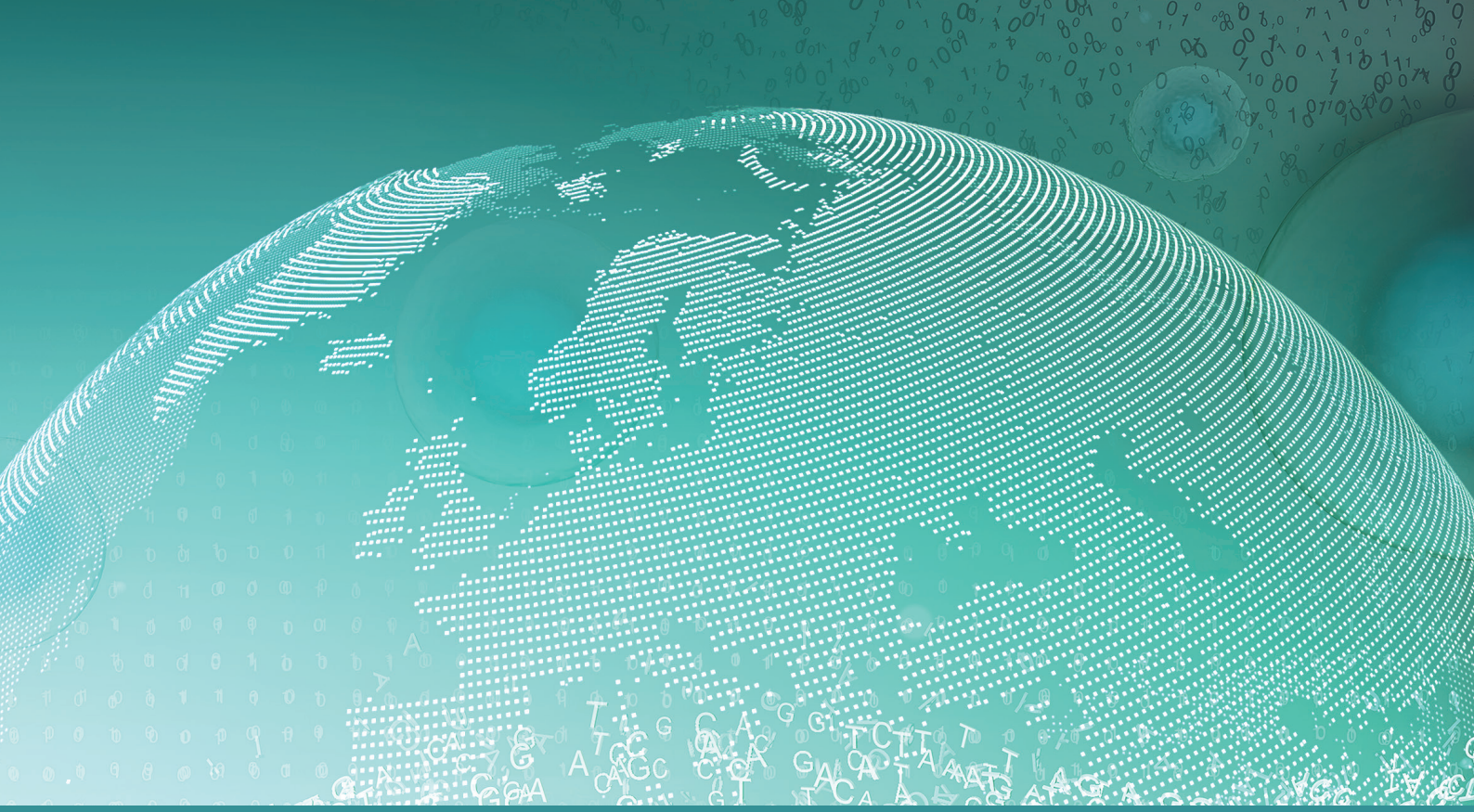


The European Bioinformatics Institute

Impact Report 2016

EMBL-EBI

On the cover:

The artwork in this publication, created by Spencer Phillips at EMBL-EBI, is inspired by the global, collaborative nature of the scientific community.

This publication was produced by the External Relations team at the European Bioinformatics Institute (EMBL-EBI). It is available online at www.ebi.ac.uk/about/our-impact

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

EMBL-EBI Impact Report 2016

Contents

Foreword	2
Our community in 2016	4
EMBL-EBI data resources	6
Data usage and growth	23
Research highlights	26
Major collaborations	32
Building skills and capacity	36
Industry engagement	42
Infrastructure	46
European coordination	48
Staff and alumni	50
Funding facts and figures	52
Looking ahead	54

Foreword

Science is global

Science is fundamentally international, and knowledge does not stop at borders. Scientists and engineers the world over share a passion for discovery, and seek knowledge wherever it may be found. At the European Bioinformatics Institute (EMBL-EBI), our mission is to fuel that passion through open data, connecting the global scientific community by providing data services and fostering collaborative research. We serve as a springboard for innovation by pooling knowledge and showing it clearly for anyone to investigate with new eyes.

This report highlights some of our contributions to science and technology in 2016, and shows the significance of our work to society.

Our member states

EMBL-EBI delivers high-impact discovery services that return value to our member states and to the global economy.* We are one of six sites of the European Molecular Biology Laboratory (EMBL), an international treaty organisation whose mission is to provide excellent research and scientific services across all aspects of molecular biology. Our site focuses on bioinformatics: data infrastructure, resource provision and blue skies research. In 2016, EMBL benefited from the participation and support of 22 member states and two associate member states. You can find out more about EMBL and its other sites in Europe at www.embl.org.

EMBL-EBI in the UK

EMBL is fortunate to have clear host-site agreements in place with the UK, Germany, France, Italy and Spain that guarantee our ability to recruit excellent scientists from around the world to deliver on our scientific missions. Because membership in EMBL is entirely separate from membership in the European Union (EU), the results of the UK's 2016 referendum have no direct impact on EMBL as a whole, or on EMBL-EBI in particular. Our success springs from collaboration, so we engage actively with policymakers to address the implications of 'Brexit' for the wider science, technology and engineering community.

**UK-based consultancy Charles Beagrie Ltd undertook an in-depth analysis of the value and impact of EMBL-EBI, which was published in February 2016. The report values the benefits to users and their funders at £1 billion per annum worldwide - equivalent to more than 20 times the direct operational cost of the institute. Executive summary: http://www.beagrie.com/EBI-impact-summary.pdf | Full report: http://www.beagrie.com/EBI-impact-report.pdf*

The hub of bioinformatics in Europe

This is an exciting time of change for the Wellcome Genome Campus, which has diversified to include EMBL-EBI, the Sanger Institute, ELIXIR, Genomics England and, at the time of this printing, eight companies in its new BioData Innovation Centre. This vibrant, international and collaborative community represents an intense concentration of expertise in genomics and bioinformatics, and is increasingly at the centre of transformative developments in the biomedical and life sciences.

Building capacity and solving problems

We make scientific enquiry easier by bringing people together, building capacity through training and knowledge exchange. We enrich scientific communities through data coordination, and by ensuring the hard-won results of scientific experiments can be integrated and re-examined by future generations.

Most importantly, we solve problems by creating new, innovative ways to analyse and compare complex, multi-layered information. This work is essential for understanding life at its most fundamental level, and translating knowledge into solutions for biomedical, agricultural and environmental science.

Sincerely,

Rolf Apweiler, Joint Director

Ewan Birney, Joint Director

Our community in 2016

We are proudly European, but our culture and community are profoundly international. In 2016 our 547 members of staff represented 64 nations, and a critical mass of expertise in bioinformatics.

Community facts and figures

- 547 members of staff (FTEs) from 64 nations, and 141 visitors from 34 nations who worked with us on specific projects
- 258 scientific publications co-authored with collaborators at 430 organisations in 45 countries
- 186 projects jointly funded with 624 institutes in 63 countries
- 23 member companies based in 8 countries in our Industry Programme, with two new members in 2016
- 17 countries and 8 companies represented on our scientific advisory boards
- Joined 21 ELIXIR members in ELIXIR-EXCELERATE, a major, EU-funded infrastructure project

EMBL member states in 2016

- Austria
- Belgium
- Croatia
- Czech Republic
- Denmark
- Finland
- France
- Germany
- Greece
- Iceland
- Ireland
- Israel
- Italy
- Luxembourg
- Malta
- Netherlands
- Norway
- Portugal
- Spain
- Sweden
- Switzerland
- United Kingdom
- Australia*
- Argentina*

**Associate member state*

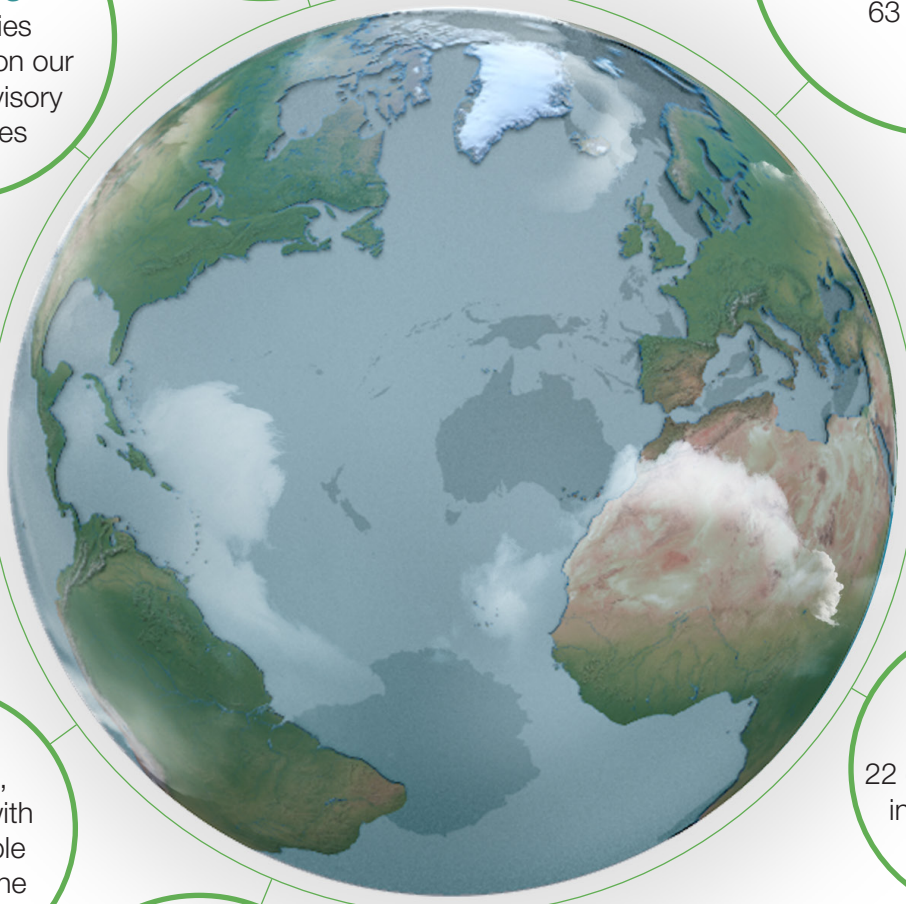
EMBL prospect member states in 2016

- Hungary
- Lithuania
- Poland
- Slovakia

Publications
258 papers
Co-author affiliations:
430 organisations
45 countries

Joint grant funding
186 projects
624 organisations
63 countries

Scientific Advisors
17 countries represented on our scientific advisory committees



Industry Programme
22 companies based in 8 countries are members

Training
350 events, connecting with 13 000 people throughout the world

Staff & Visitors
547 members of staff from 64 nations
141 visitors from 34 nations

Our impact is global, and our community is truly international. Many of our achievements are made possible by collaborations with science and technology professionals throughout the world.

