of in-house web APIs to deliver data efficiently to users. The LiteMol application now allows users to visualise genomic variation in a 3D protein structure context, seeing if a specific genomic change disrupts protein structure or ligand binding. The Epigenome Explorer has been redeveloped in response to increasing demand for regulation data. A new Ensembl Post-GWAS service permits users to upload GWAS summary statistics and compare them to available datasets.

www.ensembl.org

New Ensembl website:

<http://2020.ensembl.org>

European Nucleotide Archive (ENA)

The ENA launched a new browser, a sophisticated entry point into the archive, offering new functions including an advanced search function, embedded data quality visualisations and direct access to user support. The new browser will allow for greater flexibility and scalability in the future, enabling ENA to respond to the needs of its users by changing the features it offers and by adding new functionalities as necessary.

www.ebi.ac.uk/ena

Single-cell sequencing clusters in Single Cell Expression Atlas.

European Variation Archive (EVA)

As the sole international variation resource for human and non-human variation, the EVA made its first public data release of permanent identifiers for 750 million variants in 203 non-human species, a key milestone to realise the findability, accessibility, interoperability and reproducibility (FAIR) principles for genetic variation.

www.ebi.ac.uk/eva

Expression Atlas

Significant improvements have been made to the Expression Atlas and its sister resource, the Single Cell Expression Atlas (SCEA), which saw usage double across all continents in 2019. The SCEA now contains 132 single-cell RNA sequencing studies across 12 species, totalling 1.3 million cells. The single-cell RNA sequencing analysis pipelines have been expanded to accommodate droplet-based technologies.

www.ebi.ac.uk/gxa



SCEA

GWAS Catalog

In 2019, there was a 65% increase in associations added to the GWAS Catalog compared to 2018. The GWAS Catalog focused on improving its user interface, with new pages that allow users to directly access all associations with a particular gene. The GWAS Catalog is also working with a number of large GWAS consortia to promote the sharing of summary statistics for their studies, which are used in meta-analyses, follow-on analyses and in the integration of association data with different data types. Moreover, GWAS Catalog is collaborating with UK Biobank to increase the availability of GWAS summary statistics from UK Biobank approved studies in the GWAS Catalog, with summary statistics now available for 44% of UK Biobank publications.

www.ebi.ac.uk/gwas

InterPro and Pfam

InterPro and Pfam evolved to keep pace with expanding volumes of UniProt sequence data and data from metagenomics studies, which are identifying novel sequences within protein regions whose structure is unknown, also called protein “dark matter”. The continuous assessment and integration of protein signatures into InterPro and the targeted generation of Pfam signatures are helping to bridge the gap between the known and unknown protein universe.

www.ebi.ac.uk/interpro

<http://pfam.xfam.org>

MetaboLights

MetaboLights continues to be one of EMBL-EBI’s fastest growing data repository, boasting an extended depth of the data. The submission procedures have been redeveloped to address the needs of the diverse user community. An online editor now guides the submitter, enriching the data and making it more interoperable.

www.ebi.ac.uk/metabolights

MGNify

The number of analysed datasets in MGNify has increased by about 50%. As the world's largest microbiome analysis resource, MGNify has expanded the scope of its taxonomic and functional analyses and has notably improved the interpretation of metagenomics assemblies.

www.ebi.ac.uk/metagenomics

Omics Discovery Index (OmicsDI)

OmicsDI, an open source platform that makes data from publicly available research discoverable and reusable, has developed a set of metrics that helps quantify the impact of omics datasets. The Omics score helps researchers track the impact and reuse of their datasets, while also being a useful tool for funding agencies and the wider scientific community.

www.omicsdi.org

PRoteomics IDentification Database (PRIDE)

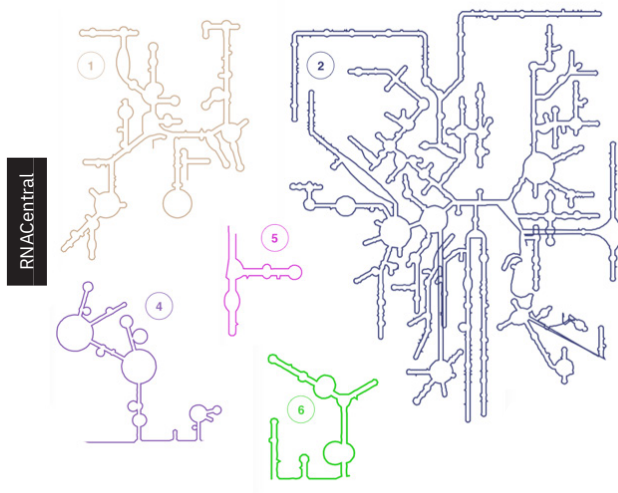
PRIDE saw a record number of over 3900 datasets submitted in 2019, and 625 Terabytes (TB) of data downloaded. The new PRIDE Archive infrastructure, including a new website and programmatic interface, was put in production, with a focus on user friendliness, reliability and scalability.

www.ebi.ac.uk/pride/archive

RNAcentral

The non-coding sequence database RNAcentral reached a milestone of 40 member databases and over 16 million non-coding RNA sequences. A new RNA visualisation pipeline generated over 8 million high-quality secondary structure diagrams. RNAcentral launched an updated, cloud-based version of the sequence similarity search, enabling users to filter search results by organism, RNA type or keyword.

<https://rnacentral.org>



Examples of human RNA secondary structures generated using the new RNAcentral pipeline.

Expanding our biocuration efforts

As one of the few places in the world that provides extensive curation and annotation for biological data, EMBL-EBI has been focussing on improving its process in order to address the needs of a growing user base.

Ensembl

The Ensembl team annotated 120 genomes in 2019, double compared to the previous year. The focus was on farmed and domestic animal genomics, as well as birds and reptiles.

The Ensembl/GENECODE project at EMBL-EBI and the RefSeq project at the National Centre for Biotechnology Information (NCBI) in the United States have provided independent high-quality reference gene datasets since the sequencing of the first human genome almost 20 years ago. A new collaboration between the two institutes, called Matched Annotation from the NCBI and EMBL-EBI (MANE), is providing a **matched set of transcripts for human protein-coding genes**. It will also define one representative transcript for each gene, in addition to the Ensembl and RefSeq alternate transcripts.

In 2019, Ensembl and RefSeq released a canonical genome-wide transcript set for 75% of coding genes with one well-supported transcript per protein-coding locus. Such a transcript set has been eagerly awaited by the clinical community and it is hoped that it will be set as a default across genomic resources.

Putting our users first



Sangya Pundir is a User Experience (UX) and Product Manager in the Protein Function Development team, which manages UniProt. After training as a bioinformatician, Sangya realised she was not just interested in the science and the data, but also in making data resources more accessible to the user community.

“My job combines UX and product management,” explains Sangya. “I speak to users about what

they want to achieve when employing our data services and I work with my team to address user needs in the most effective way.

“We are currently working on a major redesign of UniProt, driven by changes in the science, technology and user base. Traditionally, UniProt was used more by the proteomics, bioinformatics, molecular biology and structural biology communities. More recently we are seeing more genomics researchers, clinicians and clinical researchers, who tend to dip into the data to get a more comprehensive view of their gene or disease of interest.

“Our aim is to reshape UniProt so it can scale up to the increasing amount and complexity of data coming in. We aim to make it more flexible, easier to use and to update. This is more than just a face lift, it’s a major overhaul, and although it will take time, it will result in a modernised data resource that serves our users well.”

Reactome

The Reactome pathway database now enables researchers who produce or review biological pathways to link their contributions to their ORCID. The new functionality acknowledges the importance of sharing scientific data and expertise and helps researchers keep track of the contributions they make to Reactome.

UniProt

As our understanding of biology increases, the way we capture the information is changing. In collaboration with external consortia, UniProt has improved the annotation of key biological

processes, including DNA replication, gene regulation and pseudoenzyme functionality in human and key model animals.

The Protein Function Development team also **standardised and structured enzyme and sequence annotation**, and re-designed the Enzyme Portal. New data formats, such as Proteomics standard Extended Fasta Format (PEFF) were adopted.

A collaboration between Ensembl and UniProt resulted in the **creation of the GIFTS curation tool**, which allows the two databases to correct differences between their gene annotations, ensuring the consistency of data.

