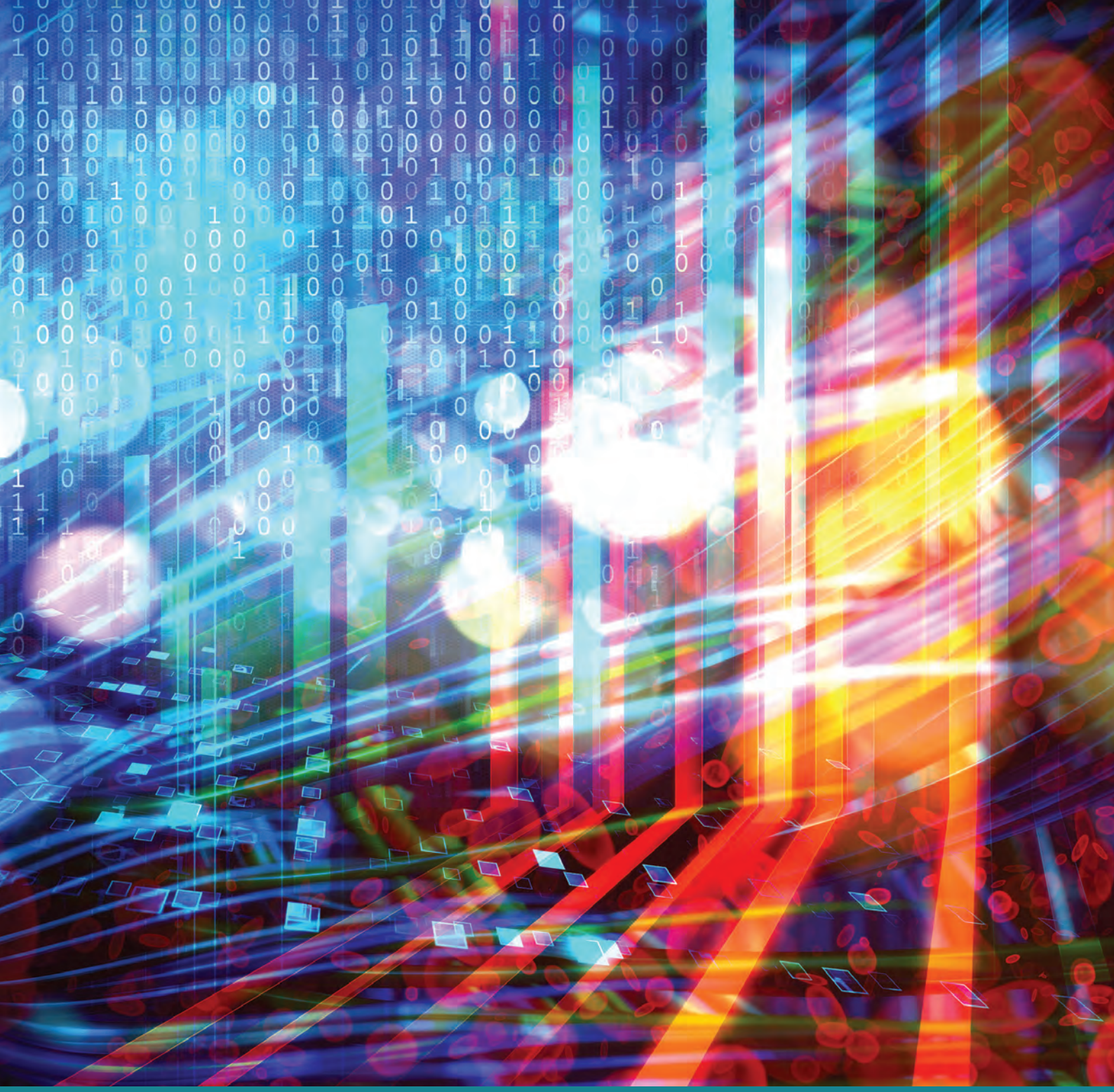


European Bioinformatics Institute · Cambridge

Annual Report 2018

EMBL-EBI

© 2018 European Molecular Biology Laboratory

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

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

Table of contents

Who we are

EMBL's European Bioinformatics Institute (EMBL-EBI) is the world's leading source of 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

As EMBL-EBI prepares to celebrate its 25th anniversary in September 2019, we look back on the things that have transformed our institute into a leading source of biological and biomolecular data for scientists all over the world, and we reflect on the strategic priorities set to guide our activities over the coming years.

Bioinformatics is an excellent spring board for discoveries in the life sciences but serving a growing user base with changing demands requires both structured planning and flexibility. This report details the progress we made in 2018 against our strategic roadmap and the many ways in which we support scientific discovery and drive societal applications in collaboration with partners, funders and the wider community.

As the life sciences continue to transform with the advent of ever cheaper and more portable technologies that can generate immense volumes of data, coupled with computational advancement, the way we currently perform bioinformatics is also radically changing.

2018 saw EMBL-EBI grow to host over 40 different biomolecular databases with data volumes continuing to rise. Data access and demand increased significantly in 2018, with the average number of daily requests to our websites almost doubling to 64 million.

In 2018, we launched the first dedicated resource for single-cell RNA sequencing data for different species, the Single Cell Expression Atlas. We also laid the foundation of the BioImage Archive, an open data resource for reference bioimages.

The growth in data submission and reuse, as well as the global community's push for repositories for new data types, reflect the shift of bioinformatics from a niche discipline to a more central space in the life sciences. From human health to agriculture and biodiversity, more researchers are turning to bioinformatics to find answers to pressing questions, and high-quality data is of the essence.

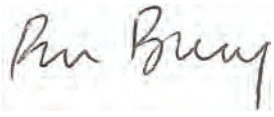
To address these changing needs, EMBL-EBI continues to play diverse roles in data collaborations, to adapt our training offerings to build capacity, and to support the global expansion of biomolecular data resources.

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



Celebrating 25 years of EMBL-EBI

Cast your mind back to 1994. It was the year Brazil won the World Cup after a penalty shoot-out with Italy, and the hit TV series ***Friends*** debuted on NBC. In the world of technology, Amazon and Yahoo had just been set up and the first commercial web browser, Netscape Navigator, was launched. The internet, previously used mostly by scientists and scholars, was beginning to look like the next big thing.

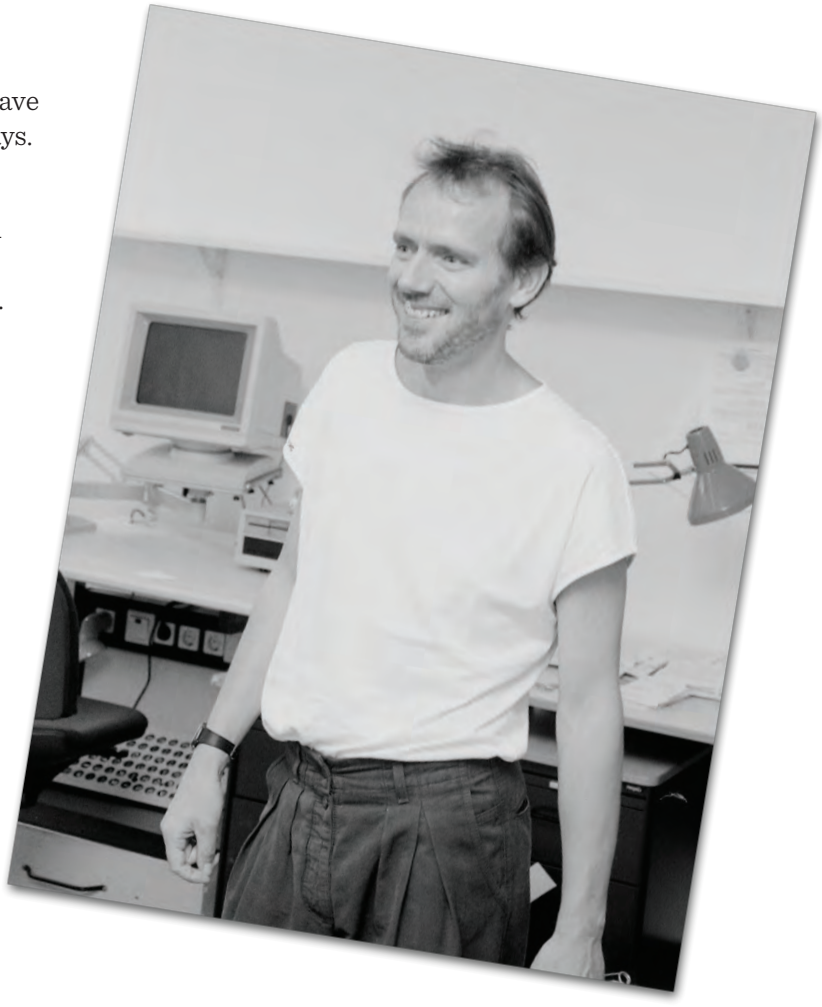
In September 1994, a small group of researchers from EMBL Heidelberg travelled to a remote campus in the Cambridgeshire countryside to set up a home for the growing volumes of biological data being generated around the world: the European Bioinformatics Institute (EMBL-EBI) was born.

Here, we present the memories and reflections of some of the people who have been with EMBL-EBI since its early days.

"The idea for EMBL-EBI was born in the mind of Graham Cameron, who ran the EMBL Data Library," explains Rolf Apweiler, Joint Director of EMBL-EBI. "He believed sequencing would be transformative for biology, but only as long as there was a place that would archive, analyse and annotate the sequences, and, most importantly, make them publicly available."

"The idea for EMBL-EBI was born in the mind of Graham Cameron, who ran the EMBL Data Library,"

Rolf Apweiler,
Joint Director of EMBL-EBI



Graham Cameron (pictured here) developed the concept for EMBL's European Bioinformatics Institute. Credit: EMBL Archive

EMBL-EBI Main Building under construction

Humble beginnings

Today, EMBL-EBI's two buildings accommodate around 800 employees from over 60 countries, but back in 1994, things were very different.

"EMBL-EBI was a couple of Portakabins and a hole in the ground that Graham Cameron very proudly gave us a tour of," remembers Claire O'Donovan, Head of Metabolomics. "What's funny is that today we still have Portakabins on site, but only because the institute is growing so fast that we often have to use them for staff overspill."

And it's not just the number of people that is on the rise. Maria Martin, who joined EMBL-EBI as a database developer in 1996 and now runs the Protein Function Development team, reflects on how much the data volumes have changed. "Back then, Swiss-Prot – today part of UniProt – had about 80 000 entries. We thought this was a lot and were wondering how to handle the amount of data that was coming in from collaborators. Nowadays we have over 150 million protein sequences, and growing."

> **"We were wondering how to handle the amount of data that was coming in,"**
>
> Maria Martin,
> Team Leader,
> Protein Function Development

All about the data

In 1994, the two data resources for EMBL-EBI were the EMBL Nucleotide Sequence Database – now the European Nucleotide Archive (ENA) – and Swiss-Prot. Alongside these, there was also a small research group and a huge sense of excitement about what was to come.

Over time, the volume and diversity of data increased significantly. "In the late 90s, the microarray revolution started in Stanford," explains Alvis Brazma, Head of Molecular Atlas Services. "I remember that industry was particularly interested in the topic. In fact, ArrayExpress was one of the first data resources set up with industry contributions."

These days, genomics, single-cell sequencing, metagenomics and imaging data are just some of the many data types EMBL-EBI resources accommodate. "We have always been very good at pre-empting the next big thing and adapting to it," says Alvis.

A computing revolution

As data volumes grew, so did the demand for infrastructure. The first computer room consisted of only a few racks. When the time came to expand the computing room, it sparked a big debate.

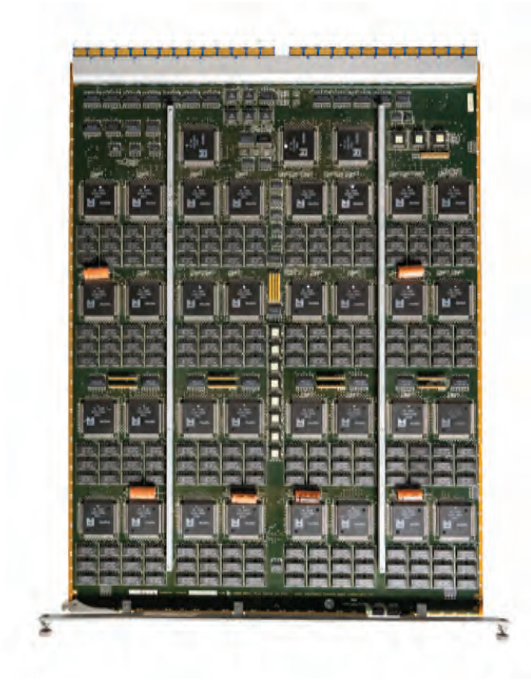
Mark Green, former Head of Administration, remembers: "The table tennis room was quite a large social area, so we thought it would be a terrible waste to convert it into a computer room, as it would take us years to fill it with kit. In the end, we bit the bullet and converted it. Within 18 months, the place was rammed full of kit and we were running out of space yet again. Soon after, we set up a data centre on campus. Now, we have three data centres plus cloud storage, which is constantly on the rise."