EMBL-EBI data resources

Enriching life-science research

EMBL-EBI provides high-impact, public data resources that support discovery and return substantial value to the economy. The value and impact report by Charle Beagrie Ltd., published in 2016, estimates that EMBL-EBI data and services contributed to the wider realisation of future research impacts worth £920 million every year. The annual direct efficiency impact was estimated at between £1bn and £5bn per annum.

During 2016 we served an average of 27 million requests per day to our websites and ran 12.7 million computational 'jobs' for our users per month. Jobs represent both organic exploration of our data resources as well as high-volume automated queries. Each month these requests were from researchers at approximately 3.2 million unique hosts, representing both commercial and academic users.

What our users say

In our annual user survey we get feedback on both the extent to which we help researchers work and their attitudes toward EMBL-EBI services. In November 2016 we had 1191 respondents, who answered 28 questions about EMBL-EBI data services. Some highlights include:

"Essential to my research"

72.42%

"Useful in my daily work"

71.66%

"Informative, helpful and logical"

83.67%

"Reduced the time required to find relevant/required data"

90%

"Improved my efficiency in exploiting publicly available data"

90%

"Enabled research to go ahead that might otherwise not have"

79%

"Enabled me to undertake a greater quantity of research activities"

75%

ArrayExpress	
BioModels Database	
ChEBI	
ChEMBL	
Ensembl	
Ensembl Genomes	
International Nucleotide Sequence Databases	
Expression Atlas	
IMEx Consortium	
InterPro	
Protein Data Bank in Europe	
PRIDE	
RNACentral	
PubMed Central International	
Reactome	
UniProt: The Unified Protein Resource	

***Major database collaborations**: Our data resource teams collaborate with organisations throughout the world, and work with each other at EMBL-EBI, to fully integrate public molecular data. Here we show a selection of collaborations for our major databases, which illustrates how a global network makes open data possible.*

Expanding our offerings

New data types

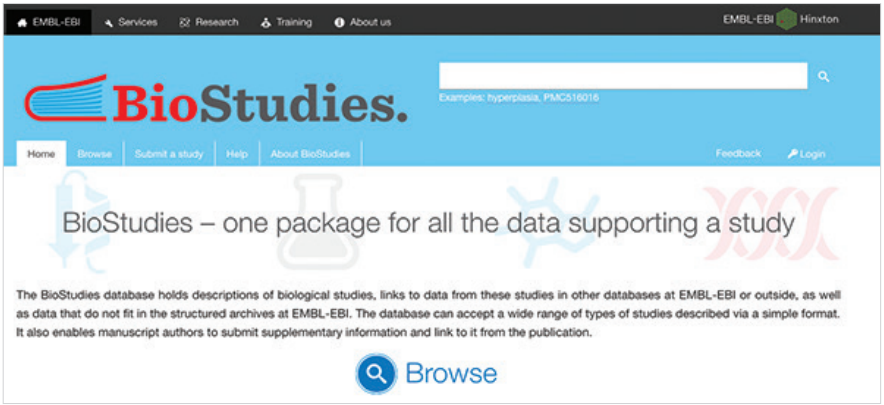
A home for unconventional data: BioStudies is our new repository for data and materials that do not fit in the 'traditional' structured databases. It gives researchers a new way to submit and share unconventional datasets—even if they are not associated with a paper. In 2016 BioStudies launched its easy-to-use, web-based data-submission tool, officially opening its doors to researchers everywhere.

BioStudies was inspired by diXa,* a project that combined heterogenous data to develop tools for predicting toxicity for human liver and heart tissue. This work continues in the context of the HeCaToS project, in which we manage toxicogenomic data.

BioStudies captures data from the EU-ToxRisk project,** and is complemented by EMPIAR and the incipient Imaging Data Repository (a collaboration with the University of Dundee funded by the BBSRC and the EU through Euro-BioImaging).

> *"EMBL-EBI resources are essential to all modern life-sciences research. We simply could not function without the core, reliably maintained data collections and the world-leading expertise that is centred on EBI for the organisation and analysis of biological data."*
>
> *– User comment in the 2016 economic impact survey by consultancy Charles and Neil Beagrie*

**diXa: the Data Infrastructure for Chemical Safety, and HeCaToS, the Hepatic and Cardiac Toxicity Systems modelling project, are funded by the EU's Seventh Framework Programme (FP7).*
*** EU-ToxRisk, a project focused on mechanism-based toxicity testing and risk assessment, is funded by the EU's Horizon2020 Programme.*



Sharper image: EMPIAR, the Electron Microscopy Public Image Archive, became firmly established in 2016, expanding to include raw data from 3D Scanning Electron Microscopy (3DSEM) and soft X-ray tomography image data. EMPIAR data is used in the on-going EMDataBank Map Validation Challenge, and provides the data needed for methods development and training in this emerging area of science.

EMPIAR: Raw EM data for validation, methods development and data mining

- 12.4 terabyte dataset: β-galactosidase structure (resolution: 2.2Å)
- First datasets from serial block face Scanning Electron Microscopy (SEM)
- First dataset from focused ion beam SEM
- First datasets from Volta phase-plate experiments

Image Data Repository: prototype

- Collaboration with the University of Dundee and the University of Cambridge, enabled the development of a scalable repository for imaging data, to be launched in 2017

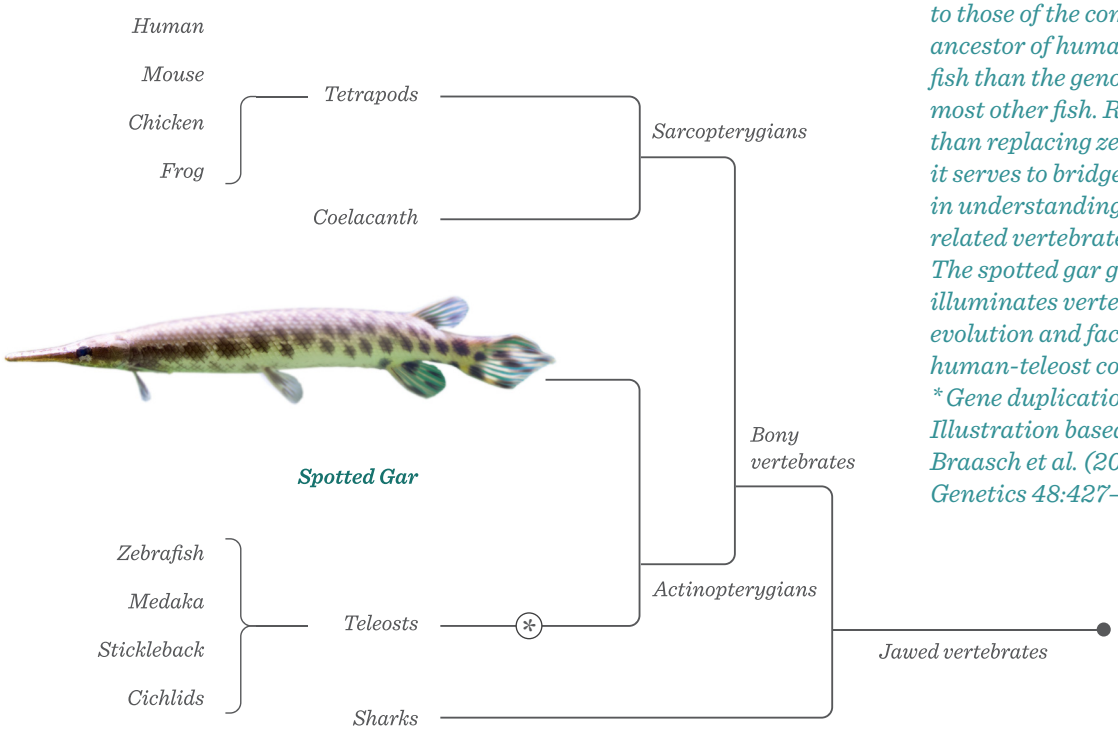
Single Cell Expression Atlas: To capture the increasing amount of single-cell transcriptomic data being generated and shared by the scientific community, we prototyped a new Single Cell Atlas resource based solely on data from single-cell RNA sequencing and featuring advanced 3D-visualisation techniques.

Gene expression regulation: As a first step in understanding tissue-specific regulatory networks, the Ensembl team performed a comprehensive analysis of the promoter capture Hi-C method. Summary statistics from this analysis, plus 9 billion gene-to-SNP Genotype–Tissue expression (GTEx)–eQTL correlations, are now available in Ensembl.

Order out of disorder: InterPro, our integrating service for protein sequences, families and motifs, welcomed three new member databases in 2016. Two are traditional protein families databases (the Conserved Domains Database and the Structure Function Linkage Database) and one is for annotating intrinsically disordered regions (MobiDB).

InterPro now calculates and displays a fine level of detail about proteins, including annotations at the 'per residue' level and information about intrinsic protein disorder. InterProScan now calculates information on intrinsically disordered regions (IDRs), and the MobiDB-Lite tool predicts IDRs in UniProt sequences.