Providing better access to literature

As one of the world's largest open databases of life sciences literature, Europe PMC serves an increasing range of researchers with different demands.

In 2019, the Europe PMC website was redesigned to offer a streamlined search and reading experience. The new design offers a more comprehensive view of a paper's abstract, full text and figures, allowing users to see key scientific findings at a glance. Additional information, such as supporting data, impact metrics, open peer review and links to protocols are easily findable in the in-page menu.

To add even more value to the literature, Europe PMC also established the annotation submission system, allowing different providers to share their text-mined or curated annotations. Users can now search by type of annotation and by the annotation provider using the Advanced Search.



31 million
abstracts



6 million
full-text articles



76 700
grants



94 000
preprints



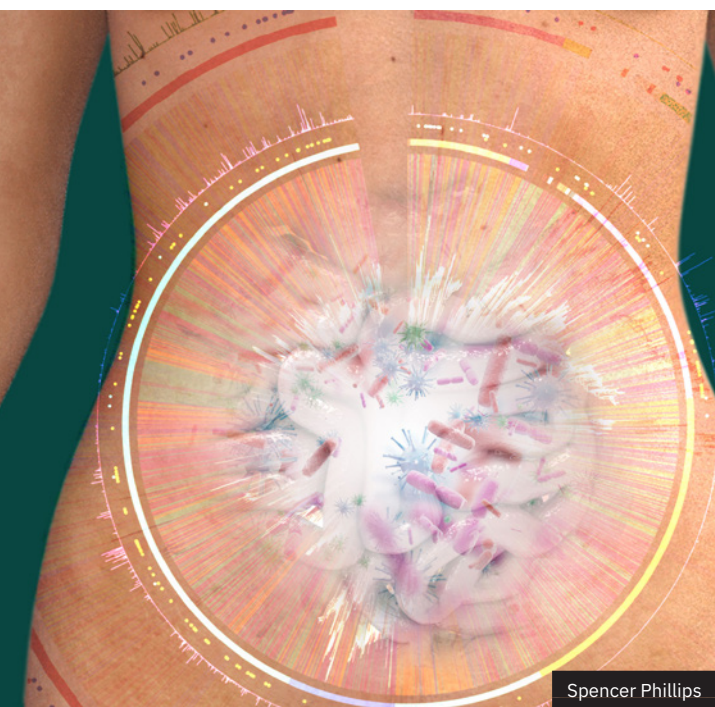
2.8 million
Open Access
articles

Research highlights

Charting the unknown

Despite extensive research in recent years, the gut microbiome remains a very prolific field of study. Researchers in the Finn group used computational methods to identify nearly 2000 uncultured gut bacterial species. The information allows researchers to characterise the gut microbiota more accurately and to further explore its links to human health and disease.

Almeida, A., et al. (2019). A new genomic blueprint in the human gut microbiota. *Nature*. DOI: [10.1038/s41586-019-0965-1](https://doi.org/10.1038/s41586-019-0965-1)



Spencer Phillips

Artist's impression of human gut microbiome

Using the Horvath epigenetic clock, the Thornton group explored the molecular mechanisms that may drive ageing in humans. They found one gene, called NSD1, that seems to be closely linked to the process. This type of research could advance our understanding of ageing.

Martin-Herranz, D. E., et al. (2019). Screening for genes that accelerate the epigenetic aging clock in humans reveals a role for the H3K36 methyltransferase NSD1. *Genome Biology*. DOI: [10.1186/s13059-019-1753-9](https://doi.org/10.1186/s13059-019-1753-9)

The Marioni group used a multiomics approach to define how cell identities are established in the early stages of mouse embryonic development. They created a single-cell resolution epigenetic map of the three primary germ layers during in embryo development. The results identify germ layer specific timings required to prime the different cell types for lineage specification.

Argelaguet, R., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*. DOI: [10.1038/s41586-019-1825-8](https://doi.org/10.1038/s41586-019-1825-8)

Understanding disease

Open Targets researchers and collaborators used CRISPR technology to disrupt every gene in over 300 cancer models. The results revealed thousands of genes that are essential for cancer survival, making them potential candidates for cancer therapeutics.

Behan, F.M., et al (2019) Prioritisation of oncology therapeutic targets using CRISPR-Cas9 screening. *Nature*. DOI: [10.1038/s41586-019-1103-9](https://doi.org/10.1038/s41586-019-1103-9)

Most human protein-coding genes are regulated by multiple promoters, but the contribution of alternative promoters has been largely unexplored so far. In a joint study with the Genome Institute of Singapore, the Brama group used RNA sequencing data to infer active promoters from 18 468 cancer and normal samples. The study shows how a dynamic landscape of active promoters shapes the cancer transcriptome.

Demircioğlu, D., et al. (2019). A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell*. DOI: [10.1016/j.cell.2019.08.018](https://doi.org/10.1016/j.cell.2019.08.018)

The Birney group in collaboration with Moorfields Eye Hospital (UK) are performing genome-wide association testing using UK Biobank data from over 65 000 individuals who have genetic data records and optical coherence tomography (OCT) scans. The researchers performed a comparison of associations with different inner retinal thickness parameters.

Khawaja, A.P., et al. (2020). Comparison of Associations with Different Macular Inner Retinal Thickness Parameters in a Large Cohort: The UK Biobank. *Ophthalmology*. DOI: [10.1016/j.ophtha.2019.08.015](https://doi.org/10.1016/j.ophtha.2019.08.015)

Exploring the protein universe

The Beltrao group in collaboration with the Proteomics and Open Targets teams, generated the largest reference phosphoproteome resource to date. They used machine learning to define a functional score that ranks 120 000 phosphosites according to functional importance. Identifying new functional phosphosites has enormous potential to progress research into many biological processes and diseases.

Ochoa, D., et al. (2019). The functional landscape of the human phosphoproteome *Nature Biotechnology*. DOI: [10.1038/s41587-019-0344-3](https://doi.org/10.1038/s41587-019-0344-3)



Spencer Phillips

Artist's impression of exploring the phosphoproteome using deep learning.

The Goldman group developed a novel mathematical model of protein sequence evolution that combines amino acid sequence information with protein structure information. The model outperformed previous approaches and improved inference of phylogenetic trees and reconstruction of ancient protein sequences.

Perron, U., et al. (2019). Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States. *Molecular Biology and Evolution*. DOI: 10.1093/molbev/msz122

Decoding cancer images

Yu Fu, a postdoctoral fellow in the Gerstung Research Group, studied applied mathematics and obtained her undergraduate degree in China. Undecided on her future career path, she seized an opportunity to study for a master's degree in disease modeling in France, where she developed her interest in cancer research and image analysis, leading her to pursue a PhD in the field. Her skills made her a perfect match for the Gerstung Group, which investigates the mechanisms of cancer development and treatment using computational biology.



At EMBL-EBI, Yu developed a machine learning algorithm that analyses tumour tissue images to distinguish healthy cells from cancer and to predict genomic changes and survival. “We couldn’t have done this without the computing resources at EMBL-EBI,” she says.

The study has been published in *Nature Cancer*, but Yu doesn’t want to stop there. “There still is a long way to go but I really hope we can implement this method in a clinical setting and help patients,” she says. “Our algorithm can analyse a tumour image in the blink of an eye and is much cheaper to use than sequencing. That would enhance the work of pathologists immensely.”

Increasing the application of bioinformatics

Interdisciplinary research and collaboration enable EMBL-EBI to develop new technologies and tools for a wide range of applications, strengthening the utility and usage of bioinformatics in healthcare, agriculture and biodiversity.