Another technical milestone was setting up the first web servers for EMBL-EBI data resources, in the late 90s, the early days of the internet. "There were lots of problems with connectivity back then, so getting data from the United States required special traffic permissions," recalls Rodrigo Lopez, Head of Web Production.

"From the beginning, the internet was all about search," continues Rodrigo. "And all of a sudden, you didn't have to go to the library, sit in a queue or wait for books. You simply sat at your desk and connected to the network. It was a huge shift in how science worked.

"We used to have these crazy coding competitions back then, to see who wrote more code, and we would count lines and mistakes to determine the winner. We had a hell of a good time; we were writing code, we were developing methods, we were doing science. It was all cutting edge and there was an amazing atmosphere. Even today I think we're just scratching the surface of what we can do with the web."

> "Even today I think we're just scratching the surface of what we can do with the web,"
>
> Rodrigo Lopez,
> Head of Web Production



Silicon Graphics hardware facilitated a significant increase in the number of sequence similarity searches that could be done in a given timeframe. Credit: EMBL-EBI

Hole-in-the-Wall gang

So what about the people who made all these things happen? "Before I joined EMBL in Heidelberg, I had been told that EMBL was a bit like the Hole-in-the-Wall gang – they didn't follow rules, they made their own. And I discovered that EMBL-EBI was a bit like the Hole-in-the-Wall gang's Hole-in-the-Wall gang," says Mark Green.

"It felt more like a group of friends working together, a young institute where everybody was a colleague and we had regular international cuisine parties," remembers Maria Martin.

As the institute grew, it became impossible to know everybody, but to this day, teams work together closely through "glue projects", ensuring that data are interoperable. Knowledge exchange and collaboration within and outside EMBL-EBI are pillars of open data for the life sciences.

Looking to the future

Much has changed in 25 years, but some things remain the same. EMBL-EBI is still collecting, analysing and opening up data for its users. It just happens at a much greater, more diverse scale. Also, while in the past, data were used mainly by bioinformaticians, they now power discoveries in human health and disease, precision medicine, agri-tech, biodiversity and beyond.

So, what's next? "The big unknown now is the functional part," says Rolf Apweiler. "We know only a small number of the functions of genes, transcripts and proteins, but we need to work out their full characterisation. Sequencing only scratches the surface. The functional question is a much bigger one and will take a very, very long time to answer. But when we crack it, it may allow us to do things we can now only dream of."

"We have always been good at pre-empting the next big thing and adapting to it,"

Alvis Brazma,
Head of Molecular Atlas Services

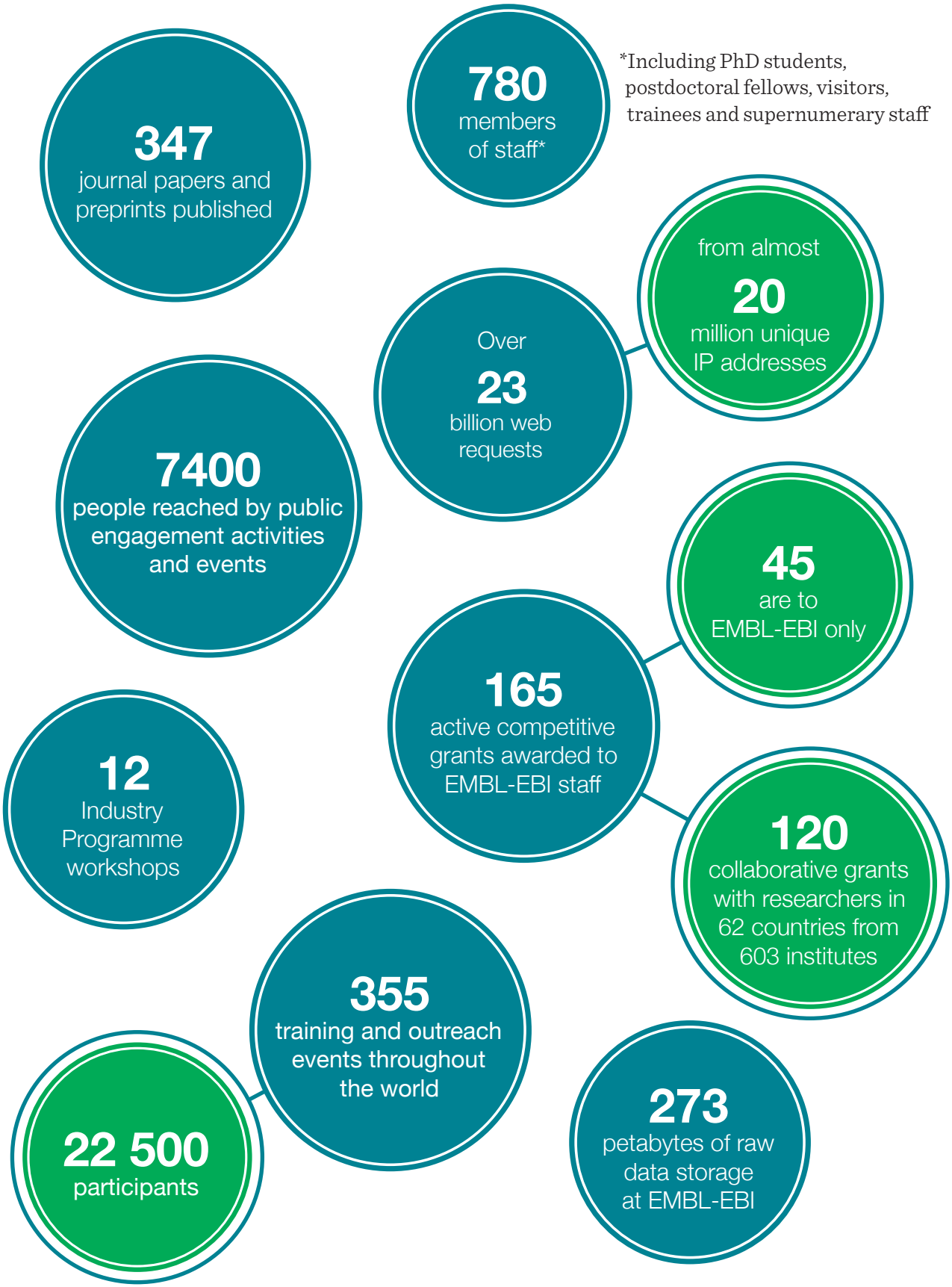
EMBL-EBI's first router, which allowed the institute to launch its first web servers. Credit: EMBL-EBI

What we achieved in 2018

Our resources are used across the globe and this only increased in 2018: we now have users in virtually every country on Earth.

To fulfil our missions, in 2018 we launched new data resources and improved existing ones, grew our research activities, expanded our training activities on other continents and multiplied our collaborations.

2018 in numbers



Highlights of the year

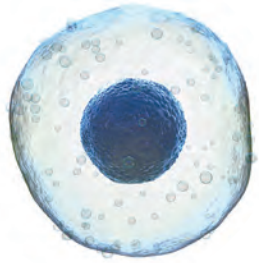
EMBL researchers design a computational method (MOFA) to jointly analyse multiple types of molecular data (page 26)

JUN

Celgene joins Open Targets (page 31)

MAY

Single Cell Expression Atlas launched (page 20)



APR

Scientists use data from human cancers and *C. elegans* to understand mutational causes of cancer (page 25)

PDX Finder is launched (page 20)

MAR

In a first for EBI Training, two students use robot avatars to study genomic medicine (page 38)

Study shows that similar characteristics in sheep and goats can result from different patterns of gene selection (page 26)

Three years after EMBL-EBI's Nick Goldman issued his 'Bitcoin Challenge' in Davos, University of Antwerp's Sander Wuyts decodes the key for DNA storage.

JAN

New technique developed to detect differences between cells (page 26)

FEB

Study identifies new diabetes genes in mice (page 25)

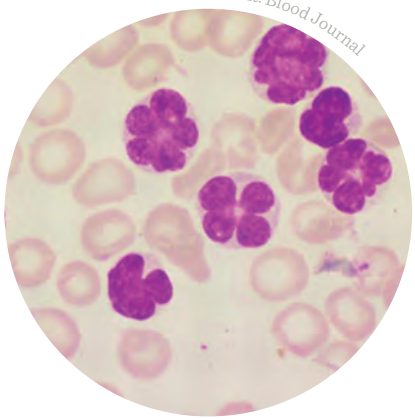
Cambridge is selected as a Health Data Research UK site, which will be jointly run by EMBL-EBI, the Wellcome Sanger Institute and the University of Cambridge.

Data resources news

Research news

EBI Metagenomics becomes MGnify (page 20)

Study shows that virus HTLV-1 changes DNA to affect thousands of genes (page 26)


Credit: Blood Journal

EMBL-EBI's BioStudies database and SourceData from EMBO announce a collaboration that is set to make published data from EMBO Press openly accessible and easy to find.

DEC

NOV

Researchers find that roots of leukaemia are detectable in blood years before diagnosis (page 25)

Europe PMC begins indexing of preprints (page 29)

PDBe celebrates its schools art project with its largest exhibition yet (page 39)

JUL

EMBL-EBI joins the Darwin Tree of Life project, the UK contribution to the global effort to sequence 1.5 million known eukaryotic species on Earth (page 28)

The Wellcome Genome Campus, home of EMBL-EBI, awarded the Silver Engage Watermark for public engagement.

OCT

Newly sequenced mouse genomes unearth unknown genes (page 27)

Sanofi joins Open Targets (page 31)

First large-scale systematic analysis explores how germline variants effect drug response of cancer cells (page 25)

AUG

SEP

Scientists find a faster, more accurate way of diagnosing and treating tuberculosis using DNA sequencing (page 26)

PhenoMeNal is launched (page 20)

EMBL-EBI joins LifeLab, a series of events to engage the public with science for European Researchers' Night (page 39)

Progress against our strategy

To deliver on our missions, we have developed a strategic roadmap that enables us to provide for the growing volumes of biomolecular data, the diversity of biological applications and the increasing needs of a global scientific community. In 2018, we made significant progress against our five strategic priorities.

Increasing usage, utility and application of bioinformatics

As bioinformatics becomes a more mainstream discipline, demands for data services are increasing significantly. One of our strategic priorities is to serve the needs of an extremely broad, diverse user base by delivering the best possible quality of data resources and research that complement our services.

Data usage and data submission

2018 was another year of intense growth in terms of data usage, reflecting our efforts in improving the experience of our users. On an average day at the end of the year, we saw more than 64 million requests to our websites. This is a considerable increase compared with the average daily web requests in 2017 (just under 38 million) and 2016 (27 million). We are seeing a clear upward trend over time as the number of users increases and the mean usage by each user also grows.

EMBL-EBI websites were accessed from almost 20 million unique IP addresses in 2018 from virtually everywhere in the world, with the heaviest usage coming from the USA (22%), China (13%), India (9%), the United Kingdom (6%) and Germany (3%).

To manage increasing data submissions, our computational infrastructure is continually expanding. At the end of 2018, we had around 273 petabytes (PB) of raw storage available, up from 260 PB in 2017.

What is a request?

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

How big is a petabyte?

A petabyte (PB) is 10^{15} bytes of data, 1000 terabytes (TB) or 1 million gigabytes (GB). Assuming that one standard definition film uses about 4 GB of storage space, a petabyte is big enough to hold 250 000 films.

EMBL-EBI's raw data storage capacity is 273 PB. If that amount of data were stored on DVDs, the resulting stack would be around 87 km high, nearly **ten times the height of Mount Everest**, the tallest point on Earth.

Data resources update

EMBL-EBI's services include the provision of biological databases and the tools to explore them. Our user community includes academic and commercial researchers throughout Europe and the world, with ever-increasing usage by medical professionals and the healthcare industry.