Evolutionary divergence: The teleost genome duplication was an event that shaped the evolutionary history of teleost fishes, a vertebrate clade that includes model organisms such as zebrafish and medaka. The spotted gar is a member of the ray-finned fish lineage, which diverged before this gene duplication event, making it extremely interesting for evolutionary studies. The spotted gar acts as a bridge in applying knowledge about zebrafish to humans. We collaborated with spotted gar researchers to create a resource for connecting human biology to fish biomedical models. We also worked with flycatcher researchers to compare collared and pied flycatchers to study gene expression evolution as species diverge (Braasch et al., *Nature Genetics*).



*The genome of the spotted gar is much more similar to those of the common ancestor of humans and fish than the genomes of most other fish. Rather than replacing zebrafish, it serves to bridge the gap in understanding these related vertebrate genomes. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. * Gene duplication event. Illustration based on Braasch et al. (2016) Nature Genetics 48:427–437.*

Reference genomes and proteomes

Mouse and Human genomes: Data-driven research relies on accurate reference data. Ensembl holds reference gene sequence and annotation for human, mouse and almost 100 other vertebrate species, including key model organisms and farmed animals. In 2016 several of these references were improved and updated by incorporating manual annotation, using long-range sequences to support alternative splicing, and adding long, intergenic noncoding RNA (lincRNA) data. The mouse reference gene resources grew significantly with the release of sequence and annotation data from 16 inbred mouse strains (Yates et al., *Nucleic Acids Research*).

Reference proteomes: To help users navigate the growing number of complete proteome sequences effectively, UniProt defined a set of reference proteomes that serve as landmarks in proteome space. The selected proteomes provide broad coverage of the tree of life, and represent a cross-section of the diverse data in the UniProt Knowledge Base. They include well-studied model organisms and other species of interest to biomedical and biotechnological research (Breuza et al., *Database*).

Epigenomes: The Ensembl Regulatory Build grew rapidly in 2016, with 68 annotated human epigenomes from data generated by the ENCODE, Roadmap Epigenomics and Blueprint projects.

Extensive automation and high-throughput algorithms such as WiggleTools make for an efficient, high-throughput analysis pipeline. The EpiRR registry for epigenomics metadata now contains references recorded by epigenomics researchers for 6109 experiments across 1934 epigenomes, and serves as a basis for the International Human Epigenome Consortium (IHEC) data portal.

Wheat: Owing to its size and complexity, the bread wheat genome has been a major, long-standing challenge in genome assembly. In 2016 we significantly improved this assembly in Ensembl Plants, based on the results of a collaborative BBSRC-funded project with the Earlham Institute.

Parasitic Worms: WormBase ParaSite, the partner resource to WormBase, grew to include data for 134 genomes of parasitic nematodes and flatworms (helminths), with new visualisation tools deployed to indicate the completeness of each genome assembly (Howe et al., Molecular and Biochemical Parasitology; Howe et al., *Nucleic Acids Research*).

Fungal Pathogens: We rely on the expertise of the user community to continually improve our genome annotation. In 2016 we supported the community in completely revising the annotation of the pathogenic fungus *Botritus cinera* (literally, 'grapes like ashes'). Botritus cinera infection can either cause wine grapes to rot, or result in distinctively sweet wines. Ensembl Genomes helps researchers understand the underlying genetics of this fungal pathogen.

Depending on conditions, grapes infected with this fungus could either be lost, or make for a delicious dessert wine. Image of Botritus cinera *from Wikimedia Commons. Photographer: Tom Maack*

Data growth

Understanding environments: EBI Metagenomics, which grew 11-fold in 2016, has become a community standard, with international policymakers and others consulting on how to maximise the potential of this emerging field. We operate a functionally rich portal encompassing metagenomics data archiving, standards compliance, functional and taxonomic analysis, data analysis, exploration and interpretation.

Samples in the spotlight: BioSamples links diverse data from over 5 million biological samples, some of which have been used in many different experiments over an extended period of time. The resource now offers sample information from the European Bank for induced pluripotent Stem Cells (EBiSC), the Functional Annotation of ANimal Genomes (FAANG) project and the Human Induced Pluripotent Stem Cell Initiative (HipSci), among others.*

Genome-Wide Association Studies Catalog: In 2016 our staff manually curated a total of 454 studies and 6766 SNP–trait associations from 273 publications. To offer better, more structured information about the samples in these studies, we also made 'ancestry' and 'country of recruitment' data available (MacArthur et al., *Nucleic Acids Research*).

**EBiSC is an Innovative Medicines Initiative addressing the demand for quality-controlled, disease-relevant, research-grade iPSC lines, data and cell services. FAANG, supported by the BBSRC, is delivering standardised phenotype and genotype datasets from animal species with reference-quality genome assemblies. The Wellcome- and MRC-funded HipSci project created a global iPSC resource and used it to carry out cellular genetic studies.*

Expression: Our gene expression resources now offer thousands of experiments and assays from 30 organisms, with the RNA-seq studies in our Baseline Expression Atlas including data from both large-scale, high-impact studies (e.g. GTEx, BLUEPRINT) and smaller-scale proteomics experiments.

Pathogens: The number of bacterial genomes available through Ensembl Bacteria increased to nearly 40 000 and number of fungal protist genomes to over 800.

Expression: We curated over 3000 transcriptomics experiments and nearly 125 000 assays. These assays included nearly 500 RNA-seq experiments, over 8000 differential comparisons across 30 organisms, and nearly 700 plant experiments.

Non-coding RNA: The RNAcentral database expanded to include an additional one million new non-coding RNA sequences, all available through a lightweight genome browser. We built 135 new functional non-coding RNA families using a new, literature-driven curation workflow in Rfam (RNAcentral Consortium, *Nucleic Acids Research)*.

Proteomics: Our PRIDE repository supports the reuse and reanalysis of proteomics data. In 2016 researchers submitted close to 2000 datasets to and downloaded approximately 241 terabytes of data from the archive. Our links with the Japanese proteomics community strengthened as we welcomed jPOST to the ProteomeXchange Consortium, of which PRIDE is a founding member (Vizcaíno et al., *Nucleic Acids Research*).

25 000 protein structures: The Protein Data Bank in Europe (PDBe) annotated their 25 000th structure in 2016. Our annotation work has grown rapidly over the past few years as PDBe has handled data submissions from all over Europe and Africa. Thanks to new features in the PDBe website, scientists can now effortlessly view X-ray and Electron Microscopy maps in conjunction with the atomic models from the archive.

Systems Biology models: We now offer over 1600 literature-based models through BioModels, including a comprehensive set of curated neurodegeneration models as part of the EU-funded AgedBrainSysBio project (Lloret-Villas et al., *CPT Pharmacometrics Systems Pharmacology*).

Target discovery: Our **ChEMBL** database of bioactive compounds now offers annotated data extracted from the literature, deposited by neglected-tropical-disease researchers and, through a data-exchange collaboration, extracted from patents thanks to the BindingDB Group at the University of California, San Diego. This adds valuable bioactivity data on biological targets early, before publication in medicinal chemistry literature.

Patented chemicals: At the end of 2016 the number of novel, annotated chemical entities in SureChEMBL was approximately 17.6 million. The resource is growing at a rate of around 80 000 novel chemicals per month curated from roughly 50 000 new patents.

InterPro: We added 1238 new InterPro entries in 2016. We have also continued to regularly release InterPro data, issuing six public releases in 2016, along with 11 internal data releases to UniProt.

Metabolism: Our MetaboLights resource supports experimental data from metabolomics experiments, and in 2016 processed mass spectrometry (MS) imaging studies in collaboration with the EMBL Heidelberg Core Metabolomics Facility.

New data types in structural biology

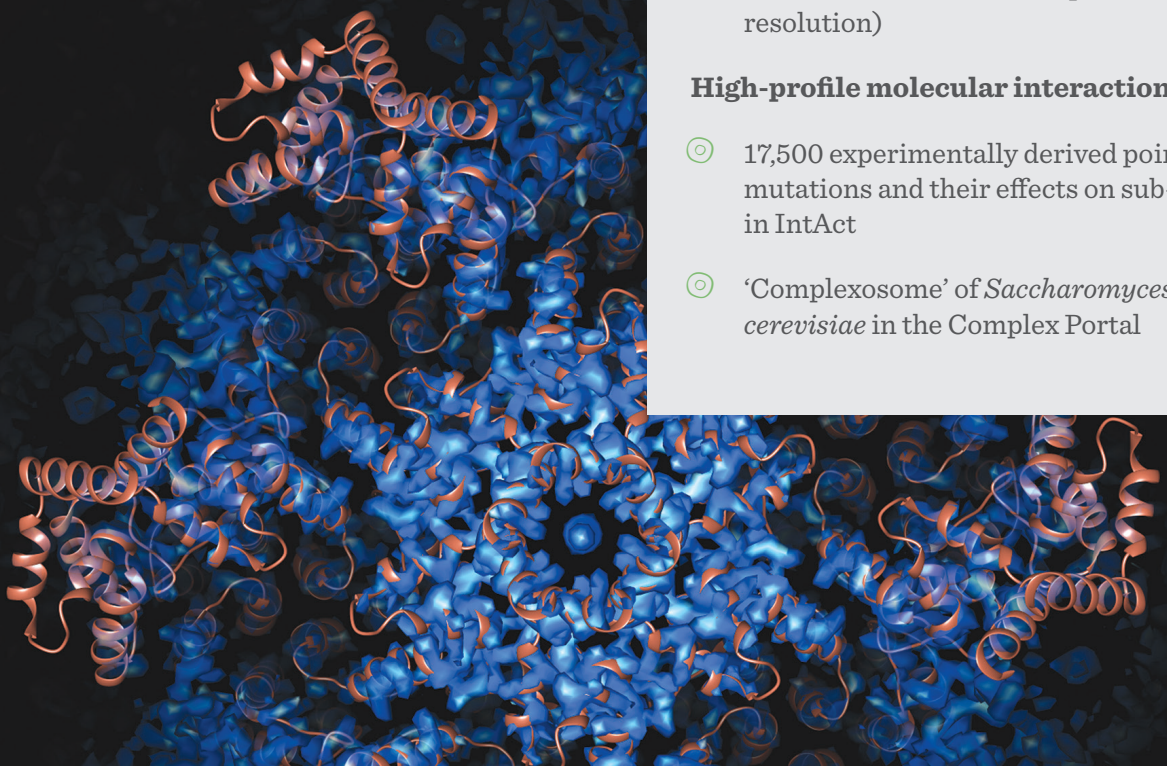
- ~20% of the 1000 new entries in the Electron Microscopy Data Bank (EMDB) were from tomography and sub-tomogram averaging experiments
- 40% of these 1000 entries were at reported resolutions better than 6Å (1 Å is about one millionth the width of a human hair)

High-profile datasets curated in Electron Microscopy Data Bank:

- Zika virus: Single-particle cryo-EM Zika virus structure (3.8Å resolution)
- HIV: structure of HIV-1 capsid-SP1 (3.9Å resolution)

High-profile molecular interaction datasets

- 17,500 experimentally derived point mutations and their effects on sub-networks in IntAct
- 'Complexosome' of *Saccharomyces cerevisiae* in the Complex Portal



Problem solving: new software and tools

Cool tools: data analytics

Variant Effect Predictor (VEP): The VEP is a powerful tool for predicting if the effect of a sequence variant is damaging or beneficial (McLaren et al., *Genome Biology*). For example, a clinician can use it to explore a patient's genome and understand possible disease risks. The VEP is built on Ensembl's comprehensive genome annotations, and can now annotate an entire human genome (~4 million variants) in under an hour and an exome in under 5 minutes (~200,000 variants).