Healthcare

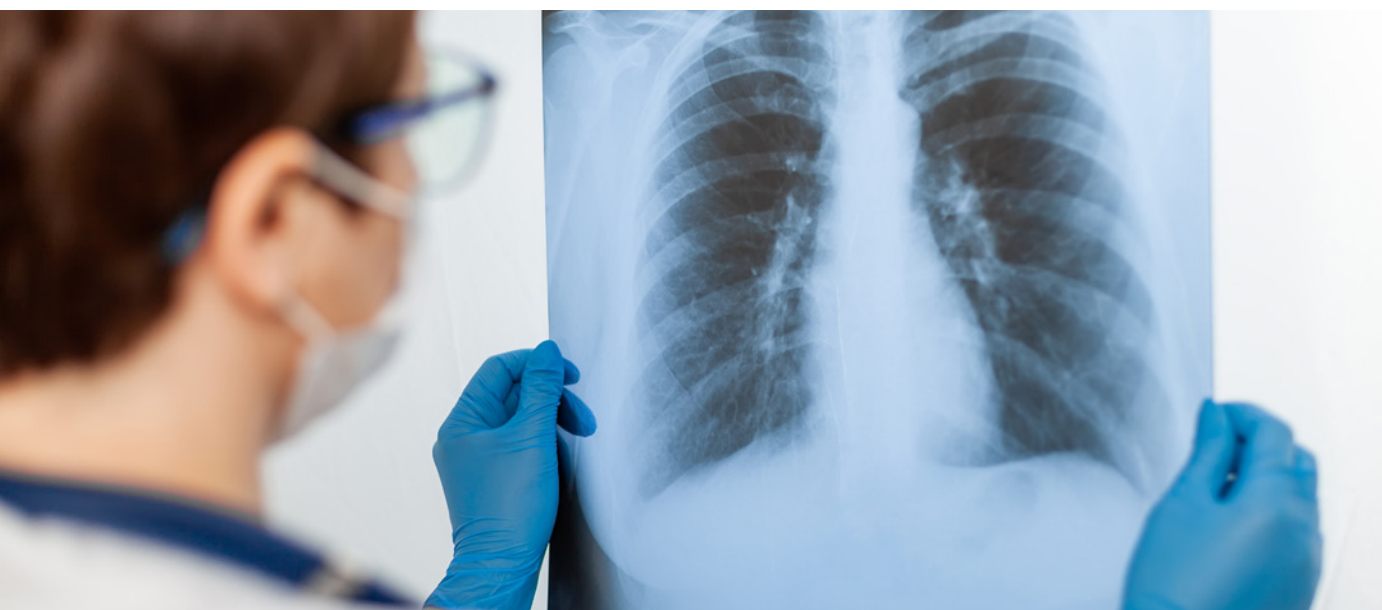
Every year, **tuberculosis** kills more than 1.5 million people. The Iqbal research group explores how to use genome sequencing to help clinicians determine appropriate treatment. The team developed a simple method that reduced the diagnostic time from two weeks to just over 12 hours. In a collaboration with Public Health England, the software developed by the group is now used when testing putative tuberculosis infections.

In 2019, the team received funding from UK Research and Innovation (UKRI) for the ARGENT project. The aim is to create a global,

live database of tuberculosis genomes, easily accessible by public health professionals. Within minutes a user can upload a tuberculosis sample from a patient and check if it is part of an outbreak cluster or a recent transmission from another country. This enables clinicians to decide which treatment is most suitable, and could help monitor drug resistance levels.

Another project from the Iqbal group combined knowledge of bacterial genetics with a web search algorithm, to design BIGSI, a search engine that allows scientists to query public microbial DNA data for specific genes and mutations. BIGSI could prove very useful for **monitoring antibiotic resistance** genes and understanding how bacteria and viruses evolve.

Funding from Health Data Research UK resulted in a collaboration to deliver the National Phenomics Resource, a pan-UK phenotype resource that will improve the interoperability of health data enabling research that will deliver improved patient care.



Agriculture

As part of the Functional Annotation of Animal Genomes (FAANG), EMBL-EBI has been involved in a coordinated international effort to accelerate genome to phenome research. We have been working closely with collaborators and animal breeders, particularly in the following areas:

- ⊙ Cattle breeding – [Bov-Reg](#)
- ⊙ Aquaculture – [AQUA-FAANG](#)
- ⊙ Pig and poultry breeding – [GENE-SWitCH](#)

In 2019, Ensembl released annotation and comparative genomics resources for 12 pig breeds, as well as a cattle hybrid. The Vertebrate Gene Nomenclature Committee (VGNC), which provides official gene names for key vertebrate species, now has an approved nomenclature for nearly 12 000 pig genes.



For the plant breeding community, EMBL-EBI's Industry Programme hosted a **hackathon** that brought together plant scientists, plant breeders, agronomists and industry partners in the agritech sector to develop tools for data integration. The aim was to identify sustainable

crop protection approaches to mitigate pest and disease issues that may arise as a consequence of climate change. The hackathon demonstrated EMBL-EBI's capability in supporting agritech beyond genomics, and the output of the event is published in GitHub.

Biodiversity

In recent years, EMBL-EBI has been increasingly exploring how bioinformatics can serve as a tool to understand and preserve the biodiversity of our planet.

EMBL-EBI is a partner in the ambitious [Darwin Tree of Life](#) (DToL) project, which aims to sequence the genetic code of 66 000 animal, plant, protozoa and fungi species found in the Atlantic archipelago of Britain and Ireland. The collaboration is part of a wider initiative called the Earth BioGenome Project, a global effort to sequence 1.5 million known species on Earth.

DToL will enhance our understanding of genomic diversity and evolution, will support ecology and environmental science, and will aid the protection and restoration of biodiversity.

EMBL-EBI's performs a number of important roles covering data archiving, coordination, annotation and delivery. The genomes generated through DToL will be annotated and made publicly available to scientists all over the world through Ensembl.

After the project received funding in 2019 to sequence the first 2000 species, Ensembl has been scaling up its infrastructure and annotation to accommodate the influx of data coming its way. The work builds on EMBL-EBI's involvement in the Vertebrate Genomes Project, an international consortium that ultimately aims to produce gold-standard genome assemblies for all known vertebrate species.

Ensembl is also trialling a new **rapid release protocol** that will speed up the publication of new genome annotation from months to just weeks. The new protocol will be delivered through a dedicated website with the ability to browse genomes, search for identifiers and perform sequence searches.



Name that gene



When researchers set out to study a novel gene, they try to give it an interesting name. The HUGO Gene Nomenclature Committee (HGNC) standardises and approves human gene names

and symbols. HGNC marked 40 years of activity in 2019. Tamsin Jones is a curator in the Vertebrate Gene Nomenclature Committee (VGNC), an offshoot from the HGNC that names non-human vertebrate genes.

Tamsin works to harmonise the names of large gene families across vertebrates. “We try and ensure every gene has a unique name that is consistent across the literature and across species,” she explains. Tamsin recently worked to create a unifying naming system for olfactory receptors. “It’s been a huge project with external collaborators, and we’re working with other nomenclature committees to adopt our system,” she says.

Tamsin studied evolution, genetics and developmental biology in New Zealand and in the United States. She then worked for FlyBase as a biocurator, sifting through *Drosophila* literature to catalogue the traits associated with different fly mutants. She joined the HGNC at EMBL-EBI to come closer to her passion for evolutionary biology.

Extending collaboration and coordination

EMBL-EBI's work has always been deeply rooted in collaboration. We work across borders and scientific disciplines. Here, we highlight some of our major collaborations in 2019, and how they are helping us achieve our second strategic priority: extending collaboration and coordination.

In 2019, we worked with 577 organisations from 62 countries. The diversity of our collaborators reflects the wide range of our projects, as well as the growing utility of bioinformatics.

UK Biobank

The European Genome-Phenome Archive (EGA) stores and shares identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA is a collaboration between EMBL-EBI and the Centre for Genomic Regulation in Barcelona.

Since 2017, the EGA has been working closely with the UK Biobank, which collects genomic and health information from over 500 000 UK volunteers. In 2019, the collaboration expanded and EGA partnered with deCODE Genetics and the Wellcome Sanger Institute to sequence, archive and distribute whole genome sequencing data from all UK Biobank participants. This new dataset, which will be made available between 2019 and 2021, is a treasure trove of data for scientists, and is expected to accelerate research into the prevention, diagnoses and treatment of a wide range of life-threatening diseases.

Human Cell Atlas

The Human Cell Atlas (HCA) is an international collaboration that aims to create comprehensive maps of all human cells, in order to help further our understanding of human health and disease. EMBL-EBI, the University of California Santa Cruz, the Broad Institute and the Chan Zuckerberg Initiative have developed the **Data Coordination Platform (DCP)** that enables researchers to organise and share HCA data.