Launching new data resources

In 2018, we released new data resources to life scientists:

PhenoMeNal

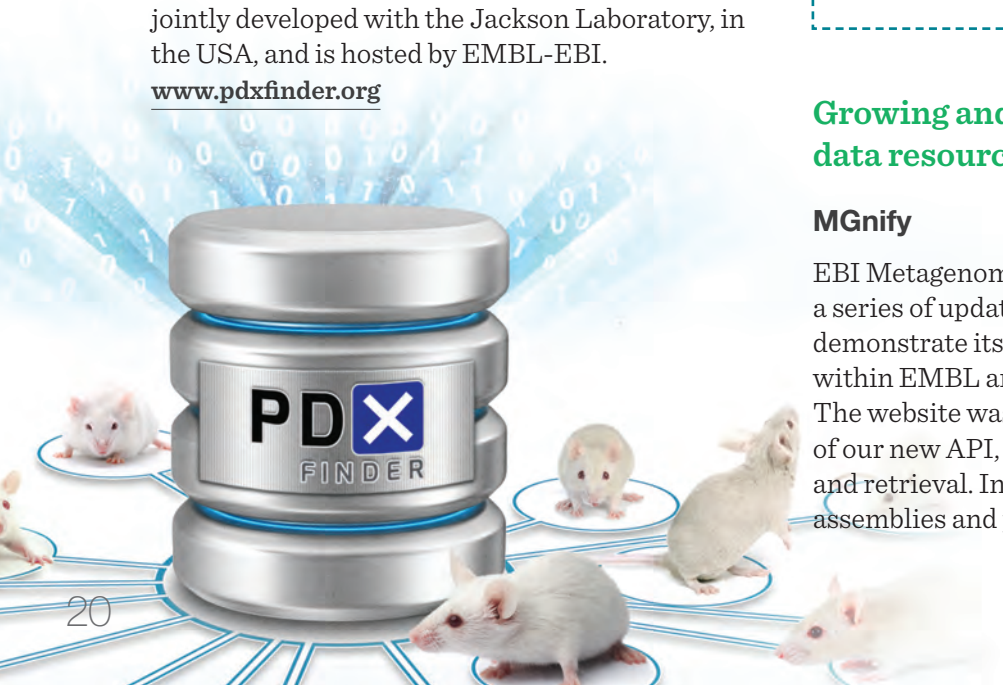
PhenoMeNal (Phenome and Metabolome aNalysis) is a comprehensive and standardised e-infrastructure that supports the data processing and analysis pipelines for molecular phenotype data generated by metabolomics applications.
http://phenomenal-h2020.eu

Single Cell Expression Atlas

This is the first EMBL-EBI resource that systematically analyses and displays single cell RNA-Seq data across different species. It was officially launched in 2018 and, as part of the Expression Atlas, it already contains over 50 datasets from 9 different species, comprising almost 40 000 cells. A new visualisation interface was developed to represent these data.
www.ebi.ac.uk/gxa/sc

PDX Finder

PDX Finder is the first open and free global online catalogue for Patient Derived Xenograft (PDX) models, which are increasingly used in cancer research. This global catalogue for PDX models was jointly developed with the Jackson Laboratory, in the USA, and is hosted by EMBL-EBI.
www.pdxfinder.org



Genome Properties (GP)

GP is an annotation system in which functional attributes can be assigned to a genome, based on the presence of a defined set of protein signatures within that genome. It is a valuable tool for the analysis and annotation of genomes. Although not a new resource, EMBL-EBI released a brand new version to the public in 2018. GP now uses InterPro entries, has a community curation interface and a new website permitting interactive genome analysis. Coverage includes more than 700 new protein properties.
www.ebi.ac.uk/interpro/genomeproperties

In the pipeline

We have also worked on developing new resources, which will be launched in 2019:

- **Protein Data Bank in Europe Knowledge Base (PDBe-KB)** is a community-driven resource, collating functional annotations and predictions for structure data in the PDB archive. PDBe-KB is a collaborative effort between PDBe and a diverse group of bioinformatics resources and research teams.
- **BioImage Archive** is a secure, searchable storage system for image data related to figures and results in publications, from the molecular to organism scale. In 2018, we have undertaken significant work to lay the foundations and customise two of our data resources, the BioStudies Database and EMPIAR, in order to build the prototype for the BioImage Archive.

Growing and improving our data resources

MGnify

EBI Metagenomics was renamed MGnify to reflect a series of updates and improvements, as well as demonstrate its increasingly collaborative nature within EMBL and the wider scientific community. The website was overhauled to take advantage of our new API, enabling enriched data searches and retrieval. In 2018, MGnify generated 27 000 assemblies and produced a metagenomics-derived

protein database, currently exceeding 1.2 billion non-redundant sequences. Also in 2018, MGnify doubled its publicly available analyses from the previous year, increasing to around 200 000. While the majority continue to be rRNA gene amplicon datasets (150 000), MGnify also hosts over 20 000 publicly available, analysed WGS datasets.
www.ebi.ac.uk/metagenomics

European Nucleotide Archive (ENA)

The ENA, one of the first universal data repositories in molecular biology, continued to grow in 2018. It now archives and presents the globally comprehensive set of $1x10^{16}$ base pairs of read data. During 2018, 5200 studies were processed including 400 000 libraries and 40 000 assemblies. ENA users include 2300 data submitters, 13 data submission brokers and tens of thousands of direct data consumers. Support was provided through 5000 helpdesk tickets.
www.ebi.ac.uk/ena

European Variation Archive (EVA)

The EVA is now the sole international variation resource for human and non-human variation, providing access to millions of sequence and structural variants. In 2018, it saw a tenfold increase in the number of submissions. The number of species represented in the EVA increased to 48 and 770 million fully-browsable variants, representing a 33% growth in variation data. Integration with Ensembl has improved during 2018, with their genome browser consuming variant and genotype data directly from EVA data sources.
http://www.ebi.ac.uk/eva

European Genome-phenome Archive (EGA)

The number of studies and datasets submitted to, and released by, the EGA increased to 1772 and 4338 by the end of 2018. The volume of data available for download from the EGA increased to 6.5PB in 2018, and it distributed 5.5PB of data. A new version of the **Experimental Factor Ontology** was developed, supporting UK Biobank phenotypes and integrated with international cross-species phenotype ontology UPheno.
https://ega-archive.org

BioSamples

The BioSamples database is now the EMBL-EBI authority for sample information and provides a centralised resource for FAIR[1] sample data with a data content of over 6.6 million samples in January 2019. In 2018, BioSamples developed and implemented the Bioschemas sample specification, allowing content to be indexed by Google.
www.ebi.ac.uk/biosamples

GWAS Catalog

The GWAS Catalog delivered in 2018 was the largest summary statistics database enabling GWAS analyses. In total 1429 studies and 40 855 variant-trait associations were added, along with full p-value summary statistics from 449 studies and improved capture of ancestry data.
www.ebi.ac.uk/gwas

Ensembl

The amount of short variant information in the Ensembl human databases doubled from 329 million to over 655 million. The number of individuals genotyped at each locus also continued to increase for human and livestock species.
www.ensembl.org

PRIDE

PRIDE received a record number of over 3000 submitted datasets in 2018 and continued to establish its place as the leading international repository for proteomics data. A total of 394 TB of data were downloaded from PRIDE, the largest volume in a single year to date. A significant amount of work was undertaken to annotate phospho-proteomics data jointly with EMBL-EBI's Beltrao research group. A new web interface and overall infrastructure for PRIDE was developed (in beta), with a focus on reliability and scalability.
www.ebi.ac.uk/pride

Expression Atlas

In 2018 the growth of high-quality bulk transcriptomic data continued. By December, the Expression Atlas contained about 3500 experiments and over 140 000 assays. These assays included over 700 RNA-seq experiments, nearly 8500 differential comparisons across 44 organisms, and nearly 800 plant experiments. At the end of

1. A set of guiding principles to make data Findable, Accessible, Interoperable and Reusable (FAIR)

2018, the Baseline Expression Atlas contained 164 RNA-seq studies, including data from many high impact studies and 21 proteomics studies.
www.ebi.ac.uk/gxa

InterPro

In 2018, InterPro's data availability and visibility were improved with a new website and associated API, supporting greater flexibility in querying, presenting and retrieving data. In total, 3555 new signatures were added, resulting in 2785 new entries. The resource now provides relationships between Homologous Superfamily and other InterPro entries, as well as improved modelling of discontinuous domains and extended intrinsic disorder annotations.
www.ebi.ac.uk/interpro

RNAcentral

RNAcentral continued growing across two major releases, bringing the total number of non-coding RNA (ncRNA) sequences to 14 million, while aggregating the data from 27 member databases. A new comprehensive genome mapping pipeline was implemented, enabling automatic mapping of ncRNA sequences to the latest genome assemblies for over 300 key species. The RNAcentral website was also improved in 2018 with the addition of a new feature viewer and significant enhancements in the text search functionality.
http://rnacentral.org

ChEMBL

ChEMBL, our widely-used curated resource of bioactive molecules, now contains more than 15 million bioactivity values from over 1.8 million compounds with a new, more flexible database schema. In 2018, ChEMBL released a new web interface with significantly enhanced functionality.
www.ebi.ac.uk/chembl

MetaboLights

MetaboLights is a global public resource for data from metabolomics experiments and derived information. In 2018, the coverage and depth of data in MetaboLights continued to expand. The team developed MetaboLights Labs – a workbench providing the infrastructure for metabolomics data processing, analysis and submission to MetaboLights – and the course *Introduction to Metabolomics Analysis* to provide guidance on the metabolomics field and to encourage data quality.
www.ebi.ac.uk/metabolights

Complex Portal

Our Complex Portal, a manually curated, encyclopaedic resource of macromolecular complexes from a number of key model organisms, released the draft of its first complete complexome, the set of annotated biomolecular complexes, for *Saccharomyces cerevisiae* (Baker's yeast). This dataset, now similar to the entire Complex Portal, also benefits from the introduction of stable identifiers, which facilitate the use of the portal as a reference resource for molecular complexes in other databases.
www.ebi.ac.uk/complexportal

In the pipeline

- We are working on a very ambitious project: a full reimplementation of the **Ensembl** website and browser. This new site will be a performant and interactive reimagining of the current browser, using modern technologies such as WebGL and JavaScript front-end frameworks. During 2018 we continued our extensive design process alongside developing our front-end framework and our genome browser.
- We have started the development of the next **UniProt** website that will take advantage of new technologies and paradigms in web development and facilitate interoperability with developments in other resources.
- Groundwork for upcoming improvements and redesign of the **EMDB** website was carried out in 2018, including improved functionality for search, statistics and validation analysis.

New data standards

Following the pilot implementation in 2017 of htsget (a code specification to provide secure streaming access to sequencing read data), 2018 saw the release of a new standard by the Global Alliance for Genomics and Health (GA4GH) called **refget**, developed by EMBL-EBI.

Refget provides access to reference sequences, an essential baseline of knowledge in bioinformatics analysis, through use of a checksum identifier derived from the sequence itself. Checksums are used to detect errors in data introduced by transmission or storage, which can verify data integrity.

Both correctly identifying a reference sequence and retrieving that sequence is an essential step when reconstructing sequencing reads held in the CRAM file format. Refget enables this by specifying a URL with a checksum identifier that returns a plain text representation of that sequence. The specification also provides access to metadata concerning a checksum identifier including its length and any known aliases. Refget is available from the European Nucleotide Archive (ENA) and provides access to ENA's catalogue of reference sequences.

Developing bioimaging resources

As part of a wider EMBL drive to improve access to imaging technology and data, bioimaging is an area of increased priority for EMBL-EBI.

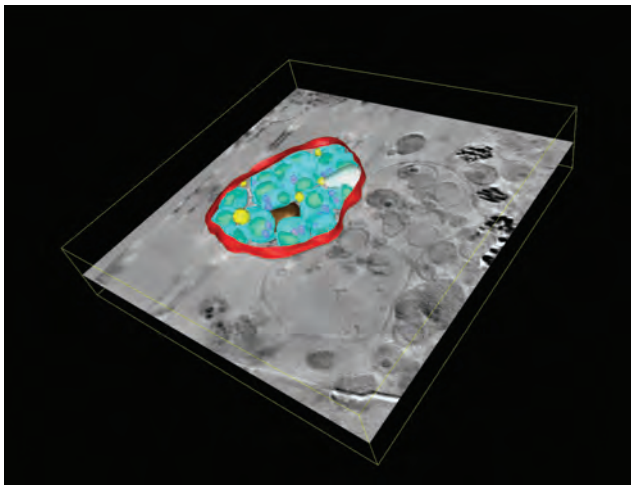
EMPIAR, our public resource for raw electron microscopy (EM) images, received 85 depositions in 2018, almost doubling its holdings to nearly 200 entries and over 100 TB of raw EM image data. As EMPIAR will be one of the pillars of our new **BioImage Archive**, a resource set to launch in 2019, preparations were made in 2018 for moving EMPIAR to an object-store back-end, so as to enable its growth to a petabyte-sized resource.

Developments around semantic segmentations of 3D bioimaging data were almost completed in 2018, including a software toolkit, a web-based Segmentation Annotation Tool and a Volume Browser through which users can interactively analyse 3D image data and segmentations with their annotations.