Viewing variation data: Ensembl builds and updates variation databases for 22 vertebrate species, totalling over 540 million variants. Variation frequency data from many sources (e.g. ExAC, Exome Sequencing Project, UniProt, dbSNP) can be searched by ontology term, synonym or accession number simultaneously, as each resource has now been mapped to an ontology of phenotypes, traits and disease descriptions. Finding such connections across related disease terms can potentially reveal important pathogenic associations.

Genome Annotation search tool: Ensembl's new Advanced Search tool, which is intended to eventually replace BioMart, has been built using new, scalable technologies. This allows comprehensive, fast queries of genomic annotation of over 80 million gene models and associated data from approximately 20 000 species. The Advanced Search retrieves variation and expression data from the European Variation Archive and Expression Atlas. Built to support the EBiSC project (see page 12), its flexibility makes it potentially useful in a variety of contexts.

Ontology Lookup Service: The OLS, a repository for biomedical ontologies, provides a single interface for searching across numerous ontologies. It includes the Human Phenotype Ontology (HPO) and the Experimental Factor Ontology (EFO), among many others, and allows users to map a term to different ontologies using the new Zooma tool.

Focusing on the right genes: Ensembl offers a new method that identifies high-confidence gene-orthologue pairs based on sequence similarity and local gene-order conservation. This will help human disease researchers identify the related genes in model species as well as understand evolutionary relationships.

Taxonomy tool: A new component of EBI Metagenomics allows users to perform eukaryotic analyses based on 18S rRNA, a subunit of the ribosome and one of the basic components of all eukaryotic cells (Mitchell et al., *Nucleic Acids Research*).

Discovery proteomics: The **PRIDE** Cluster resource brings together the millions of peptide mass spectra in the PRIDE Archive database.

Using a newly developed spectrum-clustering algorithm, members of our Proteomics team were able to detect millions of unidentified spectra across different datasets. The tool enables the discovery of previously unknown post-translational modifications or peptide variants (Griss et al., *Nature Methods*).

Text mining for key concepts: EuropePMC's SciLite platform, launched in 2016, brings the literature closer to the underlying biological data in a very transparent way. It overlays text-mined annotations on research articles, helping users find key concepts more easily and providing links to related resources or tools.

On-the-fly gene set overlap analysis: Anyone who is interested in finding out how a particular set of genes overlaps with over 7500 differential comparisons in Expression Atlas can use a new tool to perform this analysis in seconds, for up to 100 Ensembl gene identifiers (Petryszak et al., *Nucleic Acids Research*).

HMMER3: The upgraded HMMER3 algorithm offers BLAST-speed detection of distantly related proteins. It is accessible through our newly expanded HMMER website infrastructure.

UniChem web services: Our updated chemogenomics tool now links information on approximately 130 million chemical structures from 29 source databases.

Cheminformatics cocktail: New chemistry web services, based on RDkit, allow users to perform more complex queries and combine data in ChEMBL. A SOLR-based search supports those accessing ChEMBL and UniChem programmatically through APIs (Mutowo et al., *Journal of Biomedical Semantics*).

Insecticide resistance: When researchers describe their datasets well, the results of their research can be explored in new ways well into the future. Thanks to VectorBase's rich sample metadata about the vectors of human pathogens, we were able to deploy a search facility that links different data types, providing access to population-wide assay and insecticide resistance data.

Integration toolkit: A new toolkit from PDBe integrates cellular structure, molecular structure and other forms of bioinformatics information, translating between existing segmentation file formats and EMDB-SFF, the new format that supports structured biological annotations. The toolkit complements our extended SIFTS pipeline, which maps UniProt and PDB information.

Storing more using less: To prepare for future data growth, we continued to refine and implement CRAM, an innovative method for data compression. ENA's comprehensive service provides reference coordinate-based indices for CRAM data.

FAIR exchange

Enhancing data discoverability: Released in 2016, the Omics Discovery Index (OmicsDI) makes it easier to discover, access and reuse open- and controlled-access datasets covering genomics, transcriptomics, proteomics and metabolomics. It uses shared identifiers and rich metadata to highlight groups of related datasets in 11 repositories hosted by six different organisations (Perez-Riverol et al., *Nature Biotechnology*).

Omics DI enables FAIR access to multi-omics datasets.

Metagenomics Exchange: In collaboration with the MG RAST service in the United States, we launched the Metagenomics Exchange to reduce duplication and stimulate data submission globally.

Flexible format for model exchange: Researchers working in different companies face practical barriers to collaboration on pharmacometric modelling. To address this problem EMBL-EBI developed PharmML, a flexible format for exchanging computational models in pharmaceutical R&D. It is a key component of the IMI-funded DDMoRe model repository, which helps researchers collaborate on models to improve the design of cost-effective, reliable clinical trials of new and repurposed drugs (Swat et al., *CPT Pharmacometrics and Systems Pharmacology*).

Plant phenotyping experiments: Plant phenotypic data enshrouds a wealth of information which, when accurately analysed and linked to other data types, brings to light new knowledge about the mechanisms of life. Yet, the lack of common standards to describe phenotypic data has hampered data exchange and reuse. To address this, our Ensembl Plants team put forward a minimal set of guidelines for plant phenotypic experiments, specifying both the content and format of the description (Krajewski et al., *Journal of Experimental Botany*).

Cellular structure data file format: Together with community experts, we have been developing a file format for representing segmentations and the transformations between sub-tomogram averages and tomograms that also supports structured biological annotations (EMDB-SFF). This is essential to enable the integration of cellular structure data with other forms of bioinformatics information as well as with molecular structure data.

NMR data: MetaboLights includes studies in an open data standards format for Nuclear Magnetic

Resonance data, developed in the COSMOS project. The format is intended to be fully compatible with existing NMR data collected for chemical, biochemical or metabolomics analysis as well as organic synthetic experiments (Kale et al., *Current Protocols in Bioinformatics*).

Show me the data

A new view on proteins: UniProt launched the first protein sequence viewer to integrate all publicly available information, helping researchers visualise and interpret large volumes of biological data.

TrackHub Registry: This new service allows any researcher to share their genomic datasets, making them discoverable by anyone and ready to visualize as 'tracks' in Ensembl.

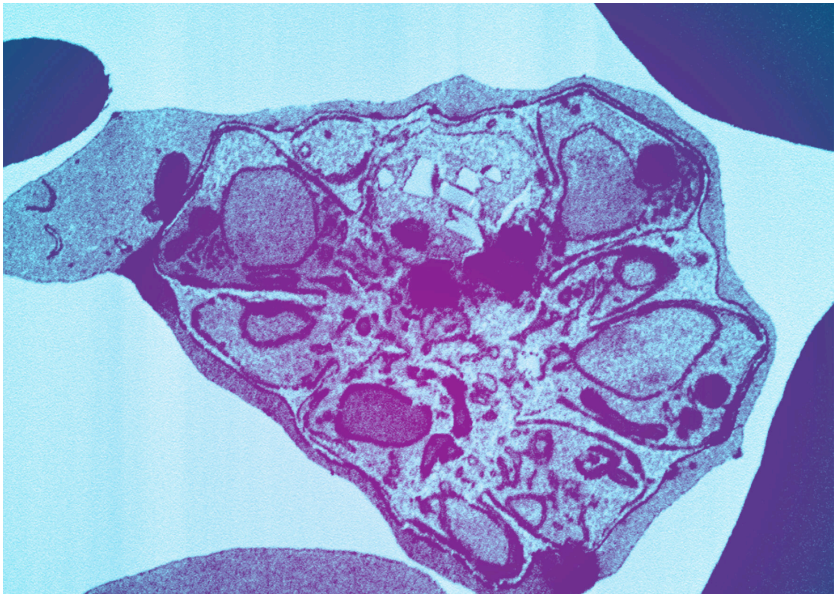
Plant and fungi expression data: A new pipeline automatically aligns public RNA-sequence data against the genome sequence. Users can now visualise expression data from over 1000 distinct experiments within Ensembl Plants and Ensembl Fungi (Bolser et al., *Methods in Molecular Biology*).

Haplotype diversity view: The Variation Annotation team team created a haplotype view per transcript using phased genotype data from the 1000 Genomes Project to view haplotype frequencies by population.

PredComp: We now offer a comprehensive, graphical report comparing predicted annotations against automated ones in UniProt-TrEMBL.

Interactive pathway diagrams: Visualising, navigating and reusing biological pathway data is now easier thanks to Reactome's new, high-quality overview diagrams and download functionality. Scientists can reuse and rearrange Reactome diagrams for use in publications and presentations, or even integrate the JavaScript-based widgets into third-party applications (Fabregat et al., *Nucleic Acids Research*).

Volume slicer: An enhanced visualisation tool for structural data in EMDB and EMPIAR displays 3D reconstructions of EM experiments as 2D 'slices'. Users can explore these slices from three different angles and navigate through the volume,



Serial block face SEM of a malaria parasite infected red blood cell. Image by Sakaguchi et al. (2016); DOI: 10.1016/j.jsb.2016.01.003. Available via EMPIAR.

without installing software or downloading large files (Salavert-Torres et al., *Journal of Structural Biology*).