The HCA DCP was released in April 2019, containing data from 29 projects. EMBL-EBI delivered **data ingestion services**, tools and processes allowing submitters to provide cellular-resolution sequencing data and associated metadata. A new programmatic submission route unifies **integrated submissions** with other EMBL-EBI resources, including BioStudies, BioSamples and ENA, with further archives to be integrated in 2020. This ensures data longevity and interconnectivity.

EMBL-EBI's Expression Atlas released an interactive analysis portal for datasets in the Single Cell Expression Atlas and the HCA DCP, facilitating interactive, flexible and scalable single cell RNA sequencing analysis workflows and data sets for the community.

Open Targets

Open Targets is a unique pre-competitive public-private partnership that uses human genetics and genomics data for systematic drug target identification and prioritisation. Founded by EMBL-EBI, the Wellcome Sanger Institute and GSK, the collaboration has grown to include Biogen, Bristol Myers Squibb, Sanofi and Takeda.

The Open Targets Platform enables researchers to **access potential drug targets** associated with disease. In 2019, the platform increased its functionality by enhancing the description of diseases, and by introducing target safety annotation and synthetic lethality data in cancer from the Open Targets experimental program.

Updates to the Open Targets Genetics portal – a processing pipeline and database for detailed analysis of human genetics evidence – included the addition of disease-molecular trait co-localisation analysis, integration of new molecular QTL datasets, including from the new eQTL Catalogue released in 2020, and integration of the GWAS Catalog summary statistics database.

Eight new Open Targets projects were initiated in 2019, including informatics projects aiming to enhance target safety assessment, define therapeutic windows for targets based on genetic data, and develop context-specific network methods.

Genomic data sharing

Within the next four years it is predicted that the majority of human genomes will be generated through national-scale healthcare initiatives. Vast and diverse datasets from millions of people are essential for genomics research, but there are many technical and ethical challenges for federated international data sharing. Facilitating safe and secure transcontinental human data exchange is essential for turning genomics into a useful tool for healthcare research.

In 2019, EMBL-EBI and partners launched a Horizon 2020 project called Common Infrastructure for Cohorts in Europe, Canada and Africa (CINECA), that will promote interoperability between 14 human cohorts consisting of over 1.4 million participants, to deploy Global Alliance for Genomics and Health (GA4GH) standards across international borders.

By enabling researchers to securely access genetic data from diverse human populations, CINECA will accelerate research in the genetic drivers of disease and support the development of treatments tailored to each individual patient's genetic profile, an ultimate goal of personalised medicine.

A collaborative model

The European Nucleotide Archive (ENA) **data submissions brokering** network has continued to mature. Data submissions brokering, in which our partners support the capture and curation of data into ENA from user communities, plays an increasingly important role in achieving breadth and quality in our data.

ChEMBL continued its collaboration with Novo Nordisk in the area of novel target discovery and initiated a similar project with the Structural Genomics Consortium.

UniProt developed new collaborations with groups working on disordered proteins and gene regulation, and continued its input into existing proteomics, protein interaction and functional annotation communities.

In a collaboration with the George Washington University and the University of Georgia, UniProt launched GlyGen, an online portal that hosts glycan data for several species.

The subscription-based Industry Programme welcomed BioMarin as a new member and delivered 12 workshops to its community.

ChEMBL and drug discovery



Nicolas Bosc is a scientist who specialises in data mining and analysis in the Chemistry Services team. Before he joined EMBL-EBI in 2016, Nicolas completed a PhD in France in cheminformatics, a field that applies computational tools to answer questions in

various subfields of chemistry, including pharmacology. “I had formal training in biology and chemistry, but by the end of my PhD, I was more of a data scientist,” he explains.

The Chemistry Services team develops and manages ChEMBL, EMBL-EBI’s database of small-molecule bioactivity data. Most ChEMBL users tackle problems related to drug discovery, and Nicolas helps them access and analyse data. He also uses the data in machine learning projects.

Like other members of the team, Nicolas contributes to external projects. “I joined a small consortium the aim of which was to provide a platform that would predict whether a molecule could help treat malaria,” he explains. “The Malaria Inhibitor Prediction Platform, or MAIP, is now freely available. It’s nice to see a project with such potential impact come to fruition.”

Global Alliance for Genomics and Health

EMBL-EBI continued to actively participate in the development of the Global Alliance for Genomics and Health (GA4GH) and the propagation of genomics data standards. In 2019, we led the Large Scale Genomics work stream, which creates standardised methods for accessing large-scale genomic data.

EMBL-EBI contributed to the development of the Variant Representation standard, which significantly reduces ambiguity in exchanging variation data. We also helped develop the RNAget API, which provides a means of retrieving data from several types of RNA.

A new GA4GH standard named htsget was implemented in the European Genome -phenome Archive (EGA), and has improved API access to genomic data. Finally, we helped develop the Cloud suite of standards, making data more accessible in a secure way. To aid in the secure distribution of genomic data files, the Crypt4GH container was developed with an EMBL-EBI centred use case. This allows archives such as EGA to share files in a safe and secure manner that permits encryption both at rest and in transit, with the access keys only known to the recipient.

Global Biodata Coalition

The Global Biodata Coalition (GBC) is a new international organisation supported by biomedical and life sciences research funders. It coordinates support for essential life sciences data resources, ensuring a sustainable future data ecosystem available to all researchers worldwide.

EMBL has been involved in starting up the GBC, and Rolf Apweiler continues to represent EMBL on the GBC steering committee and on the Board of funders. In 2019 the GBC established a secretariat and secured funding for development of full operations.

Supporting new initiatives

Our efforts to engage the global bioinformatics community amplified in 2019. We expanded our collaboration with the DNA Data Bank of Japan (DDBJ) to facilitate them reusing our database model to establish a new Japanese MetaboBank resource.

We also set up two collaborations with **NASA**, one to support them in establishing an in-house metabolomics resource, and the second to enable them to analyse microbiome samples from the International Space Station and other selected environments.

The remarkable diversity of the natural world can serve as an endless source of inspiration for science and industry. Accessing genetic resources from across the world helps us develop products and services that benefit humankind, from medicines to agricultural practices and beyond. **Access and Benefit Sharing** for biodiversity-related data is one topic on which we engaged with stakeholders in 2019, promoting and demonstrating the value of open data. Through our engagement around the Convention for Biological Diversity, we are helping the quest for models that enable open science and support fair and equitable Access and Benefit Sharing for providers of genetic resources and their derived data.

Engaging with funders, policy makers and regulators

EMBL-EBI works closely with strategic partners and funders from across the globe. In 2019, we had 18 funders providing support for 157 projects. A full list of funders is available on page 50.

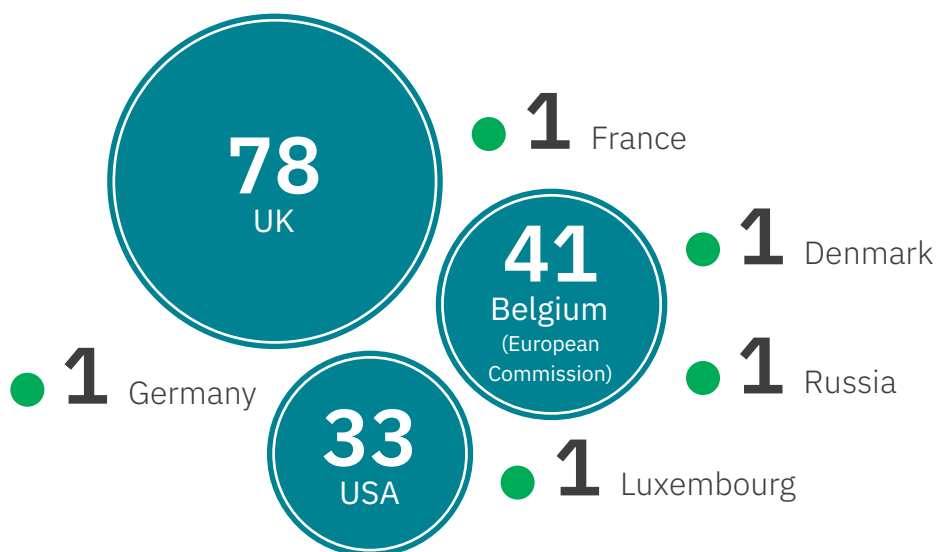
We maintain a helpful dialogue with policy makers in our host nation, Europe and the world. To this end we host inbound visits to facilitate bioinformatics knowledge exchange. In 2019, we welcomed the Luxembourg Institute of Health and we visited the French

National Institute of Health and Medical Research (INSERM), the only public research organisation in France entirely dedicated to human health.