Another pillar of the BioImage Archive is the **BioStudies** database, a resource for aggregation of all biomedical data linked to a publication. In 2018, most of the technical work on BioStudies was related to improving the scalability and robustness of its systems, to be able to ingest and manage very large datasets, as well as to manipulate a large number of datasets. One of its new imaging data sources in the past year was datasets published in the *Journal of Cell Biology*.

www.empiar.org

www.ebi.ac.uk/biostudies



Electron tomography image with aligned segmentation for EMPIAR-10087. The segmentations highlight an erythrocyte (red) infected with the malarial parasite *Plasmodium falciparum* prior to parasite vacuoles (cyan) rupturing and exiting the cell. Segmentation courtesy of Victoria Hale of MRC-LMB.

Expanding our biocuration efforts

During 2018, we have continued to improve the human and mouse annotations of these critical genomes, which has resulted in the release of several updates to the reference **Ensembl**/GENCODE annotation for the mouse GRCm38 assembly and the human GRCh38 assembly. These updates have focused on refinement of the human protein-coding gene set, the addition of many new long non-coding RNAs (lncRNAs) and completion of the first pass of the entire mouse genome.

In 2018 we also annotated over 60 genomes for insertion in Ensembl, including major updates to key farmed animal species. We made a major effort to increase our complement of fish species, working with the fish evolutionary genomics community to identify and annotate over 40 genomes in the nucleotides archives having specific evolutionary interest.

We continued to develop our annotation pipelines to support new data types and improve accuracy. Recent improvements include integration of long-read transcriptomics data, and the use of machine learning techniques in our microRNA annotation pipeline to increase precision.

In addition, we extended the **Ensembl Regulatory** annotation build from 68 to 123 human cell lines and cell types, and from 8 to 79 in mouse, thanks in particular to the integration of numerous ENCODE epigenomic datasets. As we scale up to more samples, we are improving the robustness of our methods.

We have greatly expanded the library of transcription factor binding motifs stored in Ensembl, by switching from the JASPAR database to SELEX, thus expanding our repertoire from 98 to 632 motifs in human. Considering that there are approximately 1400 transcription factors in human, this new collection is a major leap in our efforts to comprehensively annotate trans-regulatory effects.

Furthermore, we continued to support protein annotation of key model organisms. The study of the molecular biology of model organisms provides essential data to understanding human physiological and pathological processes, being well-studied, representative organisms for their own taxonomic order. In UniProtKB, this information can be transferred computationally

For the love of proteins

As a Scientific Database Curator in the Protein Data Bank in Europe (PDBe), **Deepti Gupta** ensures that the 3D protein structures submissions are consistent and of the highest quality. She curates – or "cleans up" – the molecular structures that scientists around the world submit to PDBe, making the data useful to others.

Data consistency is crucial for scientific databases, so curators have the important role of spotting, investigating and clarifying any discrepancies or errors.

Deepti has initiated work with the Complex Portal team to integrate macromolecular complexes into PDBe, giving researchers a more comprehensive biological context for the molecule they are interested in.

Deepti is originally from India and has always had a strong in interest in structural biology and proteins. She also leads the PDBe Art project, which introduces local school children in Cambridgeshire to the beauty – and importance – of proteins.

"I have always been fascinated by proteins, these stunning microscopic structures that make up our world. I love the fact that I can contribute to broadening our understanding of proteins and share my passion with the next generation."

onto similar proteins in less well-studied, related organisms of importance in human health and disease.

Research highlights in 2018

Cancer genetics

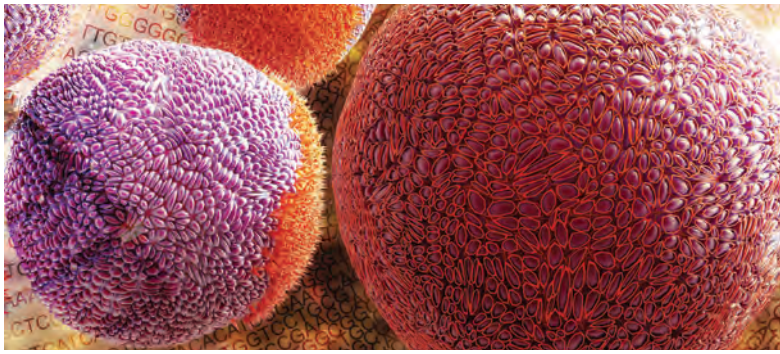
In collaboration with international scientists, EMBL-EBI's Gerstung group has discovered that it is possible to identify people at high risk of developing acute myeloid leukaemia (AML) years before they suddenly develop the disease. The study found that blood tests looking for changes in DNA code can reveal the roots of AML in healthy people. Further research could allow earlier detection and monitoring of people at risk of AML in the future, and open the prospect of identifying ways to reduce the likelihood of developing this cancer.

Abelson S, Collord G et al. (2018). Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*. doi 10.1038/s41586-018-0317-6

EMBL-EBI's Gerstung group, the University of Dundee and the Wellcome Sanger Institute have used human and worm data to explore the mutational causes of cancer. Their study shows that results from controlled experiments on a model organism – the nematode worm *C. elegans* – are relevant to humans, helping researchers refine what they know about cancer. These findings could lead to a better understanding of the causes of cancer and potentially help to identify the most appropriate treatment.

Meier B, et al. (2018). Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Research*. doi 10.1101/gr.226845.117

A large-scale systematic analysis, led by the Saez-Rodriguez and Stegle groups, found evidence that inherited genetic variants can affect drug susceptibility in cancer cell lines. The findings could help explain why cancer patients with seemingly similar tumour types respond differently to the same treatment. Instead of focusing on understanding the molecular alterations of the tumour itself, the study looked into the germline



– or inherited – component of a patient's genome, which includes the analysis of non-diseased cells. Surprisingly, results showed that the germline contribution to differences in drug susceptibility can be just as important as the contribution of somatic mutations.

Menden MP, et al. (2018). The germline genetic component of drug sensitivity in cancer cell lines. *Nature Communications*. doi 10.1038/s41467-018-05811-3

Understanding disease

Researchers in the Mouse Informatics team at EMBL-EBI, with colleagues at the Helmholtz Center Munich and the International Mouse Phenotyping Consortium (IMPC), have identified hundreds of genes that could play an important role in the development of metabolic diseases such as diabetes. Their study identified novel links to metabolic traits for 429 genes in mice, and showed that 23 of these genes may play a role in human diabetes.

Rozman J, et al. (2018). Identification of novel genetic elements in metabolism by high-throughput mouse phenotyping. *Nature Communications*. doi 10.1038/s41467-017-01995-2

For the first time, researchers proved that DNA sequencing technologies are accurate enough to predict which commonly used anti-tuberculosis drugs are most suitable for treating a specific patient. The study, a wide international collaboration that included EMBL-EBI's Research Group Leader Zamin Iqbal, could mark a shift towards a genetics-based approach to the disease, which could guide treatment decisions and, in turn, result in saving many lives.

Allix-Béguec C, et al. (2018). Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *New England Journal of Medicine*. doi 10.1056/NEJMoa1800474

Researchers from Imperial College London and EMBL-EBI's Birney group have shown that the human T-lymphotropic virus (HTLV-1) changes the folding pattern of human DNA in infected cells. They explained that the resulting disruption of gene function increases the risk of leukaemia. The research shows that HTLV-1 acts at a large number of sites across the human genome, disrupting the regulation of tens of thousands of genes.

Melamed A, et al. (2018). The human leukemia virus HTLV-1 alters the structure and transcription of host chromatin in cis. *eLife*. doi 10.7554/eLife.36245

Data analysis methods

A collaboration between the Babraham Institute, the University of Edinburgh and EMBL-EBI's Stegle group has resulted in the first method to analyse three molecular layers simultaneously during single-cell analysis. Comparing the molecular interactions of cells in this way reveals differences that could have an impact on both the early stages of life and the first stages of diseases such as cancer. The new method is called single-cell Nucleosome, Methylation, Transcription sequencing (scNMT-seq).

Clark S, et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*. doi 10.1038/s41467-018-03149-4

EMBL researchers, including EMBL-EBI's Stegle group, have designed a computational method to jointly analyse multiple types of molecular data from patients in order to identify molecular signatures that distinguish individuals. The method, first called Multi-Omics Factor Analysis (MOFA) could be particularly useful for understanding cancer development, improving diagnosis and suggesting new directions for personalised treatment.

Argelaguet R, Velten B, et al. (2018). Multi-omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*. doi 10.15252/msb.20178124

Understanding other species

Researchers at EMBL-EBI, the University Grenoble Alpes and collaborators have found that similar characteristics in sheep and goats can result from different patterns of gene selection. The study sequenced and analysed for the first time the genomes of the wild Asiatic mouflon and the bezoar ibex, comparing them to those of domestic sheep and goats.

Alberto F, et al. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*. doi 10.1038/s41467-018-03206-y

Scientists at the Wellcome Sanger Institute and the EVA team at EMBL-EBI have discovered significant diversity in the genomes of 16 laboratory strains of mouse. The research revealed for the first time notable genetic diversity in these strains. Significant areas of the genome where variation was found include regions impacting immunity, pathogen defence and sensory function. These variations also differ widely from the current reference strain, suggesting this discovery has the potential to significantly impact future human disease research.

Lilue J, et al. (2018). Multiple laboratory mouse reference genomes define strain specific haplotypes and novel functional loci. *Nature Genetics*. doi 10.1038/s41588-018-0223-8

Increasing areas of application of bioinformatics

As technology advances, human health, agriculture, biotechnology and environmental science are likely to become important areas of application for bioinformatics.

Human health

In relation to human health, one example is our involvement in the **Global Alliance for Genomics and Health (GA4GH)**, a policy-framing and technical standards-setting organisation that seeks to enable responsible genomic data sharing within a human rights framework. In 2018, apart from the release of a new data standard called refget, we have worked on the definition of the htsget specification for genomic data streaming and developed implementations on EGA and EVA.

The EVA continues to maintain the Variant Call Format (VCF) specification, which is the standard for representation of genomic variation. The EGA has deployed the Data Use Ontology (DUO), which allows Data Access Committees to categorise the data use conditions applicable to their datasets in a machine-readable way, and collaborated with the Wellcome Sanger Institute to apply DUO terms to their datasets. The EGA has contributed to the Beacon Project, a GA4GH driver project that makes it easier for researchers to find relevant datasets. BioSamples has implemented a GA4GH API in 2018 to provide standard access to sample metadata.
www.ga4gh.org

Of standards and ontologies

Melanie Courtot is a Metadata Standards Coordinator in the Archival Infrastructure and Technology team, which provides vital infrastructure for EMBL-EBI's reference archives.

Melanie oversees the BioSamples database and Data Submission Portal. She participates in several international initiatives that facilitate efficient sharing and integration of high-quality data. Melanie also leads the development of the Data Use Ontology (DUO), which enables authorised researchers to securely access genomic data.

"It can take between two and six weeks for a researcher to get access to a clinical dataset," explains Melanie. "This is partly because of the entirely justified restrictions and safeguards of accessing clinical data, but partly because the process is manual. DUO allows us to speed up the process. It essentially tells us who can use a dataset and for what purposes.

"We hope to create a system that the wider scientific community can use to accelerate scientific discovery. Ontologies and data standards are the foundations of such a system."

In 2018, our work for the **International Mouse Phenotyping Consortium (IMPC)** included adding longitudinal aging data, a tool to investigate age-related phenotypes and disorders relevant to human health, to the IMPC portal. We continued our focus on domain-specific annotation of human and mouse proteins with roles in specific areas of human health and disease. The IMPC and Patient Derived Xenograft Integrator provide large scale access to functional data necessary for translational research and drug discovery.
http://www.mousephenotype.org/

COMPARE is a multidisciplinary research network working on the rapid identification, containment and mitigation of emerging infectious diseases and foodborne outbreaks. As part of the project and working with two major European organisations involved in routine surveillance, EMBL-EBI piloted the **Pathogen Portal**, a pathogen surveillance platform. In 2018, we have extended the functionality of the portal to include a wizard-like search, management tools giving greater control over cloud analyses, data sharing controls and integration of cloud-hosted iPython Notebooks.
www.ebi.ac.uk/ena/pathogens

Agriculture

Wheat (*Triticum aestivum*) is one of the world's most important food crops and **Ensembl Plants** release 40 included the wheat reference genome (RefSeq v1.0) from the International Wheat Genome Sequencing Consortium (IWGSC). The bread wheat genome is hexaploid and five times the size of the human genome, which presented a number of unique challenges. The new assembly allows researchers to pinpoint where in the highly complex genome a certain gene is located, making it easier to accurately combine desirable traits, such as high yield and disease-resistant genes, thus producing superior wheat varieties. Ensembl Plants helped to develop new web displays optimised for polyploid genomes and to assist researchers studying EMS-induced variation.
http://plants.ensembl.org

In addition, the **Expression Atlas** was extended to include new plant species and nearly 800 plant experiments.