LiteMol: PDBe offers a new, lightweight, interactive 3D viewer that displays experimental results – electron density maps for X-ray structures and electric potential maps for EM structures – alongside the coordinate model. Developed in collaboration with Masaryk University in the Czech Republic, the viewer displays annotations such as structure quality and sequence/structure domains in context.

GO, Metagenomics! Comparing sample data in EBI Metagenomics is easier than ever, thanks to a new, specialised Gene Ontology 'slim' that creates a high-level visualisation of your sample's functional profile.

Better, faster, stronger service delivery

Protein data access: There is a greater need than ever to integrate large-scale biological data with known annotations and disseminate this information to an increasingly diverse research community. To address this need, we introduced a new API that provides access to key biological data from UniProt and Large Scale Studies data mapped to UniProt. The Protein REST API service also serves as a bridge between genomic and protein data, enabling users to retrieve genome coordinates for protein sequences.

Expression data submission: Annotare, the main data-submission tool for ArrayExpress, has greatly improved file upload functionality, integration with the customer-support system, and guidance for ArrayExpress users.

OneDep: Released in 2016, a new, integrated software system developed by the wwPDB collaboration provides a single portal for depositing structural data to the PDB, BMRB and EMDB archives, providing the user community with a more robust deposition and annotation service. OneDep captures rich metadata and streamlines the deposition and annotation of structural data from X-ray, neutron and electron diffraction, NMR and EM experiments (Young et al., *Structure*).

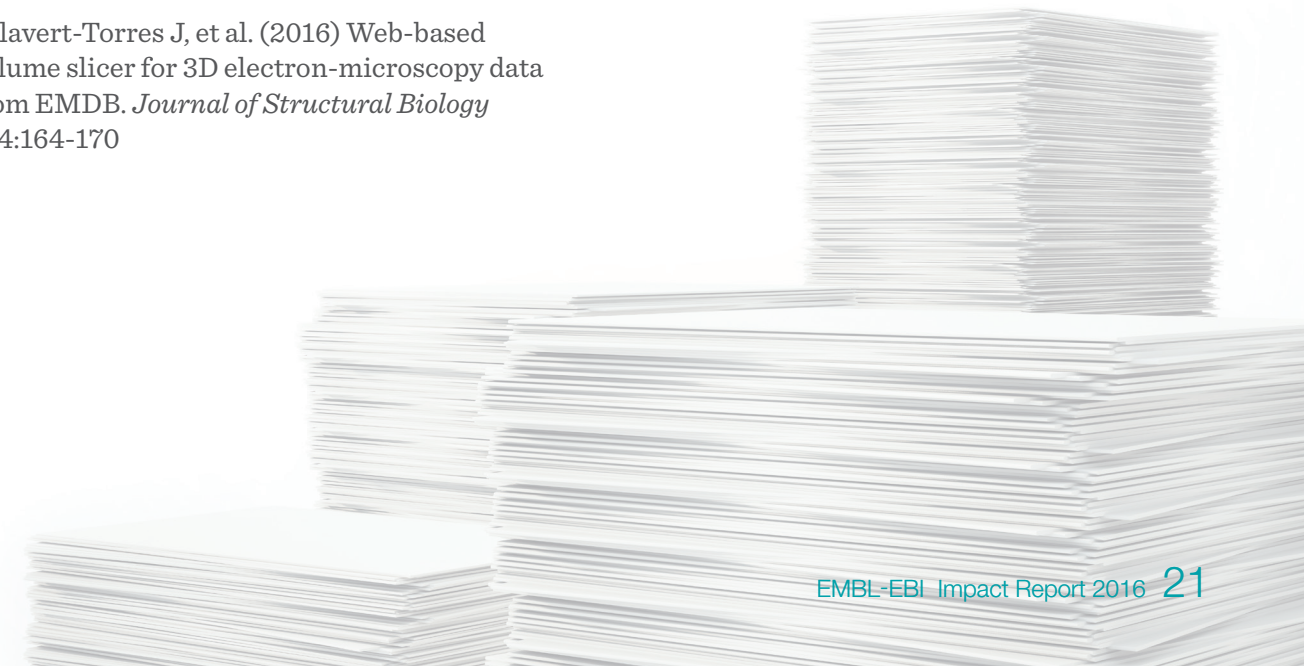
Benchmarking data analysis software: A new software tool, LFQbench, helps researchers assess mass-spectrometry-based, label-free quantitative proteomics in a comprehensive, reproducible manner. 'LFQbench' benchmarks different types of instruments, acquisition methods and data-analysis software (Navarro et al., *Nature Biotechnology*).

Metabolomics tools: A service catalogue for the PhenoMeNal e-infrastructure, coordinated by EMBL-EBI, offers 28 applications to support the data-processing and analysis pipeline for molecular phenotype data generated using metabolomics applications.

Selected publications

- Bolser D, et al. (2016) Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods in Molecular Biology* 1374:115-140
- Braasch A, et al. (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* 48:427–437
- Breuza L, et al. (2016) The UniProtKB guide to the human proteome. *Database* 2016; DOI: 10.1093/database/bav120
- Fabregat A, et al. (2016) The Reactome pathway knowledgebase. *Nucleic Acids Research* 44:D481-D487
- Griss J, et al. (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods* 13:651–656
- Howe KL, et al. (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Research* 44:D774-D780
- Howe KL, et al. (2016) WormBase ParaSite - a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology*
- Kale NS, et al. (2016) MetaboLights: an open-access database repository for metabolomics data. *Current Protocols in Bioinformatics* 53:14.13.1-18
- Krajewski P, et al. (2015) Towards recommendations for metadata and data handling in plant phenotyping. *Journal of Experimental Botany* 66:5417-5427
- Lloret-Villas A, et al. (2017) The impact of mathematical modeling in understanding the mechanisms underlying neurodegeneration: evolving dimensions and future directions. *CPT Pharmacometrics Syst Pharmacol* 6:73-86.
- MacArthur J, et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45:D896-D901; published online November 2016.
- McLaren W, et al. (2016) The Ensembl Variant Effect Predictor. *Genome Biology* 17:122
- Mitchell A, et al. (2016) EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research* 44:D595-D603
- Mutowo P, et al. (2016) A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. *Journal of Biomedical Semantics* 7:59

- Navarro P, et al. (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology* 34:1130-1136
- Perez-Riverol Y, et al. (2017) Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology* 35:406-409
- Petryszak R, et al. (2016) Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research* 44:D746-D752
- RNAcentral Consortium (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research* 45:D128-D134; published online October 2016.
- Salavert-Torres J, et al. (2016) Web-based volume slicer for 3D electron-microscopy data from EMDB. *Journal of Structural Biology* 194:164-170
- Swat MJ, et al. (2015) Pharmacometrics Markup Language (PharmML): Opening New Perspectives for Model Exchange in Drug Development. *CPT: Pharmacometrics & Systems Pharmacology* 4:316-319
- Vizcaíno JA, et al. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* 44:D447-D456
- Yates A, et al. (2016) Ensembl 2016. *Nucleic acids research* 44:D710-D716
- Young JY, et al. (2017) OneDep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* 25:536-545



What our users say

"The resources provided by EMBL-EBI are critical to biology and are typically very high quality."

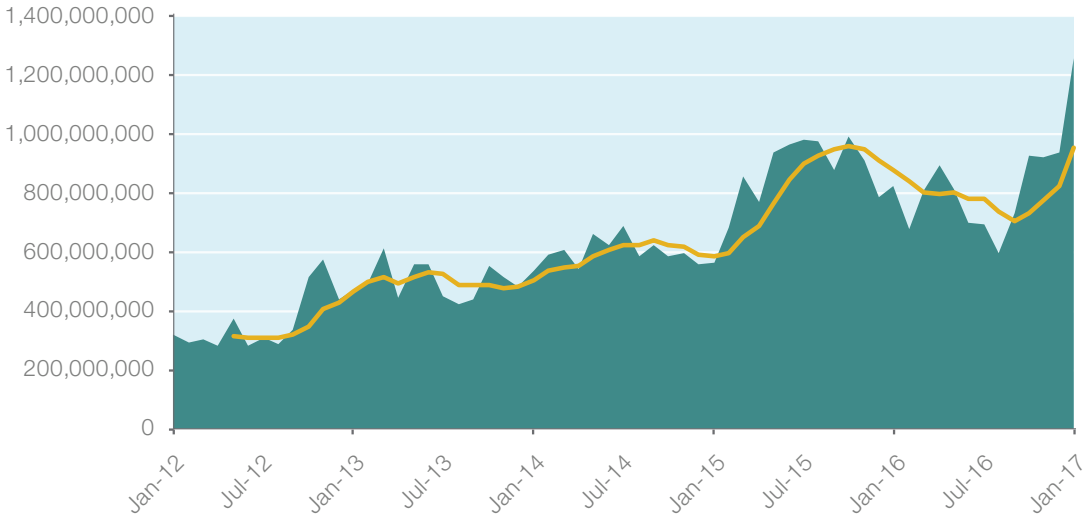
"EMBL-EBI services are the backbone of most if not all computational biology projects. It's an invaluable resource that many researchers don't often think about. Countless applications and software frameworks rely on databases and back-end web services provided by the EMBL-EBI."

"My work entirely depends on services offered by EMBL-EBI, without which I would be a painter with both hands tied behind!"

"EMBL-EBI services are essential to micro and small biotech enterprises who do not have the financial resources to afford expensive software and database access for their knowledge base and operations."

"Although I might only use certain EMBL-EBI resources a few times a year, many are essential resources to have, containing data that simply could not be collated by an individual lab or researcher."