As part of Tara Ocean Foundation's Mission Microplastics, a voyage around Europe to study the nature of plastic pollution entering the ocean, the London stop-over included a joint Tara and EMBL-EBI stakeholder event.

Globally funded

Number of grants received by EMBL-EBI in 2019 by funding body location



Continuous improvement, maximising efficiency

To support the advancement of bioinformatics and the growth of EMBL-EBI, we are striving to combine increasing scale with responsible and planned growth. This feeds into our third strategic priority: continuous improvement, maximising efficiency.

Reshaping our data resources

Several of our key data resources have seen significant improvements in 2019. Researchers submitting data can now use the newly launched [Data Submission Portal \(DSP\)](#) to send data to multiple EMBL-EBI archives. At launch, the service supported BioStudies, BioSamples and the European Nucleotide Archive (ENA). In 2019, users for the Human Cell Atlas and IMAGE consortium have used the DSP to successfully submit over 300 datasets into our archives. Additional archives will be integrated in 2020.

As we prepare to gradually retire the ArrayExpress database, we have migrated much of the infrastructure to the Biostudies database. BioStudies now accepts and distributes ArrayExpress functional genomics datasets. This required iterative refinement of the translation of the Array Express data format into the BioStudies model. The **BioStudies user interface** also enables clear and useful display of rich metadata. The team has started work on providing access to data in BioStudies from cloud environments, especially important for large imaging datasets that are too large for download.

With the aim of **disseminating proteomics data** to other EMBL-EBI resources, in 2019 the PRIDE and Expression Atlas teams worked together to reanalyse 25 quantitative proteomics datasets, integrating them into the Expression Atlas.

Developing our technology infrastructure

Our technical services teams focused heavily on improvements in **information security, storage capacity and green storage**. The appointment of an IT Security Officer helped us evaluate and improve existing security posture and protocols.

To facilitate more work on cloud infrastructures, our service teams can now deploy and manage containers as part of EMBL-EBI services and pipelines.

To help our users **submit an increasing amount of data**, the FIRE service continued to evolve. FIRE is an established software-defined storage system based on open source software, aiming to provide fast data archival and retrieval at scale, for the many physical, virtual and container platforms managed by EMBL-EBI. The yearly average for uploaded data grew from 15 TB per day in 2018 to 25

TB in 2019, reaching a record of 33 TB the month of October 2019. New methods of FIRE are now being developed to help speed up access and lower-data centre traffic.

The FIRE File Replication service has been in continual use since its creation in 2008. It has grown to be one of the EMBL-EBI flagship platform resources and, during 2019, its data ingestion and data serving capacity have increased in order to keep up with demand. FIRE is now also being used by EMPIAR for bioimage deposition and the EGA has increased its capacity as it stores new data arising from the UK Biobank.

File Transfer services have also undergone considerable consolidation and improvement, with the public data mounted on different end points standardised and redundancies removed.

In the pipeline

- ⦿ In 2020, EMBL-EBI will undergo a necessary data centre move from Hemel Hempstead to Harlow (UK), expanding its capacity by 150 racks.
- ⦿ The user is central to all EMBL-EBI Services. EMBL-EBI will continue User Experience staff training workshops covering aspects such as: conducting user interviews; designing a validating persona; running design workshops; data visualisation methods; gathering feedback; and remote user testing.

Solving the problems of the cloud



A Site Reliability Engineer, Anh Khoa Phan Le, or Dolphy as colleagues call him, recently joined the Software Development and Operations (SDO) Team.

The SDO Team develops, adapts and operates software to meet the internal needs of EMBL-EBI's computer cluster – a network of computers that act as a single, more powerful machine – and to support collaborations with external projects such as the Human Cell Atlas.

Dolphy comes from Vietnam and worked several years as a Senior Site Reliability Engineer for private companies in Singapore and Hungary. He joined EMBL-EBI to apply his skills to automate and improve internal and external transfer services.

He redesigned the application that manages EMBL-EBI's cloud user data, network, resources and access permissions. "I'm a problem solver, and this project was a great challenge for me," he says. He modernised the application with the latest technologies, making sure that sensitive user data was safely protected.

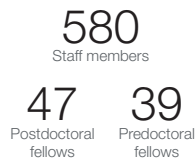
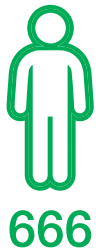
"I really enjoy working for EMBL-EBI because of the vast amount of data we deal with. I have a broad environment to explore and to learn with."

Building capacity and capability

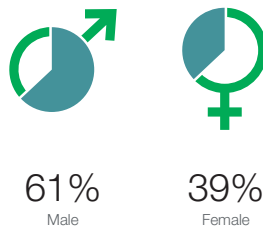
As a world-leading institute in bioinformatics, we play a major role in training and enabling life science professionals to access, analyse and share data to push the boundaries of our collective knowledge. We also recruit and develop a multidisciplinary work force that supports our fourth strategic priority: building capacity and capability.

Our people

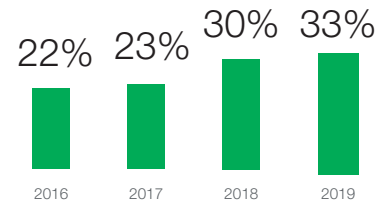
Staff in 2019 FTE (full-time equivalent)



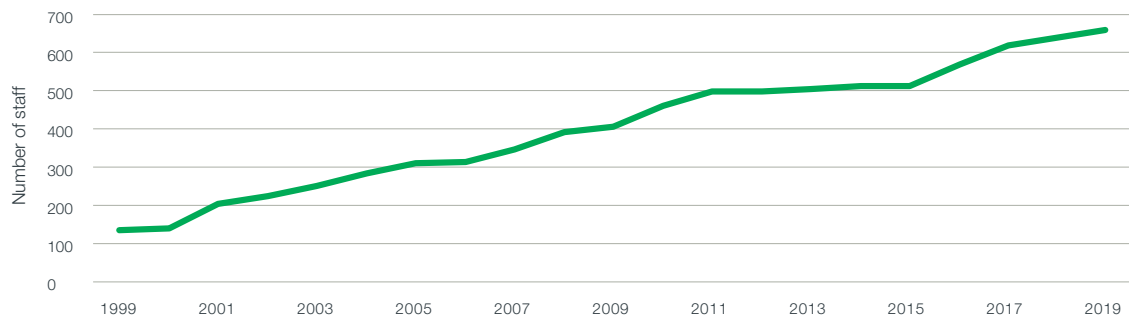
Gender distribution of staff in 2019



Senior roles held by women



Staff growth at EMBL-EBI, 1999 - 2019



In 2019 we welcomed **235 new starters**, including 101 staff members, 23 fellows, 43 trainees, 10 supernumeraries and 58 visitors. As EMBL-EBI continues to grow, we are seeing increased pressure on our office space capacity. Immediate temporary office accommodation on campus has been used and we are investing in modular office capacity for 100 members of staff until further expansion plans can be realised.

EMBL-EBI Administration coordinated preparations for Brexit and provided practical support to staff with information about the settled status scheme and immigration. EMBL-EBI supports **equality and diversity** and as such, recruits exceptional people from all over the world. Even though the majority of staff members are male, we are seeing steady progress in relation to gender balance, especially in leadership positions.

One area of recruitment that has proven challenging is the **technical careers** and software engineering roles. In 2019, we took advantage of our location in the Cambridge life sciences cluster and participated in the Agile Cambridge event. We also organised DevDay, EMBL-EBI's first dedicated recruitment event for software developers and engineers. This was an excellent opportunity for knowledge exchange, and it positioned EMBL-EBI as an employer of software engineers in the local area.



New leadership

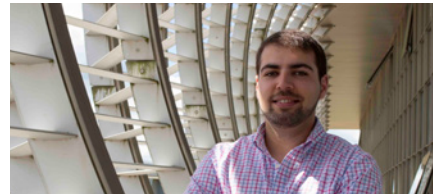
Paul Flicek and **Jo McEntyre** were appointed Associate Directors of EMBL-EBI Services. They are responsible for the institute's extensive suite of data resources and related activities, with the appointment set to consolidate EMBL-EBI's role as an international infrastructure for the life sciences.