Environmental sciences

In 2018, EMBL-EBI continued to play a role in data coordination and analysis services for sequence and related data across a breadth of scientific areas (from livestock genomics to marine microbiology) offering web portals, APIs and dedicated support and training services.

In November, we announced our participation in the **Darwin Tree of Life Project**, a major collaboration to sequence the genetic code of 66 000 UK species. The initiative is the UK part of a global effort – the Earth BioGenome Project – to sequence all 1.5 million known species of animal, plant, protozoa and fungi on Earth. The global project will ultimately create a new foundation for biology to drive solutions for preserving biodiversity and sustaining human societies.
www.earthbiogenome.org

Providing better access to literature

Europe PMC is an open database of worldwide life sciences literature. In 2018, following emerging trends in scholarly publishing, Europe PMC incorporated preprints – i.e. non-peer reviewed scientific manuscripts – into its database, alongside peer-reviewed journal articles. Over 60 000 records from 10 different preprint repositories, such as *bioRxiv, Preprints.org and ChemRxiv,* were indexed. Preprints in Europe PMC are clearly labelled and linked to their final published version. Users can now access preprint citation information, find links to open peer reviews and comments, and claim preprints to their ORCID records.

By the end of 2018, Europe PMC offered free full text for 5.2 million of around 35 million records. To extend access to open content, the database has integrated with Unpaywall – a tool harvesting legally-uploaded content from over 50 000 open data sources. Thanks to this feature over 50% more open content is now easily discoverable through Europe PMC.

Furthermore, a new public beta site was launched in 2018 offering streamlined user experience. A major improvement is the merger of the abstract and full text pages, removing the need to switch between different records. A preview for all publication figures was also added, which gives an overview of the key findings at glance. Other improvements concern search results layout and design, as well as better user interface for the annotations feature.
https://europepmc.org/

35 million
abstracts

5.2 million
full-text articles

62 000
preprints

4.2 million
patents

708 000
Agricola records

53 600
biomedical theses

865
clinical guidelines

Extending collboration and coordination

Collaboration is the lifeblood of EMBL-EBI and nearly all of our activities happen in conjunction with one or more partners and with entities of all types. We highlight here some of our major collaborative activities in 2018.

Human Cell Atlas Project

The Human Cell Atlas (HCA) project will organise and standardise terabytes of data for billions of cells, across multiple modalities, generated by hundreds of labs around the world. This data will be open and easily accessible to all researchers, enabling the scientific community to innovate rapidly without barriers to data access.

At EMBL-EBI, we are building a modern, cloud-based, modular **Data Coordination Platform (DCP)** for organising and sharing HCA data. We are also developing the metadata standards, the ingestion services and the techniques to wrangle single-cell data from a variety of established and emerging sequencing and imaging technologies.

The DCP started development in 2017, as an international collaboration between four world-leading bioinformatics and engineering institutes: EMBL-EBI, the University of California Santa Cruz (UCSC), the Broad Institute and the Chan Zuckerberg Initiative (CZI).

The team at EMBL-EBI has worked to address a huge gap in the community and has produced brand new single-cell sequencing metadata standards to define how to best describe experimental techniques and protocols. The first datasets were made available in April 2018. To ensure we can collect data at the scale we expect, we have built the DCP Ingestion Service, which can be used to submit and validate data and metadata to the HCA. This service supports direct submission from labs, using spreadsheets and web interfaces with wrangler assistance, or programmatically, to ensure large-scale sequencing efforts can upload data directly to the HCA for storage and analysis in the cloud. In 2018, the first release of the DCP infrastructure was released to beta testers.

www.humancellatlas.org

Opening doors to data

A software developer in the Human Cell Atlas team, **Alegria Aclan** works on building the 'ingest' infrastructure for the project's data submission platform.

After studying for a degree in Computer Science and working for financial companies in the Philippines, Alegria made the move to EMBL-EBI as she wanted to leave the commercial sector and make a contribution to humanity.

The team is working hard to make the submission infrastructure reliable and scalable, before making it freely available to researchers all over the world.

"We are expecting a huge amount of data to come into the HCA," she says. "We want to have a nice application for submitters to use and through which to submit their data. The application should be easy to navigate and we need to address how we are going to encourage users to submit their data to our particular platform.

"When you are working in this area, it's not just the computer and technology aspect, you also get to learn the world of the individuals using the software."

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 and Takeda and, in 2018, was joined by two new partners: Celgene and Sanofi.

In 2018, the first version of the Open Targets Genetics portal was launched, which processes GWAS summary statistics to provide a view on likely candidate explanatory genes for each GWAS locus. This has gained much traction in the community.

The Open Targets Platform, a comprehensive and robust data integration tool for access to and visualisation of potential drug targets associated with disease, released an update every two months during 2018. The platform introduced new data, including tractability scores and data from systems biology experiments.

Additionally, in 2018, 14 new experimental and informatics Open Targets projects were initiated within EMBL-EBI and the Wellcome Sanger Institute in collaboration with the industry partners.
http://genetics.opentargets.org
www.targetvalidation.org

EMBL-EBI Industry Programme

The purpose of our subscription-based Industry Programme, and part of the mission of EMBL-EBI, is to **disseminate cutting edge technologies to industry**. Member companies primarily represent the pharmaceutical sector (most of the top 20 pharma companies) but also the agrifood, nutrition and healthcare industries.

During 2018, we organised our first ever workshop on the West Coast of the USA, hosted by Pfizer, with the topic Cancer Systems Biology, which had more than 100 delegates. We also welcomed Daiichi Sankyo as a member of the Industry Programme and organised two seminars in Japan on Artificial Intelligence (AI) and Big Data in Drug Discovery.

During the year, we delivered 12 successful workshops and symposiums, representing both technical and more therapeutically focussed areas, in addition to our well-established quarterly meetings for programme members and our annual meeting for small and medium enterprises.

Other collaboration work in 2018

- **Semantics resources** from our Molecular Archives cluster, including ontologies and annotation tools, are now used by Open Targets, the European Rare Disease Community, an Innovative Medicines Initiative project (FAIRPlus) and by pharma and agrifood companies, after a successful collaboration with the Pistoia Alliance.
- EMBL-EBI researchers joined the **DepMap** (depmap.sanger.ac.uk) analysis workgroup to study what determines differences in gene essentiality and drug sensitivity across different cancer cells.
- Collaborations with Addenbrooke's Hospital, in Cambridge, UK, were set up for using electronic healthcare resources (EPIC) to predict important factors related to patient treatment plans.
- We engaged with the European Open Science Cloud (EOSC) through strategic and technical work within the EOSC-Hub and EOSCpilot projects.
- As participants in the Pan-Cancer Analysis of Whole Genomes initiative, we lead a working group on data integration jointly with our collaborators from the University of California Santa Cruz and the University of Zurich.
- We established a collaboration with research groups in India to develop capacity in data archiving and dissemination.

CABANA

CABANA is a project that aims to address the slow implementation of data-driven biology in Latin America by creating a **sustainable capacity-building programme**. Funded by the UK Government's Global Challenges Research Fund (GCRF) and led by EMBL-EBI, the project is driven by an international consortium of ten organisations. CABANA combines research secondments, workshops, train-the-trainer activities and new e-learning content to strengthen bioinformatics research in three challenge areas of special relevance to Latin America: communicable disease, sustainable food production and protection of biodiversity.

In 2018, the first full year of CABANA's operation, all of its major strands of activity were up and running. Our first three secondees arrived at EMBL-EBI and started their projects; a further eight followed in early 2019. Launching the workshop programme, we delivered two workshops in Argentina and one in Colombia, plus our first train-the-trainer workshops in Chile and Colombia. We also delivered a series of interactive webinar-based tutorials from August to November 2018.

www.cabana.online

Aquiring new skills

In the first group of CABANA secondees arriving at EMBl-EBI in 2018 was **Guillermo Rangel**, a chemical engineer and microbiologist from the Universidad de los Andes, in Bogota, Colombia.

With an interest in bacteriophage biology and evolutionary ecology, Guillermo was drawn to the opportunities that bioinformatics and big data analysis offer to the study of bacteriophage–host interactions at the microbiome level. In his research secondment project, supervised by Rob Finn (EMBL-EBI) and Alejandro Reyes (Universidad de los Andes), he is developing a pipeline for taxonomic annotation of viral sequences in metagenomic datasets. The output of the project is expected to contribute to the expansion of the analysis pipeline currently employed by MGnify, the EMBL-EBI open-source metagenomic analysis tool.

"The CABANA project allowed me to acquire a set of new skills that are of enormous relevance for working with large omics datasets", says Guillermo. "Participating in this project has been a critical step which has boosted my confidence to work in bioinformatics. This will undoubtedly open many doors for me in the foreseeable future".

Continuous improvement, maximising efficiency

To support the growth of bioinformatics and meet the demands in the field, given our current growth rates, we expect EMBL-EBI as an organisation to increase by a third over the next ten years. As we grow as an organisation, we are constantly striving to combine increasing scale with responsible and planned growth.

Engaging with funders, policy makers and regulators

EMBL-EBI continues to work closely with our strategic and life science funding partners from across the globe. In 2018 we had 26 separate funders providing support for **165 active external projects.** We welcomed Novo Nordisk Foundation, Russian Foundation for Basic Research, Save the Tasmanian Devil Program Appeal and The Genetics Society to our expanding supporter base. We continued our close collaboration on the large Human Cell Atlas project with the Chan Zuckerberg Initiative as a new strategic partner. A full list of our funders can be found at the end of this report (page 47).

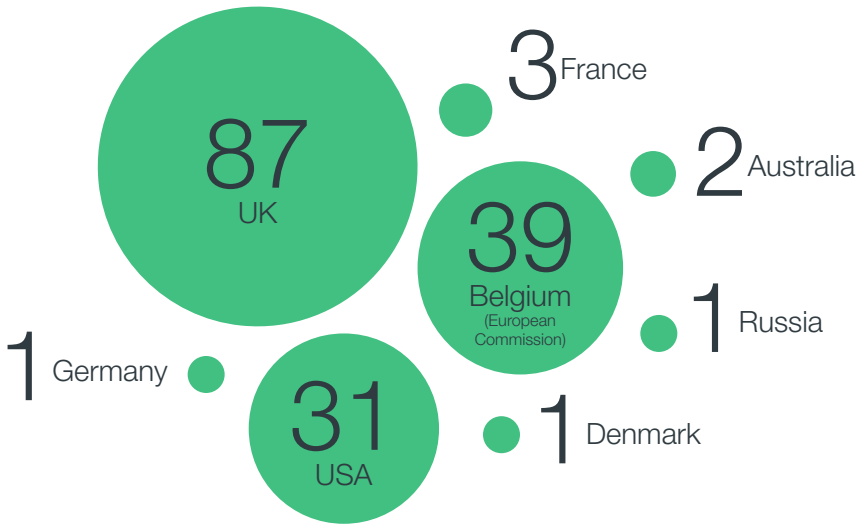
Collaboration is key to all that we do. In 2018, we worked with 603 collaborating institutes, ranging from Aarhus University Hospital to Zurich University of Applied Science and spanning 62 countries, including Taiwan for the first time. This growth in collaborators also reflects the growing diversity of scientific domains utilising the data EMBL-EBI resources provide. Analysis of citation data[2] in recent years highlighted a near tripling of distinct journal categories referencing our data resources, including key articles, resource names and accessions.

EMBL-EBI maintains a helpful dialogue with policy makers in the UK and Europe. To this end we host inbound visits to the Institute and in 2018 welcomed key scientific and political representatives, including Vince Cable, British politician; Matt Hancock, the UK Secretary of State for Health and Social Care; Fiona Watt, Executive Chair of the UK Medical Research Council; and Sarah Wilkinson, Chief Executive of NHS Digital.

The continued support from our host nation, the United Kingdom, has been key to EMBL-EBI's ongoing growth and success. The sustained capital infrastructure support, allowing EMBL-EBI technical and physical infrastructure to efficiently scale to cope with the continued dramatic rise in biological data production and use, has been successfully extended.

Globally funded

Number of grants received by EMBL-EBI in 2018 by funding body location:



2. From Europe PMC data journal catagorisation available for 30% of citing journals

Managing our growth

As we grow, we continue to adapt and scale our processes. In 2018, we reviewed internal management of information for areas such as office utilisation and how we track change.

With a **predicted 12% growth** in personnel in the period 2018-2021, current and future space requirements needed to be reviewed. An audit was completed in 2018, resulting in recommendations to maximise the efficiency and comfort of current space, allowing for growth across all areas of the organisation.