Data usage in 2016

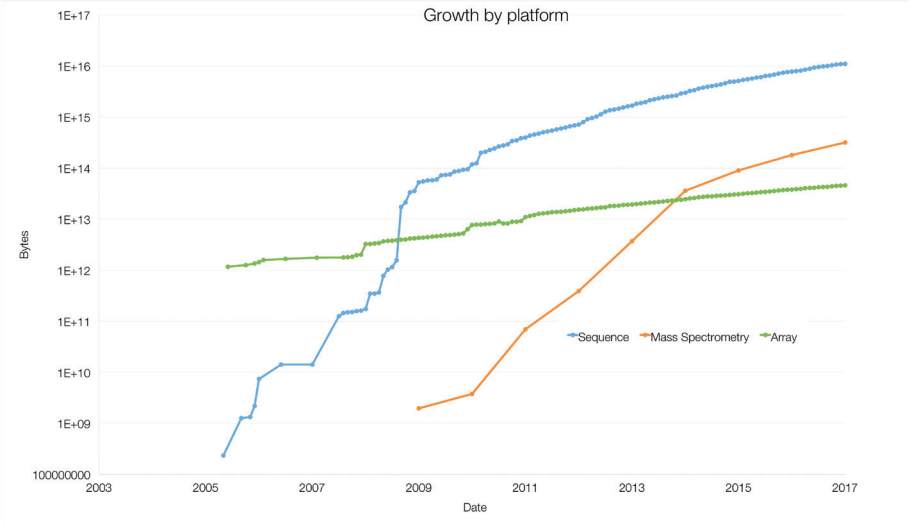


EMBL-EBI web requests per month, 2012-2016.

- Average requests per day to EMBL-EBI websites: 27 million
- Total volume of data downloaded: 3.5 petabytes
- Unique IP addresses accessing EMBL-EBI websites every month: 3.2 million
- Automated usage: 82% (56% via RESTful APIs, 26% via SOAP web services)
- Usage via web interfaces: 18%

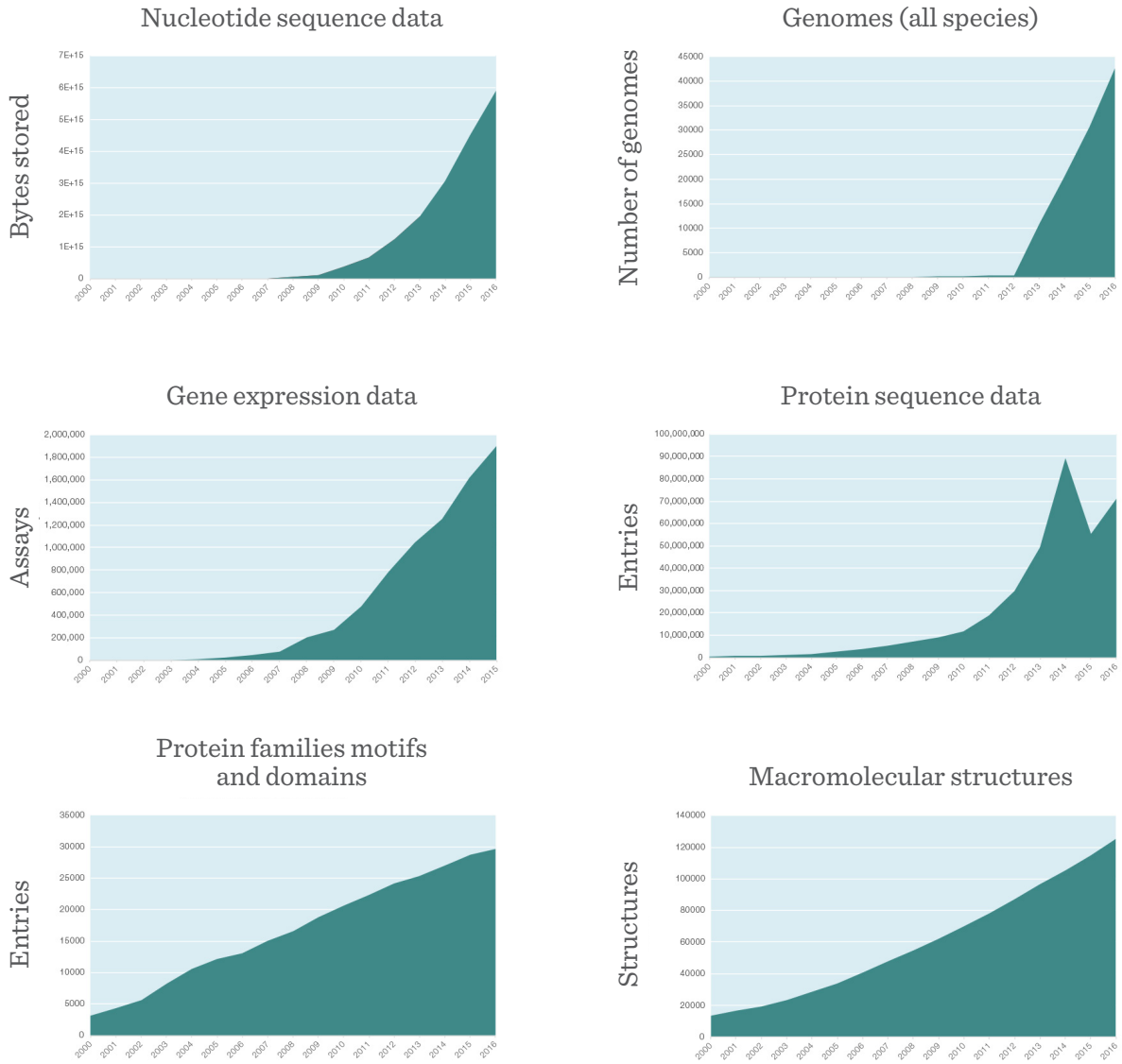
Data growth in 2016

As the cost of generating data continued to fall in 2016, our core data resources continued their steady growth. Two areas of very rapid growth were metagenomics and metabolomics data. Our Metagenomics data service grew by 11-fold in 2016 and, in collaboration with its US counterpart MG-RAST, launched the Metagenomics Exchange. MetaboLights became the recommended metabolomics repository for *Nature Scientific Data*, BioMed Central, *Metabolomics*, PLOS , *F1000* and the EMBO journals.



A comparison of data growth by data type. Note the y-axis is logarithmic, so growth in all data types is still exponential. Growth in the European Nucleotide Archive (ENA) is still exponential, but storage challenges have been mitigated by compression techniques such as CRAM. Growth in array data is slowing, but mass spectrometry data volumes continue to climb.

- Nucleotide sequence data: 5.91 petabytes stored (compare to 4.52 PB in 2015)
- Genomes – all species and strains: 42 529 (compare to 30 674 in 2015)
- Metagenomic samples: Over 90,000 datasets
- Gene expression assays: 2.2 million (compare to 1.90 million in 2015)
- Protein sequences: 71 million (compare to 55.2 million in 2015)
- Protein families, motifs and domains (in InterPro): 29 700 (compare to 28 678 in 2015
- Macromolecular structures: 125 463 (compare to 114 691 in 2015)



Data growth in EMBL-EBI's public archives, 2000 through 2016.

EMBL-EBI Research

Collaboration and career development

Our researchers work in a uniquely bioinformatics-focused environment, but interact daily with experimentalists on the Wellcome Genome Campus and elsewhere. They have the advantage of world-class computational infrastructure on campus, close integration with EMBL's research programme and facilities on five other sites in Europe. This blend fosters interdisciplinarity, and gives our scientists the tools they need to explore new frontiers in science.

In 2016 we recruited two new research group leaders, who bring new perspectives and translational focus to our programme.

Zamin Iqbal, who joined us from the University of Oxford develops computational methods for representing genetic variation, and uses them to study bacteria and parasites. His translational work focuses on applying whole-genome sequencing to pathogens in a clinical setting, and developing tools to support outbreak surveillance.

Evangelia Petsalaki, who joined from Lunenfeld Tanenbaum Institute at Mt. Sinai Hospital, Toronto studies human cell signalling, creating predictive and conditional whole-cell signalling models to gain insights into basic cell functions and disease mechanisms. This work aids the design of new, precise, therapeutic approaches, and facilitates the discovery of reliable biomarkers of disease.

Research highlights in 2016

Human disease

Cancer cell lines predict drug response: New research shows that patient-derived cancer cell lines harbour most of the same genetic changes found in patients' tumours, and could be used to learn how tumours are likely to respond to new drugs. The findings, published in *Cell*, will help to increase the success rate for developing new, more personalised cancer treatments (Iorio et al., *Cell*)

Breast cancer: towards personalised treatment: The largest-ever study to sequence the whole genomes of breast cancers, led by the Wellcome Trust Sanger Institute with significant involvement of EMBL-EBI scientists, uncovered five new genes associated with the disease and 13 new mutational signatures that influence tumour development. Published in *Nature* and *Nature Communications*, the research pinpoints where genetic variations in breast cancers occur. The findings provide insights into the causes of breast tumours and demonstrate that breast-cancer genomes are highly individual (Morganella et al., *Nature Commun.*; Nik-Zainal et al., *Nature*).

Cancer signatures: In the era of personal genomics, our research is increasingly translational and related to problems of direct significance to medicine and the environment. The Gerstung group, which was established at EMBL-EBI in September 2015, has been tackling computational cancer biology by developing new methods to support research on local, national and

international levels. They are developing tools to decipher mutational signatures associated with different, often newly identified cancers.

Acute Myeloid Leukaemia: Acute Myeloid Leukaemia (AML) is not one, but at least 11 different diseases. Published in the *New England Journal of Medicine*, research from Gerstung and colleagues shows that different 'constellations' of genetic changes can explain why survival rates vary among AML patients. A new clinical tool from the collaboration incorporates a patient's individual genetic details into a knowledge base to predict the outcome and treatment choice for that patient (Gerstung et al., *Nature Genetics*; Papaemmanuil and Gerstung, *NEJM*).

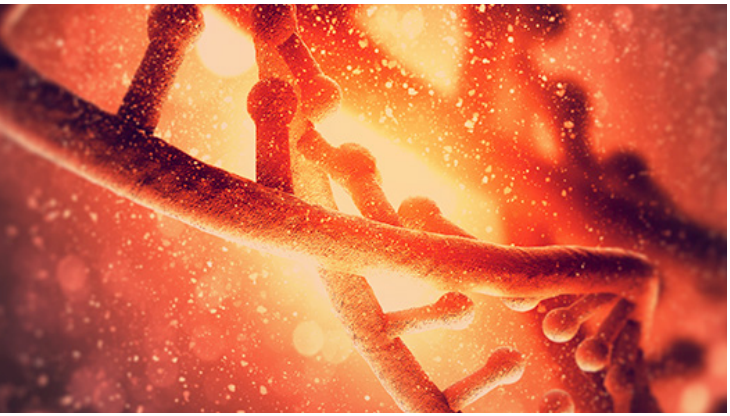
Immunology meets single-cell sequencing: A series of RNA sequencing experiments by the Teichmann group at EMBL-EBI and the Sanger Institute led the group to develop a new technique for understanding T-cell receptors. TraCeR, a single-cell sequencing tool, allows researchers to determing both the sequence of T-cell receptors in individual cells, along with each cell's gene expression profile. This opens up new possibilities for developing rapid diagnostics based on the genetic profile of blood cells (Stubbington et al., *Nature Methods*).