Tony Burdett was appointed to lead the newly created Archival Infrastructure and Technology (AIT) team, which develops services and technology for EMBL-EBI's molecular archives, including for data submission, storage and coordination. AIT is also responsible for data ingestion into the Human Cell Atlas Data Coordination Platform.



Isidro Cortes-Ciriano joined EMBL-EBI as a new Research Group Leader, focusing on Cancer Genomics. His group develops computational tools to characterise the patterns of mutations and genome instability processes in human cancers, by analysing sequencing data from clinical samples and preclinical models.



Ian Dunham was appointed as Open Targets Director. He hopes to bring new technologies and approaches to the public-private partnership, including increased use of single-cell sequencing, CRISPR and artificial intelligence.



Jessica Vamathevan became Head of the Strategic Partnership Office. In her new role, she is responsible for strategy development, planning and industry initiatives to drive broader academic and industry engagement, including management of the EMBL-EBI Industry Programme.



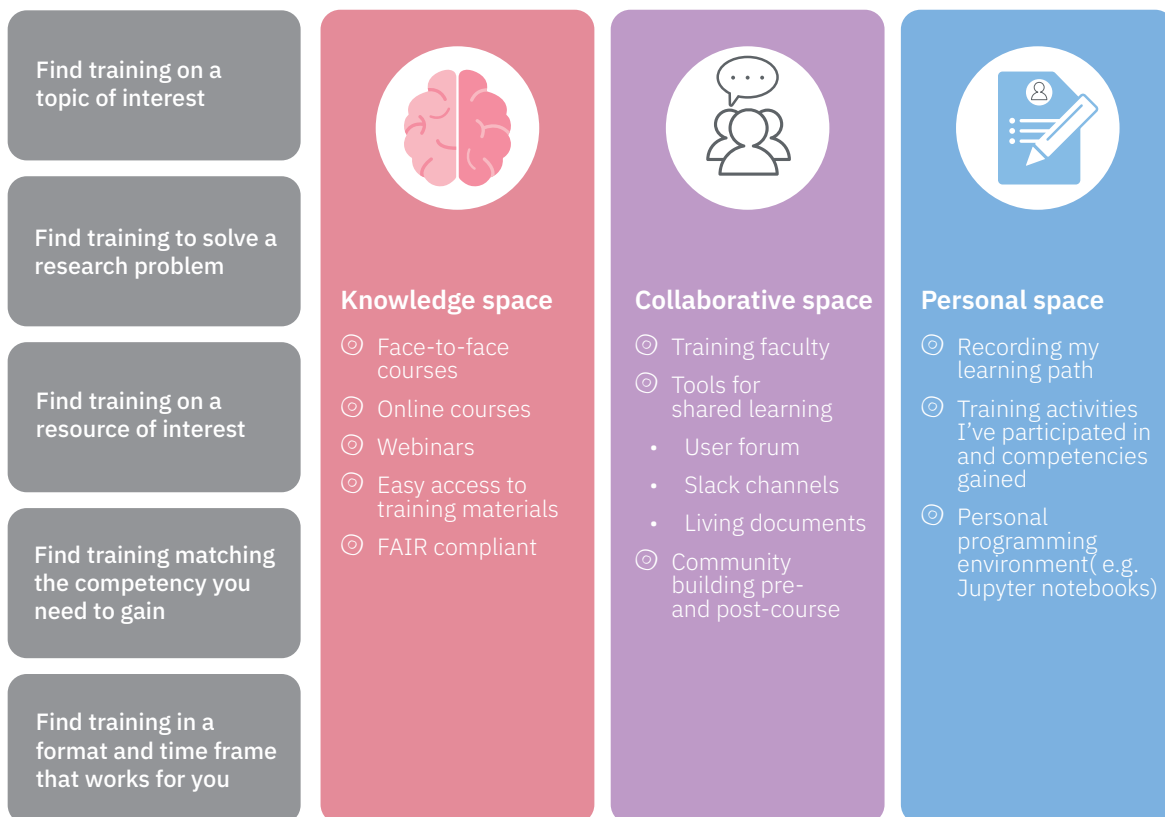
Providing bioinformatics training

The mission of the EMBL-EBI Training Programme is to deliver **world-leading training in bioinformatics** and scientific service provision to the research community. The aim is to empower scientists at all career stages to make the most of biological data, and to strengthen bioinformatics capacity across the globe.

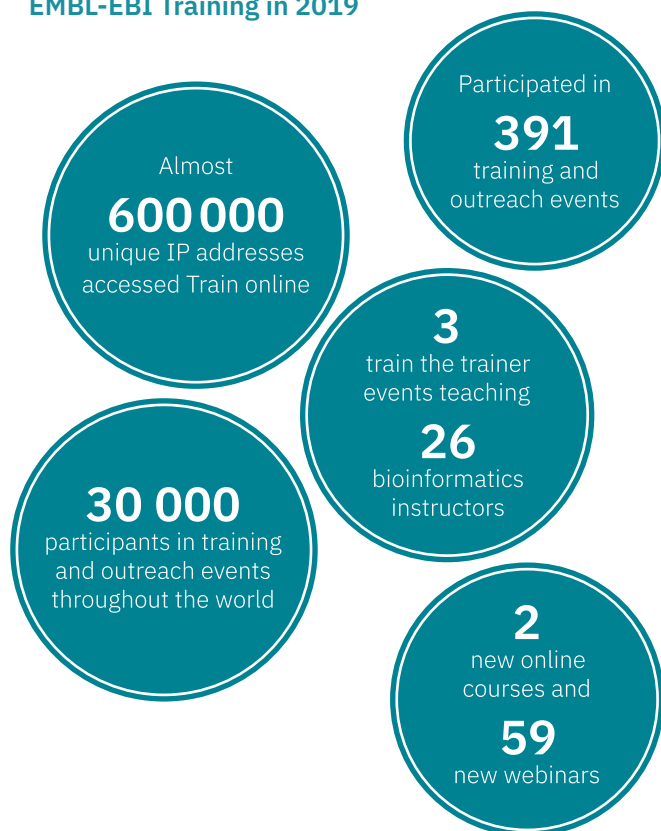
The team forms part of EICAT – the EMBL International Centre for Advanced Training – and works closely with colleagues at other EMBL sites. The team coordinates external training activities at EMBL-EBI to create a coherent, high-quality programme with global reach.

In 2019, the Training and Web Development teams launched a project to reimagine the **EMBL-EBI Training website**. This includes the development of new user-friendly course authoring tools and a new format for online courses.

EMBL-EBI leads the **CABANA project**, funded by the Global Challenges Research Fund (GCRF). CABANA aims to address the slow implementation of data-driven biology in Latin America by creating a sustainable capacity-building programme.



EMBL-EBI Training in 2019



In 2019, the CABANA project hosted 11 secondees from Latin America and concluded a very competitive second call for secondees. The team also delivered 11 workshops and supported a further three throughout Latin America. Train-the-trainer activities took place in Brazil, Peru and the UK. To support online learning, CABANA also developed a new eLearning portal and delivered a series of interactive webinar-based tutorials.

EMBL-EBI contributes training and outreach expertise to an ever-shifting portfolio of bioinformatics projects. In 2019, the Expression Atlas team delivered its first course on analysis of single-cell RNA sequencing data and Ensembl continued its training programme in low to middle income countries (LMIC). Since the start of the program in 2018, Ensembl has taught over 3000 participants in 94 training events in South America, Africa, Asia and the Middle East.

Trained and training others to train



Melissa Burke started her career as a researcher. After a PhD in parasitology in Australia and a postdoc in the UK, she joined EMBL-EBI in 2015. “I’d been on a course

at EMBL-EBI and heard about the curators who work behind the scenes. I joined EMBL-EBI first as a curator and then as a scientific training officer.”

In the Training Team, Melissa looks after all of EMBL-EBI’s webinars and online tutorials. She was instrumental in the redesign of EMBL-EBI’s online training catalogue and helps improve the experience of both trainers and trainees. She regularly coaches EMBL-EBI resource teams on how to develop their own training content.

EMBL-EBI’s new and improved online training platform is now in the works. “Online learning is really important for anyone who can’t attend a face-to-face course,” she says. “We get lots of positive feedback, saying how helpful our materials are, because it allows people to learn at their own pace.”