In light of the expected growth, in 2018 we also launched an EMBL-wide, criteria-based procurement system to facilitate large-scale procurements, and developed an automated system for tracking key performance indicators to support our reporting of objectives, which are part of the infrastructure programme.

Developing our technology infrastructure

The last year has seen improvements in the **security policy and security infrastructure** across EMBL-EBI, expansion of our application and infrastructure collaborations, laying the foundations of a cloud native environment that can operate across hybrid cloud resources, and improvements in the presentation and transparency as to how we operate our information technology (IT) services.

Our collaborations inside and outside EMBL-EBI continue to help develop our services. Application collaborations with BioExcel and the EBI Training team are providing new requirements for the EBI Cloud Portal. Work with ELIXIR and Health Data Research UK (HDR UK) has led to a distributed workflow execution environment using specifications from the Global Alliance for Genomics and Health (GA4GH).

New funds for infrastructure

In the spring of 2019, UK Research and Innovation (UKRI) awarded £45 million to EMBL-EBI to enhance the institute's **technical infrastructure.** The funding, which comes from the UKRI's Strategic Priorities Fund, will support EMBL-EBI's existing and emerging data resources, including in areas of major interest, such as genomics and bioimaging.

To obtain the funding, we worked in 2018 in partnership with the Biotechnology and Biological Sciences Research Council (BBSRC) to develop a business case, including the strategic vision for the project, its wider economic impact, financial and commercial viability and governance structures. The document was submitted and reviewed by UKRI in the summer of 2018 and formally announced in the 2019 UK Spring Budget.

The Strategic Priorities Fund supports multidisciplinary and interdisciplinary programmes at the cutting edge of research and innovation.

A **hybrid cloud strategy** has been developed during 2018 that builds upon the infrastructure collaborations that we have with cloud infrastructures within the community (e.g. ELIXIR), commercial providers (e.g. Helix Nebula Science Cloud) and with the European Open Science Cloud (e.g. EOSC-Life). Scalability and portability across this hybrid environment will be driven through the adoption of a Cloud Native Environment encompassing common continuous integration/continuous delivery (CI/CD) tools and services and an internal container execution environment that is comparable to external cloud providers.

For our internal users, a new implementation of our service portfolio and service catalogue provides a comprehensive description of our services, links to documentation, self-management pages, support, and performance metrics. Behind-the-scenes work has continued to improve the resilience of our services driven by feedback from our users as to which resources are the most critical for them. Work continues to record the use of these individual services and how much they cost to deliver.

The introduction of the EMBL's Internal Policy on Personal Data was supported with the development of coordination tools, explanatory presentations, workshops and face-to-face discussions and support.

Maintaining vital tools

Originally from South Korea, **Youngmi Park** is a software engineer in EMBL-EBI's Web Production team. She leads a four-person team dedicated to developing and maintaining one of the organisation's most vital tools, the EBI Search application programming interface (API).

EBI Search acts as a gateway to the vast amount of data available in EMBL-EBI resources by providing a fast and uniform way to access them.

Youngmi has helped improve the representation of cross-references between data resources to help users better understand the relationships between the data objects they are viewing and those in other related domains.

"I really like my job, it gives me a strong motivation to learn something new and to make the system better", says Youngmi. "It's always an exciting challenge to keep the API stable while also making it better by integrating features such as cross-referencing between databases."

Implementing an improvement culture

A series of **cross-departmental projects** (called "glue projects") are initiated each year to develop tools and infrastructure of mutual value to the participating research and service teams. These projects bring together expertise from many teams to focus on a shared challenge driven by user needs. In 2018, two glue projects were completed: Complex Portal and Resource Allocation Portal, an internal tool to link the use and cost of technical services by teams and data resources.

In order to improve the digital experience of our users, we also ensure that we implement robust and responsive software development practices with website designs and search systems. An in-house team of user experience (UX) professionals help ensure software development is combined with iterative feedback from our users in a way that is structured, comprehensive and actionable. In 2018, we applied UX design principles to several projects, including the Human Cell Atlas and PDX Finder.

Building capacity and capability

As an institute with one of the world's highest concentrations of bioinformaticians and related career professionals, we play a major role in training and growing generations of skilled people. Recruiting and developing a workforce of multidisciplinary skilled staff is essential in tackling the pressing challenges of bioinformatics.

Our people

Staff in 2018 in full-time equivalent (FTE)

645

567
Staff members

45
Postdoctoral fellows

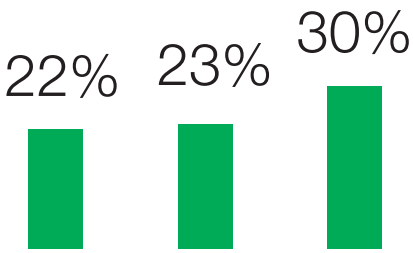
33
Predoctoral fellows

Gender distribution of staff in 2018

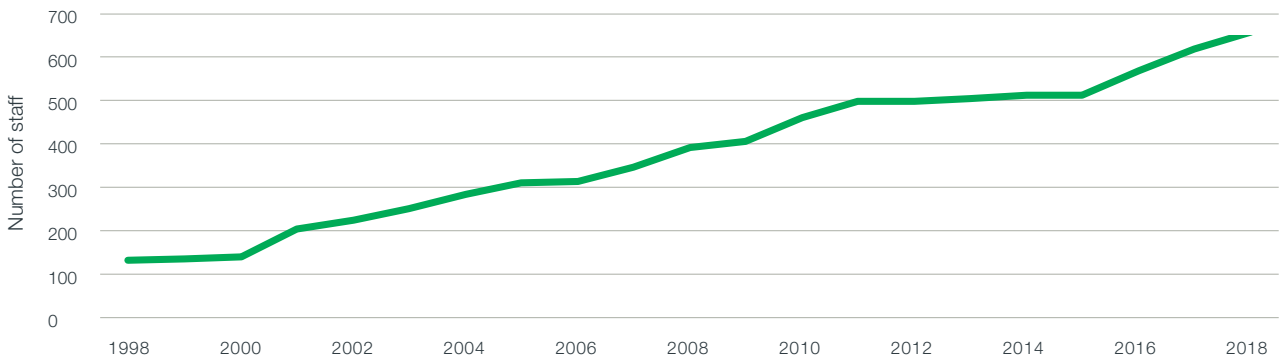
63%
male

37%
female

Senior roles held by women



Staff growth at EMBL-EBI, 1998 - 2018



2018 was another busy year for recruitment at EMBL-EBI. Overall, we welcomed **200 newcomer**s during the year, including 115 staff members, 26 fellows, 40 trainees, 12 supernumeraries and 7 visitors.

EMBL-EBI strongly values and supports equality and diversity. Even though the majority of our staff members are male, we continued making progress in relation to gender balance, especially in leadership positions. By the end of the year, almost one-third of our senior roles were held by women, the highest ever ratio in our institute. We are committed to creating and maintaining a culture in which equality of opportunity exists for all staff, prospective and existing, and will continue our work to improve gender balance in the institute.

Across EMBL, work has been undertaken to update employment terms and conditions, including improvements to parental leave, which were agreed by EMBL Council in November 2018 for implementation in early 2019. Flexible working guidelines were also implemented across the organisation, after a successful two-year pilot at EMBL-EBI.

With the Cambridge area rapidly developing into a biotech hub and data science skills in demand globally, competition for skilled technical

professionals, such as software engineers and database administrators, has increased significantly. In addition, uncertainty caused by Brexit also contributed to the challenge of recruiting talented professionals in 2018.

To overcome these difficulties, we developed communication campaigns, using video and social media, to promote recruitment of fellows and software developers. We have also participated as exhibitors in scientific conferences – Intelligent Systems for Molecular Biology (ISMB) and European Conference on Computational Biology (ECCB) – and job fairs, such as JOBIM (France) and Nature Jobs (UK). On our website, we reassure job seekers that we recruit globally and Brexit does not affect our recruitment conditions.

New leadership

Five new group and team leaders joined EMBL-EBI in 2018, working in a range of areas, from administration to research, IT and data resources.

Sarah Butcher, Team Leader, Software Development and Operations. Sarah leads a new technical team at EMBL-EBI, which develops, adapts and operates software to meet our internal needs. The team also supports engagement with external projects, such as the Human Cell Atlas, through software and services.

Rachel Curran, Head of Administration and Operations. Rachel's team is responsible for the overall delivery of administrative support at EMBL-EBI, and includes Finance, Human Resources, Strategic Project Management and Facilities/Health and Safety Teams.

Kevin Howe, Team Leader, Eukaryotic Annotation. Kevin's team is responsible for the gene annotation and comparative genomics components of Ensembl, and EMBL-EBI's contribution to WormBase and the Alliance of Genome Resources.

Irene Papatheodorou, Team Leader, Gene Expression. Irene's team focuses on gene expression analyses at tissue and single-cell level across species. The team delivers tools and services for the submission, archiving, analysis and visualisation of functional genomics experiments.

Virginie Uhlmann, Research Group Leader. Virginie's research focuses on bioimage analysis, specifically continuous representations for image analysis. Her group works on collaborative, interdisciplinary projects with biologists and software developers.

Awards and honours

Ewan Birney, Joint Director of EMBL-EBI, was made a Commander of the British Empire (CBE) in the Queen's 2019 New Year's Honours list. He was recognised for his services to computational genomics and leadership across the life sciences. Ewan was also awarded a prestigious Synergy grant from the European Research Council in 2018.

Janet Thornton, EMBL-EBI's Director Emeritus, was named joint Vice President of the European Research Council's (ERC) governing body, the Scientific Council, in December 2018.

Providing bioinformatics training

One of EMBL-EBI's missions is to deliver world-leading **training in bioinformatics** and scientific service provision to the research community. To this end, in 2018, we participated in 355 training and outreach events, supporting biomedical and life-science professionals throughout the world.

In 2018, our on-site training programme delivered 25 courses, including new ones, such as bioinformatics and functional genomics in zebrafish. In collaboration with the University of Cambridge Institute of Continuing Education, we launched the Postgraduate Certificate in Biocuration, designed to educate new biocurators with the requisite skills to work in this field and provide established biocurators with a recognised qualification that reflects their diverse life-science and computational skills.

We have also successfully piloted the use of **robot avatars** and remote access to our training compute to enable the participation of course delegates who cannot travel. In 2018, two students on the genomic medicine master's programme, who were new mothers, were able to participate in course activities remotely, enabling them to continue their training while caring for their babies.

Online, we delivered our largest ever webinar programme, including eight new courses and a new webinar series on data management for the biomolecular sciences. In total, 64 webinars were delivered live throughout the year and then made available through Train online, our freely available e-learning portal. Train online was accessed by over 485 000 unique IP addresses.

As part of an award from the Wellcome Trust, EMBL-EBI conducted over **100 Ensembl training workshops** in low-to-middle income countries (LMIC) in 2018, including Argentina, India, Malaysia, Morocco, Nigeria and Rwanda. Besides training on the Ensembl browser and its associated web APIs, these activities included a train-the-trainer component to ensure that the legacy of our LMIC training focus continues into the future and will be developed to allow participants to keep informed of Ensembl developments. In addition, the EMBL-EBI Training Programme team has also delivered eight train-the-trainer events, training 89 bioinformatics instructors.

A training journey

As an Outreach Officer, **Ben Moore** delivers some of the worldwide training for the Ensembl Genome Browser, which offers open access to genome, gene, variation and comparative genomics data.

The Ensembl Outreach team travels to the furthest corners of the globe to deliver up to 100 workshops every year. Thanks to a recent Wellcome Trust grant, the team has been able to provide more training in developing countries.

"We work closely with the host institute to tailor the content to their need," says Ben. "One week I may be going to a salmon research facility where they use genomics to understand and monitor breeding, while the next week, I might go to a hospital and deliver a workshop focused on cancer.

"I love the variety of the work but mostly, I like seeing people get excited and inspired about what they could do with our data."

Engaging with the public

EMBL-EBI's work is increasingly of direct importance to biomedical, agricultural and environmental research, and we are keen to convey our impact and relevance to a broad, general audience. To achieve this, our staff members participate in a range of activities that seek to engage people of all backgrounds with the boundless possibilities offered by genomics.

We were involved in numerous public engagement events in 2018, including two exhibitions and numerous student visits as part of the **PDB Art** project (PDBe.org/art), which collaborates with several schools and art societies in Cambridgeshire. In this project, scientists help students explore our protein structure database, PDBe, and ways of depicting molecules, using structure as inspiration for the artworks. The project reached over 100 secondary students across four different schools in 2018, and received around 60 pieces of artwork from the students for the exhibition.