Tapio Lönnberg of EMBL-EBI and collaborator Mike Stubbington of the Wellcome Trust Sanger Institute, explaining TraCeR, their new single-cell sequencing tool for determining T-cell receptor sequence and gene expression profile in individual cells.

Blueprint of immune cells: One output from the International Human Epigenome Consortium (IHEC) in 2016 was a unique study about the interplay between the genome and the epigenome, in the context of molecular phenotypes and disease. Led by Nicole Sorazano at the Sanger Institute with substantial input from the Stegle group at EMBL-EBI, the collaborators explored the role of epigenetics in the development of immune cells (CD14+ monocytes, CD16+ neutrophils, naïve CD4+ T cells). They identified hundreds of regions on the genome where they could pinpoint the likely molecular causes underlying an individual's predisposition to immune-related diseases. The findings, published in *Cell*, bridge a major gap in our understanding of how genotype and the epigenome interrelate, and affect phenotype (Chen et al., *Cell*).

Developmental biology

Anatomy of a decision: Using single-cell sequencing, the Marioni group and collaborators at the Wellcome Trust–MRC Cambridge Stem Cell Institute gained new insights into how a mouse embryo first begins to transform from a ball of unfocused cells into a small, structured entity. Published in *Nature*, the analysis of over 1000 individual cells of gastrulating mouse embryos



provides an atlas of gene expression during very early, healthy mammalian development (Scialdone et al., *Nature*).

Foetus, or placenta? Research from the Marioni group, published in *Cell*, clarifies the subtle differences between seemingly identical cells at a very early stage of development (only four cells), when they are poised to become either foetal or placental cells. Using single-cell genomics, they showed that some genes in each of the four cells behaved differently. The activity of several genes that form part of the 'pluripotency network' differed the most between cells. When activity of such genes was reduced, the activity of a master regulator, which directs cells to develop into the placenta, was increased. The genetic and epigenetic signatures revealed in the study indicate the tendencies of early embryonic cells (Goolam et al., *Nature*).

Cellular processes

Parallel single-cell profiling: Single-cell sequencing is used to study how gene expression profiles ('transcriptomes') vary between cells, and to explore chemical modification of DNA ('epigenetics'), which changes to gene expression. A new method from the Stegle group and their collaborators makes it possible to study the epigenome and transcriptome of a single cell at the same time, in parallel. Published in *Nature Methods*, the protocol gives the first direct view on the relationship between heterogeneity in DNA methylation and variation of expression in specific genes across single cells (Angermueller et al., *Nature Methods*).

Drivers of evolution hidden in plain sight: New research led by the Beltrao group showed that the biological diversity needed for evolution can be generated by changes in protein modifications. The findings, published in *Science*, provide valuable insights into how different species adapt to different environments and could shed light on how pathogens evolve and become resistant to drugs (Studer et al., *Science*).

A clear view on signalling pathways: Combining the power of 27 data resources, Omnipath helps researchers see biological signalling pathways with unprecedented accuracy. Developed by the Saez-Rodriguez group at EMBL-EBI and published in *Nature Methods*, OmniPath offers a comprehensive, unified collection of literature-curated signalling pathways based on an

analysis of 41 000 scientific papers. It also provides comprehensive guidelines, based on an extensive examination of more than 50 data resources, to help researchers select the most appropriate data resource for their work (Turei et al., *Nature Methods*).

Biotechnology

Finding enzymes by function: The Thornton group uses structural data to explore enzymes, and develops robust computational tools to improve enzyme design. In 2016 they published a new method for identifying enzymes with similar function, which can be the starting point for engineering a new enzyme. Searching by function can be problematic in the case of isomerases, a class of enzyme-transforming molecules with the same molecular weight but different 3D structures. The Thornton group created a new way to describe function that highlights relationships between enzymes in different classes, applied their approach to a reliable knowledge base of isomerases, verified their findings using EC BLAST and delivered an effective tool for identifying proteins with similar function. Published in *PNAS*, the new method characterises the chemistry and catalytic function of isomerisation, and allows researchers to explore the diversity of the chemistry of life more easily (Martínez Cuesta et al., *PNAS;* Dönertaş et al., *PLoS One*).

Genotype to phenotype

What variation reveals: New methods mean new questions, surprising answers and deeper insights into the workings of life. The Stegle group's methods for revealing hidden sources of variation, and for filtering out 'technical noise', continue to fuel discoveries linking genotype and phenotype. Their novel statistical methods for studying gene regulation, GWAS and causal reasoning in molecular systems are elucidating how genetic background and environment jointly shape healthy and diseased states. Their machine-learning methods are also improving the accuracy and resolution of single-cell biology (Angermueller et al., *Molecular Systems Biology*).

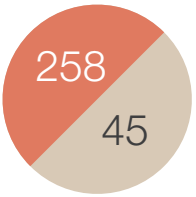
Stress-testing statistics

Maximum Likelihood: One of the methods of choice for estimating evolutionary relationships among various biological species is Maximum Likelihood (ML) inference. The Goldman group published a 'proof of consistency' supporting the method, taking into consideration how the presence of gaps in a multiple sequence alignment (usually the result of insertions or deletions of genetic material) affects the accuracy of inferred phylogenies. The group showed that the sacrifice of information does not ultimately lead to inconsistency, and delivered the clearest proof to date of statistical consistency of ML phylogenetic reconstruction for 'ungapped' or 'gapped' alignments (Truszkowski et al., *Systematic Biology*).

Research in 2016: facts and figures

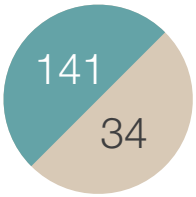
- Welcomed 141 visitors who worked with us on specific projects
- Visitors in total represented 34 nationalities
- Co-authored 258 scientific publications with collaborators at 430 organisations in 45 countries
- Joint grant funding on 186 projects with 624 institutes in 63 countries

Co-authored scientific publications



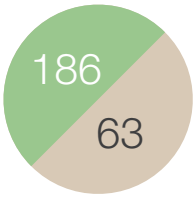
Collaborating countries

Visitors



Visitor nationalities, including trainees

Joint grant funding



Collaborating countries

Open author profiles

In 2016 Europe PMC launched author profile pages, which provide a graphical overview of your publications and your citation rate over time based on your ORCID record.

Selected publications

- Angermueller C, et al. (2015) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* 13:229–232
- Angermueller C, et al. (2016) Deep learning for computational biology. *Molecular Systems Biology* 12:878
- Dönertaş HM , Martínez Cuesta S , Rahman SA , Thornton JM (2016) Characterising complex enzyme reaction data. *PLoS One* 11: e0147952
- Gerstung M, Papaemmanuil E, et al. (2017). Precision oncology in acute myeloid leukemia using a knowledge bank approach. *Nature Genetics* 49:332–340
- Goolam M, et al. (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in four-cell mouse embryos. *Cell* 165:61-74
- Iorio F, et al.(2016) A landscape of pharmacogenomic interactions in cancer. *Cell* 166:740-754
- Martínez Cuesta S, Rahman SA, Thornton JM (2016) Exploring the chemistry and evolution of the isomerases. *Proc Nat Acad Sci USA* 113: 1796–1801
- Morganella S et al. (2016) The topography of mutational processes in breast cancer genomes. *Nature Communications* 7:11383
- Nik-Zainal S et al. (2016) The somatic genetics of breast cancer revealed by 560 whole genome sequences. *Nature* 534:47-54
- Papaemmanuil E and Gerstung M (2016) Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine* 374:2209-2221
- Scialdone A, et al. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature* 535:289–293
- Stubbington MJ, Lönnberg T, et al. (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods* 13:329–332
- Studer RA, et al., (2016) Evolution of protein phosphorylation across 18 fungal species. *Science* 354:229-232
- Truszkowski J and Goldman N (2016) Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Systematic Biology* 65:328-333
- Turei D, Korcsmaros T and Saez-Rodriguez J (2016) Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods* 13:966–967

Major collaborations

Global Alliance for Genomics and Health

In 2016 Ewan Birney was appointed Chair of the Global Alliance for Genomics and Health (GA4GH), leading the global effort to accelerate medical and research advancements through the responsible sharing of genomic and clinical data. Birney, who is also a member of the Scientific Advisory Board at Berlin Institute of Health, was also appointed non-executive Director of Genomics England in 2016. This provides an opportunity to bridge the gap between our research institute and the needs of national healthcare systems in Europe.

GA4GH counts over 300 members from the public and private sectors, and EMBL-EBI leads the way in realising interoperability. In 2016 our Variation Annotation team added new functionality to the GA4GH Application Program Interface (API). Users can now access GA4GH sequence features, genotype calls, variant annotation, lists of reference sequences and metadata in 20 different ways, with endpoints available via EMBL-EBI REST servers.

DECIPHER

Deciphering Developmental Disorders (DDD), a rare-disease project funded jointly by Wellcome and the UK Department of Health, used whole-exome sequencing to diagnose 27% of 1133 children with developmental disorders who had been examined but remained undiagnosed. Most of the diagnostic variants the researchers identified in known genes were novel, and were not present in databases of known disease variation.

In 2016 EMBL-EBI collaborators released a supporting database for DDD called Gene2Phenotype, which offers manually curated gene-to-phenotype associations or disease associations. They also created a Variant Effect Predictor (VEP) plugin that uses these data to filter results from sequencing tests. These tools are invaluable to clinicians seeking a rapid diagnosis for their patients.

Data coordination

The low cost of generating high-throughput data has spurred large, international research consortia to tackle profound scientific questions. This work often involves several institutions producing different types of data at scale, which then needs to be curated, integrated and publicly hosted for further analyses and reuse by the broader scientific community.

Our data coordination expertise allows us to support research collaborations of all sizes. This work is vital to the success of large-scale research initiatives that deliver lasting value to science and society. It ensures research outputs are seamlessly integrated into reference databases.