Engaging with the public

As bioinformatics, data analysis and digital science enter the mainstream, EMBL-EBI is growing its suite of activities for **engaging with non-scientific audiences**. The aim is to highlight the importance of our work in understanding human health, agriculture, biodiversity and our world in general.

In 2019, EMBL-EBI colleagues engaged with primary and secondary schools, university students, teachers, and community groups. As part of European Researchers' Night, the consortium in which EMBL-EBI was involved – called LifeLab – supported local events in three cities, which attracted 3400 attendees.



UNICEF, EMBL-EBI and partners encoded the Convention on the Rights of the Child in synthetic DNA.

Several new public engagement activities were developed to showcase the work of the institute. They were unveiled during EMBL-EBI's 25th anniversary celebrations and the most popular ones were repurposed for further activities, visits and events.

In a novel collaboration, EMBL-EBI teamed up with UNICEF and scientific partners Twist Bioscience and Imagene to make the 30th anniversary of the Convention on the Rights of the Child. As part of a campaign to highlight the importance of the document, they stored the text of the convention in synthetic DNA, which was deposited in the Global World Archive, in Svalbard, for safe keeping.

Public engagement in 2019



46 visits to campus
spanning **2400** people



50 activities off-campus
reaching **11 600** people

Supporting global expansion of biomolecular resources

As a world-leading centre for bioinformatics, we sit at the heart of a complex global network and work with collaborators across Europe and the world to advance science and technology. This feeds into our fifth and final priority: supporting the global expansion of biomolecular resources.

ELIXIR

EMBL-EBI serves as the European Node of ELIXIR, an international consortium with 22 member countries. ELIXIR brings together over 600 experts from more than 200 research institutes. ELIXIR connects the major centres for bioinformatics in Europe offering a collaborative platform for a united European operation. EMBL-EBI's continuous support since the inception of ELIXIR has resulted in a firm footing in the bioinformatics community and provided a formal mechanism for collaboration in data provision and setting data standards.

The Horizon 2020-funded ELIXIR-EXCELERATE project successfully concluded in August 2019. The project transformed ad-hoc links between European bioinformatics institutes into a systematic and long-term collaboration.

The implementation of **the second Scientific Programme** began in June 2019 with the launch of 28 new technical and scientific projects. These will extend the portfolio of existing ELIXIR services and develop new activities to address existing bottlenecks in working with life science data, such as responsible sharing of sensitive human data, research reproducibility, and reusability of bioinformatics workflows.

In 2019, ELIXIR expanded the portfolio of ELIXIR Communities, setting up the following ones:

- ⊙ Structural Bioinformatics
- ⊙ Intrinsically Disordered Proteins
- ⊙ Microbial Biotechnology
- ⊙ Human Copy Number Variation Community

EMBL-EBI is involved in all of the new Communities and is co-leading the Structural Bioinformatics Community.



The five ELIXIR Platforms: Data, Tools, Compute, Interoperability and Training are complemented by ELIXIR Communities.
Credit: ELIXIR

In November 2019, ELIXIR brought together over 150 bioinformaticians, trainers and software developers in Paris, for the second **BioHackathon-Europe**. The participants collaborated on over 30 projects aimed at improving software and methods for life science resources.

ELIXIR 2019 all-hands meeting.

ELIXIR also continued its collaboration with GA4GH, setting up a new strategic partnership to develop technical standards and regulatory framework for responsible sharing of genomic data across national borders. With healthcare providers soon to become the main producers and consumers of genomic data, research service operators face novel challenges in facilitating cross-border collaboration, data exchange and re-use.



Key information



Financial figures

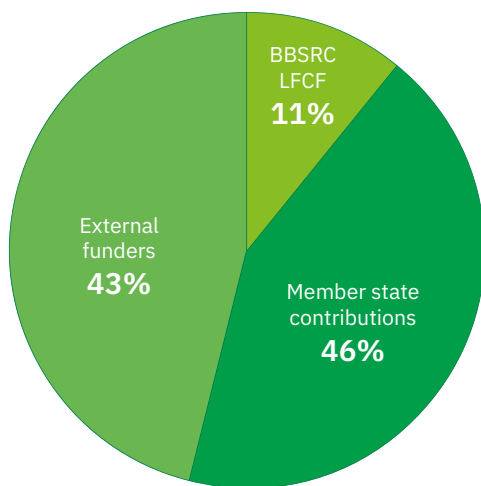
We are grateful for the continued support of our member states and other funding bodies, which in 2019 helped us develop our data resources, perform research, and respond to the requirements of the international scientific community.

The work at EMBL EBI is funded by its member state contributions and also other external funders (for example through grants). In addition to these sources, during the 2019 financial year, we received further capital investment of €9.6m from the UK Research and Innovation’s Biotechnology and Biological Sciences Research Council (UKRI-BBRSC), as part of the Large Facilities Capital Fund (LFCF).

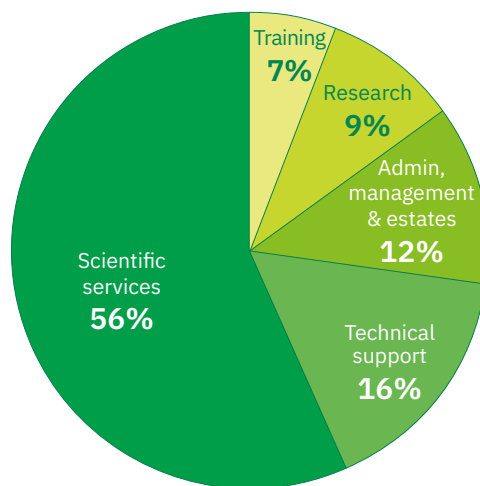
In 2019, EMBL-EBI was also awarded £44.5 million from the UK Research and Innovation’s Strategic Priorities Fund (SPF) to enable technical infrastructure development. The first phase of the expansion runs from February 2019 until March 2024 and will enable EMBL-EBI data resources to grow in response to the need for biological data management and analysis. The funding will support data

hosting for large-scale science programmes, including UK Biobank’s whole sequence genome analysis, the Human Cell Atlas and the establishment of the BioImage Archive, a unique biological imaging repository. This award builds on investment provided through UK Governments Large Facilities Capital Fund, which was completed in 2019.

The total operating expenditure of EMBL-EBI in 2019 was €85m, offset by miscellaneous income of €-4.8m. The expenditure reflects the increases in activity required to maintain EMBL-EBI resources at a time when data depositions and usage in the field of molecular biology continues to grow globally.