In September 2018, scientists from EMBL-EBI participated in the **LifeLab** consortium, a public engagement project funded by the European Union as part of the European Researchers' Night initiative, a one-day Europe-wide celebration of science in society. The project, led by the Wellcome Genome Campus (home of EMBL-EBI), delivered a series of activities in shopping centres, cafes and public spaces across Cambridge and Peterborough, in the UK, reaching over 3000 people.

Supporting global expansion of biomolecular resources

ELIXIR

EMBL-EBI serves as the European Node of ELIXIR, an international consortium with 22 member countries, bringing together over 200 research institutes and over 600 experts. ELIXIR connects the national nodes in bioinformatics infrastructure in much of Europe with EMBL-EBI and provides an excellent model for how national research infrastructures can work together to form a united European operation. Our deep involvement in ELIXIR since its inception has given this new infrastructure a firm footing in the bioinformatics community and provided us with a formal mechanism for collaboration in data provision and standards setting.

To develop standards, services and training within specific life science domains, ELIXIR establishes communities, which bring together experts across ELIXIR Nodes in a particular domain. Throughout 2018, ELIXIR established **three new communities** – for proteomics, metabolomics and for the Galaxy platform. The Proteomics and Metabolomics teams at EMBL-EBI played an important role in establishing the corresponding communities and in driving their development.

In 2018, ELIXIR continued its collaboration with the Global Alliance for Genomics and Health (GA4GH), in particular on the Beacon project, a data discovery resource to facilitate the sharing of genomics data. The Beacon API released in October 2018 became the first data interoperability standard from the GA4GH 2018 Strategic Roadmap.

In December 2018, the ELIXIR Interoperability Platform announced the initial set of ELIXIR Recommended Interoperability Resources, a set of key resources that facilitate the FAIR-supporting activities in scientific research. Two EMBL-EBI services – Identifiers.org and Ontology Lookup Service – were included in the initial portfolio.

Advancing secure human data sharing

In collaboration with the Centre for Genomic Research (CRG) in Spain, EMBL-EBI is working on 'federated access' models for its data resource European Genotype-phenotype Archive (EGA), which holds data from human studies with research usage conditions based on consent permissions in the datasets.

This model will expand as more national EGA nodes are established to support the ethical, legal and social requirements of countries that would need to retain data within their jurisdiction while building capacity for secure data sharing.

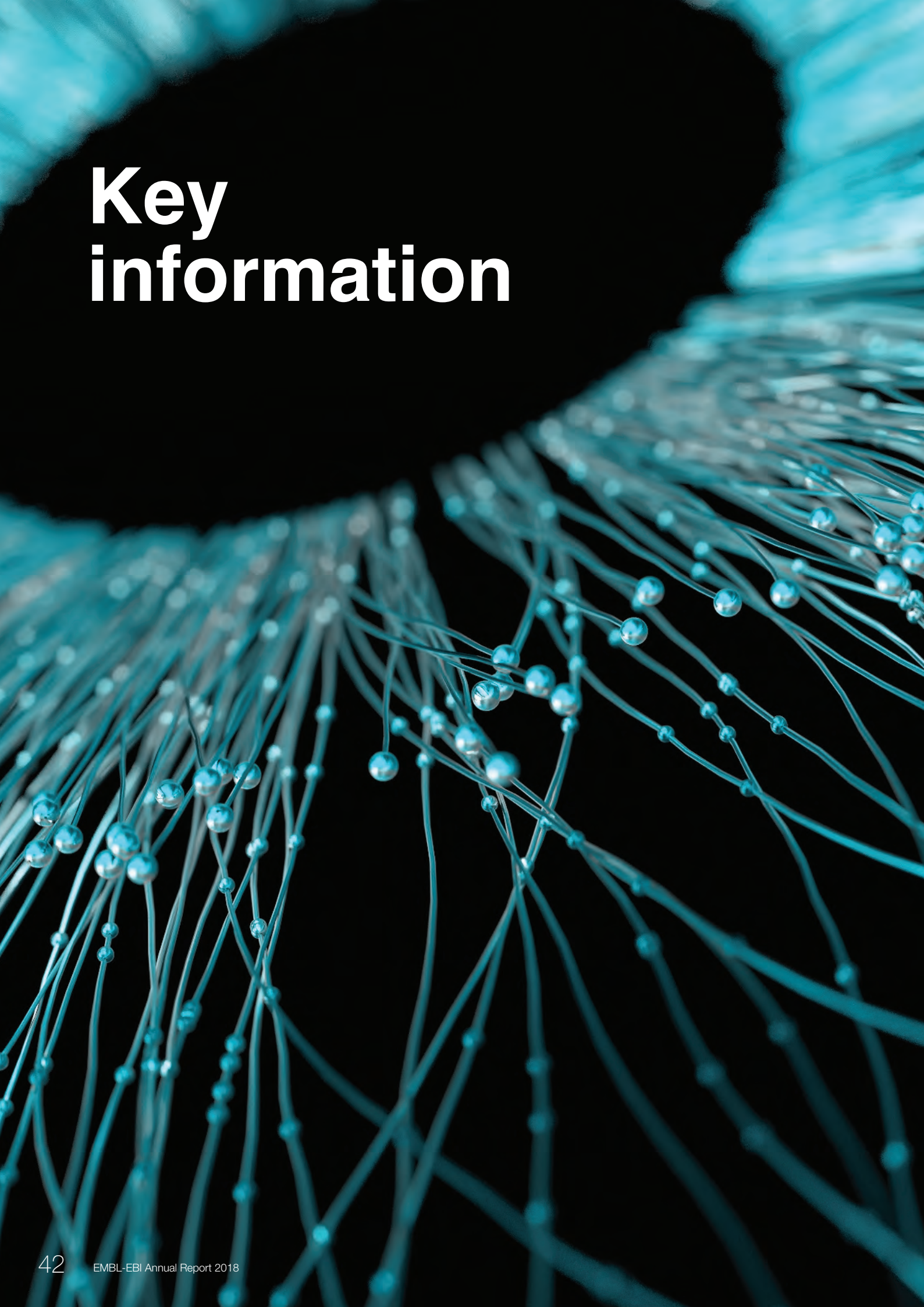
Attendees at the annual ELIXIR All Hands meeting held in Lisbon, Portugal, June 2019. Credit: ELIXIR

EMBL-EBI will provide strategic and operational support in establishing local (community) EGA roll outs that involve backend design, policy development and installation, as well as standardised tiered agreements for new partners.

Supporting new initiatives

The **Global Biodata Coalition (GBC)** is a nascent international organisation that will coordinate support for essential life sciences data resources, ensuring a sustainable future data ecosystem available to all researchers worldwide. During 2018, EMBL-EBI staff, including Joint Director Rolf Apweiler, have been active in helping to develop the organisation through membership in the GBC steering committee and helping develop the business case and financial model for the GBC.

Key information



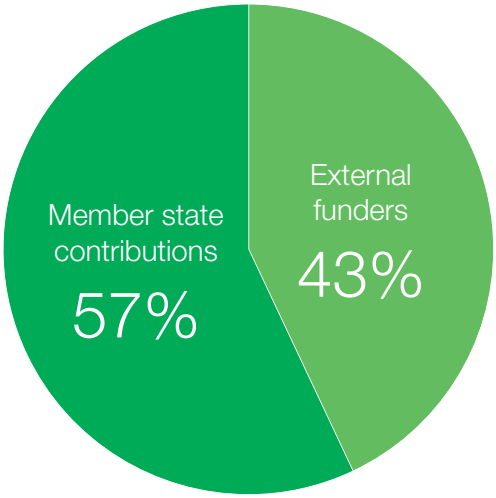
Financial figures

We are grateful for the continued support of our member states and other funding bodies, who in 2018 helped us maintain our data resources, retain staff and respond to the requirements of the international scientific community.

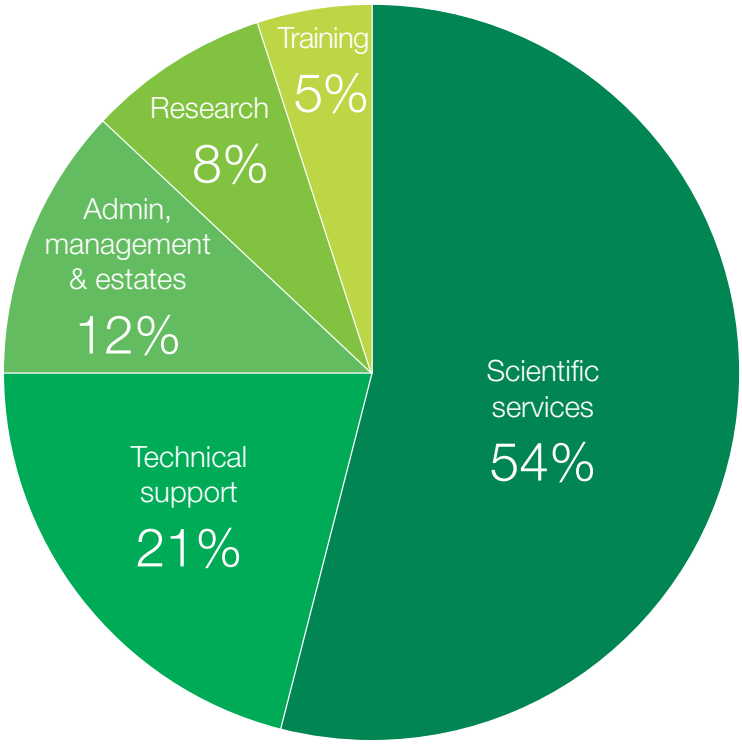
The total operating expenditure of EMBL-EBI in 2018 was €79.3 million, which is higher than previous years. The expenditure reflects the increases in activity required to maintain EMBL-EBI resources at a time when data depositions and usage in the field of molecular biology continue to grow globally.

In addition to the €79.3 million expenditure in 2018, there was further capital investment expenditure of €7.8m from the UK Research and Innovation's Biotechnology and Biological Sciences Research Council (UKRI-BBRSC), as part of the Large Facilities Capital Fund (LFCF).