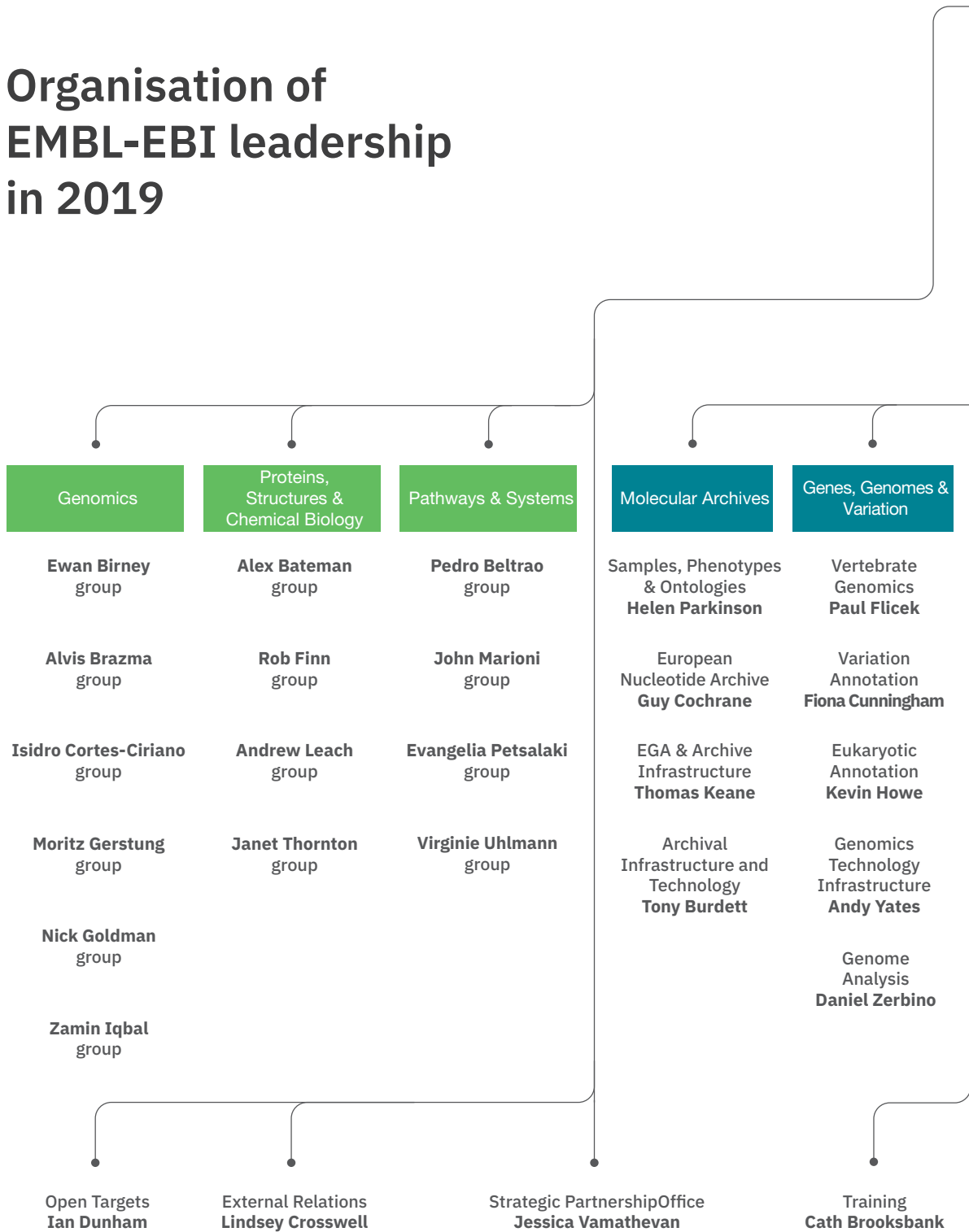


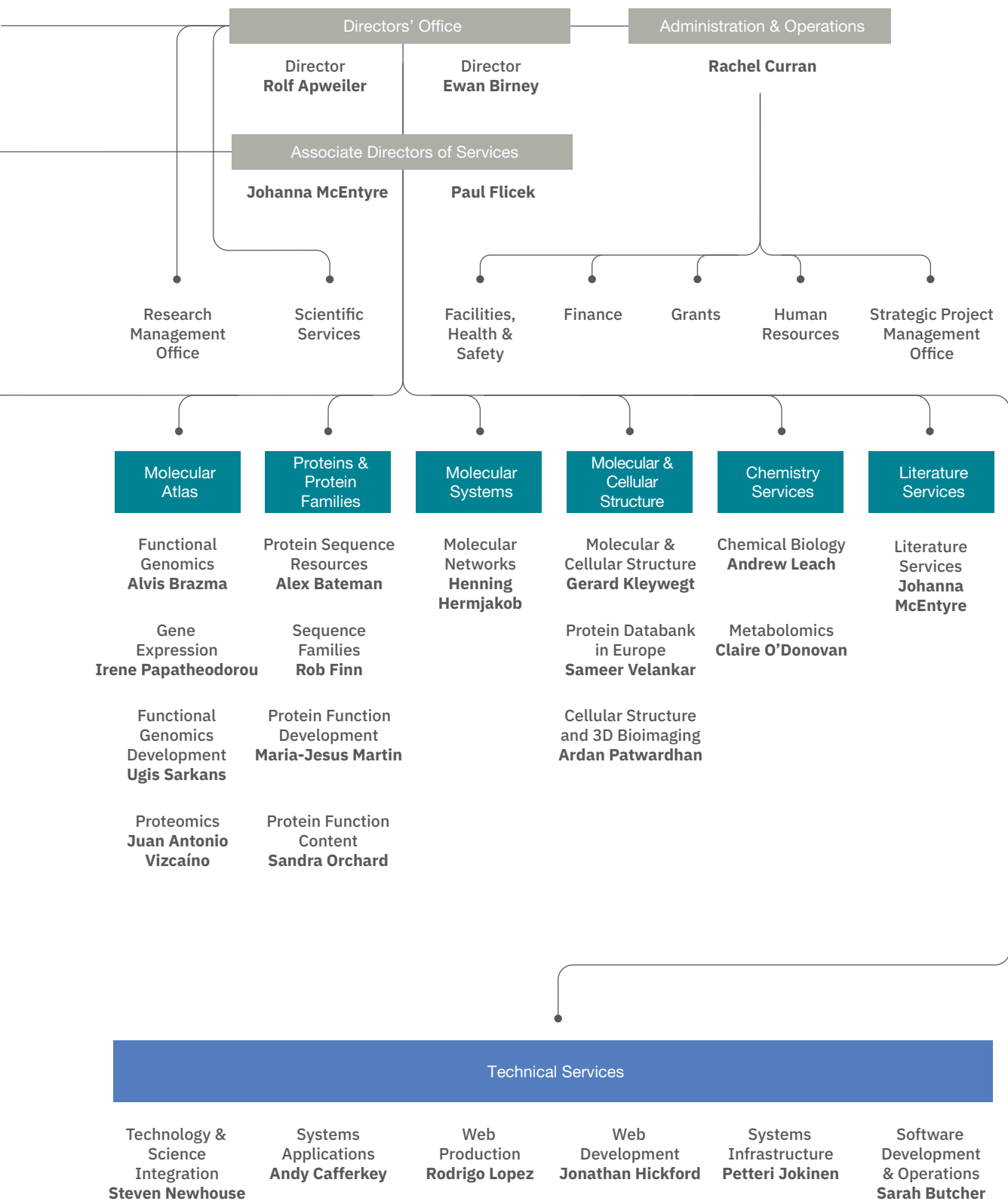
Source of EMBL-EBI funding



Use of EMBL-EBI funds

Organisation of EMBL-EBI leadership in 2019





Our governance

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

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

In 2019, EMBL-EBI was led by joint Directors Rolf Apweiler and Ewan Birney, supported by two joint Associate Directors for EMBL-EBI Services, Paul Flicek and Johanna McEntyre, as well as the Strategy and Management Committee (SMC) and 46 Group and Team Leaders (GTLs).

Strategy and Management Committee (SMC)

SMC deliberates on strategic decisions including budget, funding and the creation of new posts. Members in 2019 were (shown in alphabetical order):

Rolf Apweiler - Director

Alex Bateman - Head of Protein & Protein Family Services

Ewan Birney - Director

Alvis Brazma - Head of Molecular Atlas Services

Cath Brooksbank - Head of Training

Lindsey Crosswell - Head of External Relations

Rachel Curran - Head of Administration

Paul Flicek - Associate Director of EMBL-EBI Services, Head of Genes, Genomes & Variation Services

Nick Goldman - Research Group Leader and Head of Research

Henning Hermjakob - Head of Molecular Systems Services

Gerard Kleywegt - Head of Molecular and Cellular Structure Services

Andrew Leach - Head of Chemistry Services

Johanna McEntyre - Associate Director of EMBL-EBI Services, Head of Literature Services

Steven Newhouse - Head of Technical Services

Helen Parkinson - Head of Molecular Archive Resources

Our funders

We are extremely grateful to all of our funders for their crucial support to our work in 2019.



EMBL member states, associate member states and prospect member states:

Argentina (associate), Australia (associate), Austria, Belgium, Croatia, Czech Republic, Denmark, Estonia (prospect), Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia (prospect), Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland and United Kingdom.

Alzheimer's Research UK

Biotechnology and Biological Sciences Research Council

The Company of Biologists

Cancer Research UK

European Commission

European Molecular Biology Organization

Foundation for the National Institutes of Health

Fonds National de la Recherche Luxembourg

Human Frontier Science Program

The Moore Foundation

Medical Research Council

National Institutes of Health

Novo Nordisk

National Science Foundation

Research Councils UK

Russian Foundation for Basic Research

The Genetics Society

Wellcome Trust




List of acronyms




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

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

**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
 /EMBLEBI

**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:

Argentina, Australia, 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

Prospect member states: Estonia, Latvia