Expenditure from funders



Allocation of funding in 2018



Organisation of EMBL-EBI leadership in 2018

Genomics

Ewan Birney
group

Anton Enright
group*

Moritz Gerstung
group

Nick Goldman
group

Zamin Iqbal
group

Proteins, Structures & Chemical Biology

Christoph Steinbeck
group*

Janet Thornton
group

Sarah Teichmann
group*

Pathways & Systems

Pedro Beltrao
group

John Marioni
group

Evangelia Petsalaki
group

Oliver Stegle
group

Virginie Uhlmann
group

Molecular Archives

Samples, Phenotypes & Ontologies
Helen Parkinson

European Nucleotide Archive
Guy Cochrane

EGA & Archive Infrastructure
Thomas Keane

Genes, Genomes & Variation

Vertebrate Genomics
Paul Flicek

Variation Annotation
Fiona Cunningham

Eukaryotic Annotation
Kevin Howe

Genomics Technology Infrastructure
Andy Yates

Genome Analysis
Daniel Zerbino

Training
Cath Brooksbank

External Relations
Lindsey Crosswell

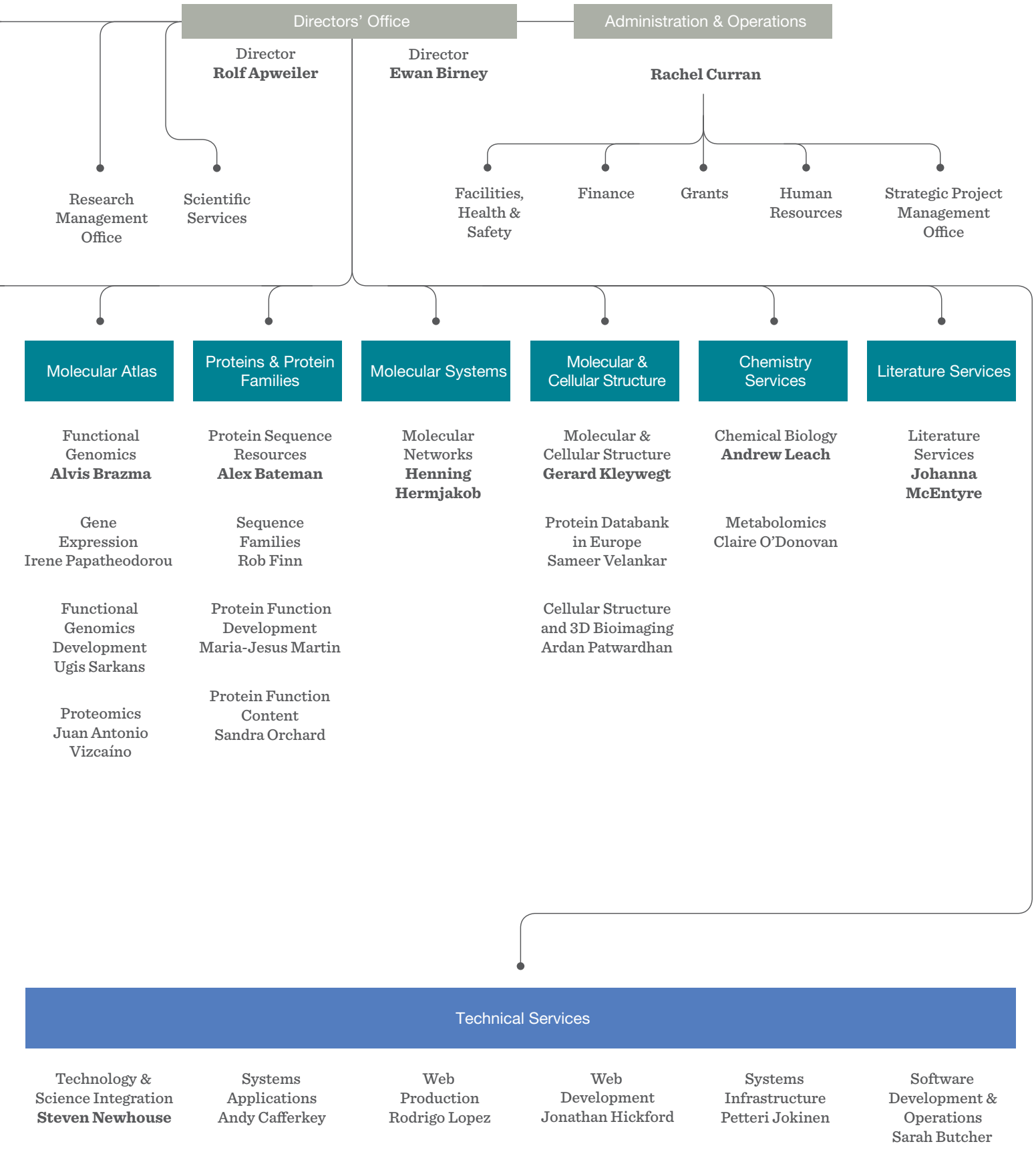
Industry Programme
Dominic Clark

Open Targets
Rolf Apweiler (interim)
(Ian Dunham in 2019)

RESEARCH GROUPS
* Visiting research group

SERVICE TEAMS

TECHNICAL SERVICES


Directors' Office
Director
Rolf Apweiler
Director
Ewan Birney
Administration & Operations
Rachel Curran
Research Management Office
Scientific Services
Facilities, Health & Safety
Finance
Grants
Human Resources
Strategic Project Management Office
Molecular Atlas
Functional Genomics
Alvis Brazma
Gene Expression
Irene Papatheodorou
Functional Genomics Development
Ugis Sarkans
Proteomics
Juan Antonio Vizcaíno
Proteins & Protein Families
Protein Sequence Resources
Alex Bateman
Sequence Families
Rob Finn
Protein Function Development
Maria-Jesus Martin
Protein Function Content
Sandra Orchard
Molecular Systems
Molecular Networks
Henning Hermjakob
Molecular & Cellular Structure
Molecular & Cellular Structure
Gerard Kleywegt
Protein Databank in Europe
Sameer Velankar
Cellular Structure and 3D Bioimaging
Ardan Patwardhan
Chemistry Services
Chemical Biology
Andrew Leach
Metabolomics
Claire O'Donovan
Literature Services
Literature Services
Johanna McEntyre
Technical Services
Technology & Science Integration
Steven Newhouse
Systems Applications
Andy Cafferkey
Web Production
Rodrigo Lopez
Web Development
Jonathan Hickford
Systems Infrastructure
Petteri Jokinen
Software Development & Operations
Sarah Butcher

Our governance

EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an inter-governmental organisation with over 20 member states and two associate member states. EMBL is led by a Director General (Iain Mattaj in 2018, Edith Heard in 2019), 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 2018, EMBL-EBI was led by joint Directors Rolf Apweiler and Ewan Birney, supported by two committees and 42 Group and Team Leaders (GTLs).

Executive Management Committee (EMC)

EMC is represented by members from the full range of EMBL-EBI activities, including technical services and administration. The committee handles key executive aspects of the institute's activities and consults closely with GTLs. Members in 2018 were:

Rolf Apweiler	Director
Ewan Birney	Director, Joint Head of Research
Alex Bateman	Head of Protein & Protein Family Services
Alvis Brazma	Head of Molecular Atlas Services
Rachel Curran	Head of Administration & Operations
Paul Flicek	Head of Genes, Genomes & Variation Services
Nick Goldman	Research Group Leader and Joint Head of Research
Johanna McEntyre	Head of Literature Services
Steven Newhouse	Head of Technical Services

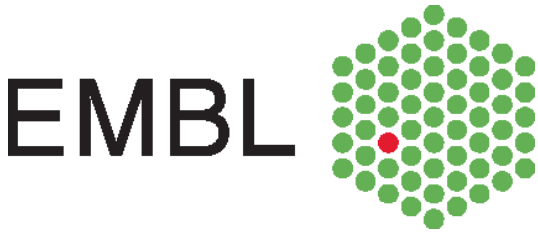
Strategy and Management Committee (SMC)

SMC deliberates on strategic decisions including budget, funding and the creation of new posts. Members in 2018 were (in alphabetical order):

Rolf Apweiler	Director
Ewan Birney	Director, Joint Head of Research
Alex Bateman	Head of Protein & Protein Family Services
Alvis Brazma	Head of Molecular Atlas Services
Cath Brooksbank	Head of Training
Lindsey Crosswell	Head of External Relations
Rachel Curran	Head of Administration & Operations
Paul Flicek	Head of Genes, Genomes & Variation Services
Nick Goldman	Research Group Leader and Joint Head of Research
Henning Hermjakob	Head of Molecular Systems Services
Gerard Kleywegt	Head of Molecular and Cellular Structure
Andrew Leach	Head of Chemical Biology Services
Johanna McEntyre	Head of Literature Services
Steven Newhouse	Head of Technical Services
Helen Parkinson	Head of Molecular Archive Resources

Our funders

We gratefully acknowledge our funders for their crucial support to our work in 2018.



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 and United Kingdom.

Alzheimer's Research UK

Australian National Health and Medical Research Council (NHMRC)

Biotechnology and Biological Sciences Research Council (BBSRC)

Bill & Melinda Gates Foundation

British Council

British Heart Foundation

Chan Zuckerberg Initiative

Cancer Research UK (CRUK)

European Commission

European Molecular Biology Organization (EMBO)

Foundation of the National Institutes of Health (FNIH)

Human Frontier Science Program (HFSP)

Innovate UK

Longitude Prize Discovery Awards

Gordon and Betty Moore Foundation

Medical Research Council (MRC)

National Institutes of Health (NIH)

Novo Nordisk

National Science Foundation (NSF)

UK Research and Innovation (UKRI)

Russian Foundation for Basic Research (RFBR)

Save the Tasmanian Devil Program Appeal

The Genetics Society

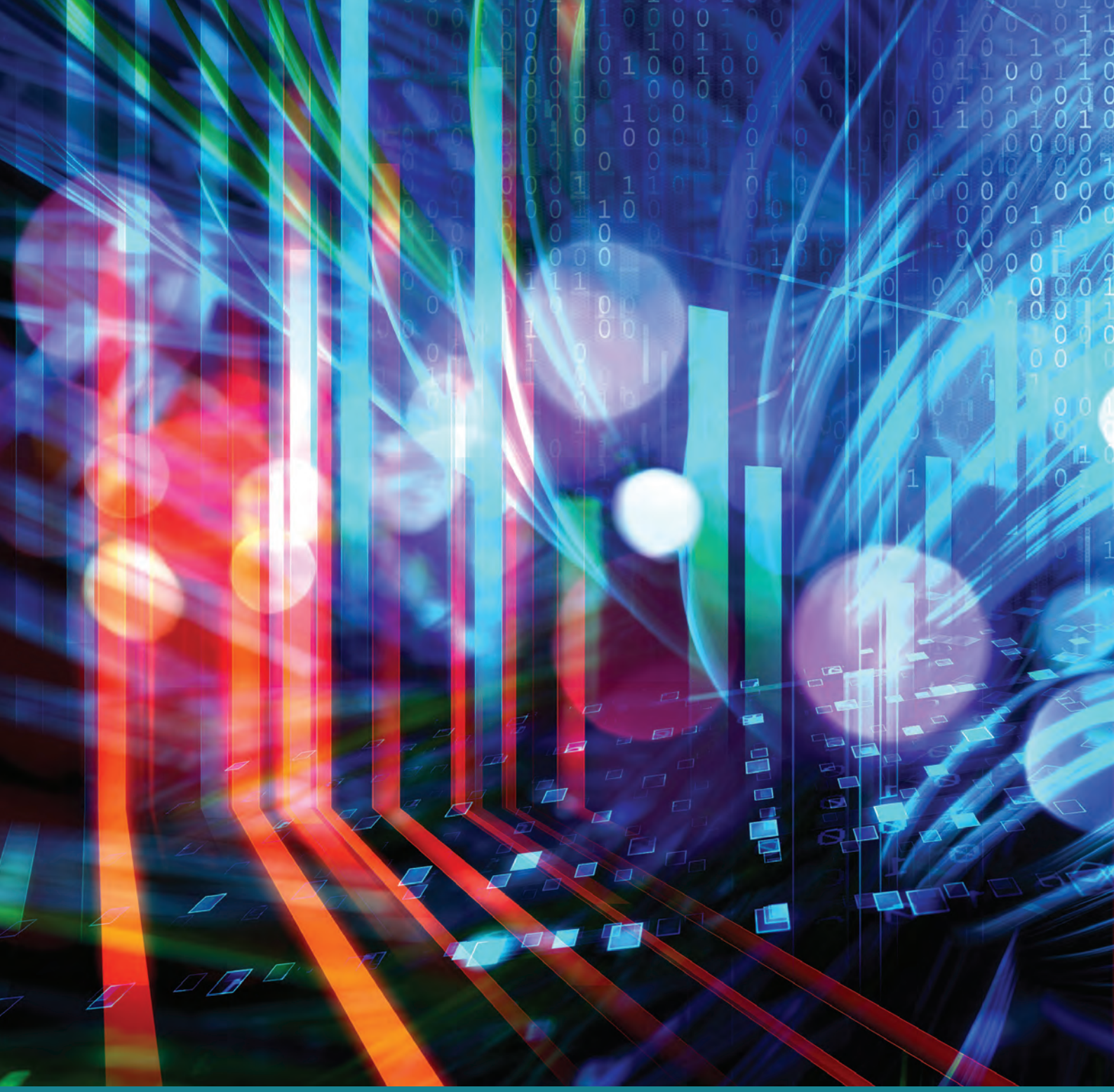
Weizmann UK

Wellcome Trust

List of acronyms

AI	Artificial Intelligence
AML	Acute Myeloid Leukaemia
API	Application Programming Interface
ASHG	American Society of Human Genetics
BBSRC	Biotechnology and Biological Sciences Research Council
CBE	Commander of the British Empire
CI/CD	Continuous Integration/ Continuous Delivery
CRG	Centre for Genomic Regulation
CZI	Chan Zuckerberg Initiative
DCP	Data Coordination Platform
DNA	Deoxyribonucleic Acid
DUO	Data Use Ontology
ECCB	European Conference on Computational Biology
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
EOSC	European Open Science Cloud
ERC	European Research Council
EVA	European Variation Archive
FAIR	Findable, Accessible, Interoperable and Reusable data
FTE	Full-Time Equivalent
GA4GH	Global Alliance for Genomics and Health
GB	Gigabytes
GBC	Global Biodata Coalition
GCRF	UK Government's Global Challenges Research Fund

GP	Genome Properties
GWAS	Genome-Wide Association Study
HCA	Human Cell Atlas
HDR UK	Health Data Research UK
HTLV-1	Human T-Lymphotropic Virus
IMPC	International Mouse Phenotyping Consortium
IP	Internet Protocol
ISMB	Intelligent System for Molecular Biology
IT	Information Technology
IWGSC	International Wheat Genome Sequencing Consortium
LMIC	Low-to-Middle Income Countries
PB	Petabytes
PDBe	Protein Data Bank in Europe
PDBe-KB	Protein Data Bank in Europe Knowledge Base
PDX	Patient Derived Xenograft
RNA	Ribonucleic Acid
TB	Terabytes
UCSC	University of California Santa Cruz
UX	User Experience
VCF	Variant Call Format
WGS	Whole Genome Sequencing



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 state: Estonia