Major data-coordination projects 2016

- **BLUEPRINT**: 47 scientific papers and 1000 datasets representing over 50 primary blood cell types from healthy individuals as well as the neoplastic counterparts of those cell types.
- **IMPC**: The International Mouse Phenotyping Consortium identified 410 genes that are 'essential to life' in the mouse, providing tremendous insights into gene function and congenital disorders.
- **COMPARE**: New cloud platform supports autonomous sequence search and analyses of incoming data to identify and type bacterial and viral samples from the environment.

BLUEPRINT

Epigenomics research centres on understanding how a single genome gives rise to all the different cell types in the human body, and how and when different cells interpret their instructions. The International Human Epigenome Consortium (IHEC) coordinates efforts to establish epigenomic maps, or 'reference epigenomes' for myriad human cell types.

EMBL-EBI provided data coordination for BLUEPRINT, an EU-funded IHEC initiative to generate the reference epigenomes of cell types from human blood. In 2016 the project delivered an invaluable resource to the scientific community, releasing over 1000 datasets that support research into blood-based diseases such as myeloid and lymphoid leukemias. The data represent over 50 primary cell types from healthy individuals as well as the neoplastic counterparts of those cell types. It also includes data from patients suffering from type-1 diabetes.

Data coordination often goes hand-in-hand with data analysis. On its completion, BLUEPRINT partners published 47 scientific papers in *Cell* and other high-profile journals, with EMBL-EBI researchers contributing to analyses. One of these studies featured a new technique that identifies parts of the genome that are in physical contact with one another. Using this technique, the researchers pinpointed hundreds of thousands of regions involved in switching genes on and off.

Blood, big data and epigenetics

IHEC and BLUEPRINT support biomedical research:

- 50+ primary cell types from healthy individuals, the 50+ neoplastic counterparts of those cell types, and several more from patients suffering from type 1 diabetes
- 47 scientific papers in *Cell* and other high-profile journals
- 8000+ datasets freely available to all through Ensembl

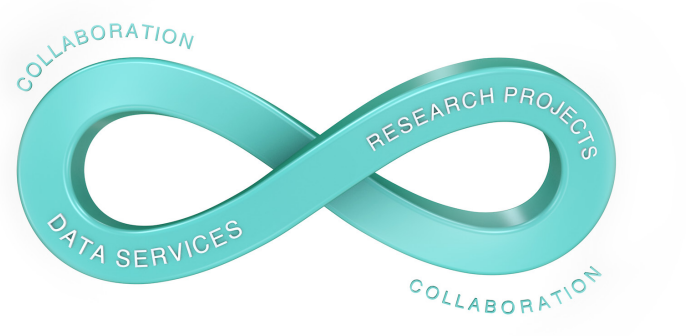
IMPC

EMBL-EBI manages data for the International Mouse Phenotyping Consortium (IMPC), an exemplar of replacement, reduction and refinement in animal research. The IMPC's goal is to determine the functions of some 20 000 genes common to mice and humans. EMBL-EBI, in collaboration with MRC Harwell and the Sanger Institute, delivers IMPC informatics.

Over 4600 human diseases are associated with IMPC mouse models. In 2016 the IMPC portal offered over 3300 phenotyped genes, and integrated new brain histopathology data from the BBSRC-funded PhenoImageShare project.

A 2016 study by the IMPC provided insights into the functions of 1751 genes, and identified 410 that are 'embryonic lethal' – that is, essential for life. Using ExaC Consortium data, the researchers showed that human orthologues of the 'essential' genes were more intolerant to variation compared to nonessential genes, which makes them promising candidates for studying human disease.

The IMPC incorporates high-resolution 3D imaging and automated, computational analysis of these images. They share all of the data and images they generate through an open-source, web-based resource, without embargo. The mouse strains are available through public repositories such as EMMA and KOMP, which enables users to mine the data, generate insights about mouse phenotypes in relation to human phenotypes, and design their own experiments.



The data produced in research projects is maintained in public data services, where it can be reused alongside new data in future research projects.

COMPARE

In 2016 the European Nucleotide Archive (ENA) team launched a platform that makes it easier for researchers within the COMPARE project to analyse pathogen genomes and feed the results into the public archives. This platform sits on top of ENA content and allows sequencing data to be routed autonomously through the appropriate analysis pipelines, with outputs going directly back into the ENA.

One analysis workflow offered in 2016 supports comprehensive bacterial identification, typing and antimicrobial resistance prediction, all via the Danish Technical University's Centre for Genomic Epidemiology (CGE) system. This allows rapid identification of outbreak samples by public health and other professionals and comparison to existing sequence information.

Another analysis pipeline uses the Erasmus MC tools to identify and type mixtures (metagenomes) of viruses. By supporting this work, we help scientists measure and understand the composition and structure of viral outbreaks.

Data integration for cancer genomics

As part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project of the International Cancer Genome Consortium, we led a working group on RNA and DNA data integration in collaboration with the University of California Santa Cruz and the Memorial Sloan Kettering Cancer Center.

Together with Peking University, we performed an integrated RNA/DNA analysis of over 1000 samples of 27 different cancer types to discover novel gene fusions. This uncovered new mechanisms for gene fusion formation via composite events of genome rearrangements and read-throughs. The results, including an analysis for expression Quantitative Traits (eQTLs), are scheduled for publication in 2017.

Industry collaborations

Benchmarking knowledge: The structure of molecular networks plays a vital role in determining how our cells behave, but in the ever-changing, endlessly complex cellular environment, they are a bit of a moving target. There are challenges on every level: pinning down the interplay between proteins, piecing together 'healthy' networks, and dissecting the myriad changes that disease can cause.

A community effort to assess the feasibility of using computational methods to piece together biological networks showed that solutions to this puzzle may be coming into reach. This DREAM challenge, published in *Nature Methods*, was coordinated by EMBL-EBI, the Medical Research Council Biostatistics Unit (MRC-BSU) and IBM.

Predicting toxicity: We are partners in several EU-funded projects that aim to better identify and curate toxicity data and apply it to the prediction of toxicological endpoints. One of these, EU-ToxRisk, aims to integrate advancements in cell biology, 'omics technologies, systems biology and computational modelling to define the complex chains of events that link chemical exposure to toxic outcome.

Diabetes database: We launched the first implementation of a federated database for Type 2 Diabetes, in the context of an NIH Accelerating Medicines Partnership with the Broad Institute and the University of Oxford.

Innovative Medicines Initiative projects

Our Industry Programme helps its partners and EMBL-EBI researchers establish collaborations in the context of the Innovative Medicines Initiative (IMI). The IMI is a partnership between the European Union (represented by the European Commission) and the European pharmaceutical industry (represented by the European Federation of Pharmaceutical Industries and Associations).

See the Industry section for a list of active projects.

Building skills and capacity

Training the next generation of computational biologists

The EMBL International PhD Programme at EMBL-EBI in 2016

The students in the EMBL International PhD Programme at EMBL-EBI receive advanced, interdisciplinary training in molecular biology and bioinformatics, studying full-time at EMBL-EBI while registered with, and receiving their degree from, the University of Cambridge. We provide theoretical and practical training to underpin an independent, focused research project under the supervision of a faculty member and monitored by a Thesis Advisory Committee composed of EMBL faculty, local academics and, where appropriate, industry partners.

In 2016 we benefited from the presence of 28 PhD students, welcoming seven newcomers. Six of our students successfully defended their theses.

Theses defended in 2016

- Francesco Paolo Casale (supervised by Oliver Stegle): Multivariate linear mixed models for statistical genetics
- Emanuel Gonçalves (supervised by Julio Saez Rodriguez): Modelling regulatory interactions between metabolism and signalling
- Kevin Gori (supervised by Nick Goldman): Clustering approaches for incongruent phylogenies
- Tommaso Leonardi (supervised by Anton Enright): Insights into the function of non-coding RNAs
- Tom Rensch (supervised by Paul Flicek): Applications for ChIP-sequencing data reusability
- Michael Schubert (supervised by Julio Saez-Rodriguez): Gene expression signatures for cancer cell line drug sensitivity and patient outcome

Postdoctoral research and career development

EMBL-EBI participates in the EMBL Postdoctoral Programme, and as such benefits from a multicultural community of excellent researchers. There were 44 postdoctoral fellows working at EMBL-EBI in 2016. Our postdocs benefit from the rich scientific and technical exchange, collaborative culture and creative environment of the Wellcome Genome Campus. Our three collaborative, interdisciplinary postdoctoral research schemes combine wet- and dry-lab approaches, consistently demonstrating the power of using bioinformatics in collaboration with traditional, laboratory-based science.

EMBL's interdisciplinary EIPOD programme, funded by EMBL and an EU Marie Skłodowska-Curie Actions Cofund grant, encourages research that crosses scientific boundaries. The six EIPOD fellows at EMBL-EBI work in scientific fields that are usually separate, and often transfer techniques to a novel context. EIPOD fellow Dénes Türei

"EMBL-EBI is involved in international projects and gives you both access to leading scientists and opportunities to publish - it's the front line of research. It is also good for people who may be thinking of moving to private companies. [EMBL-EBI] manages all these databases and has that aspect of huge data analysis and infrastructure, so you can work on something technical as well, not just academic science."

– Santi Gonzalez Rosado, postdoc EMBL-EBI

in the Saez-Rodriguez group helped develop a method that shows biological signalling pathways with unprecedented accuracy. His perspective on the project: "It has been exciting to work together with people from so many disciplines, and produce this concise view into the collective, current knowledge of signalling pathways."

The postdocs in our EBPOD programme have joint fellowships between EMBL-EBI and the Cambridge Biomedical Research Centre. Their projects apply computational approaches to translational clinical research involving human subjects, for example global cancer collaborations. EBPOD fellow Hanna Najgebauer in the Gerstung group says that she chose EMBL-EBI because "many of the world's large cancer projects deposit their data here, from blood cancers, tumours and circulating tumour DNA, so on the computational side it is great. But it's the collaboration between wet and dry science that is really unique."

The four postdocs in our ESPOD programme have joint fellowships between EMBL-EBI and the Sanger Institute, which allows them to combine experimental and computational approaches seamlessly. For example, one project launched in 2016 centres on predicting the pathogenic impact of genetic variants, combining experimental work with analyses of domain architecture and protein structure.

Henry Wellcome postdoctoral fellow and EIPOD alumna Amelie Baud: "Doing science is made easy at EMBL-EBI. The computational power we have here is amazing and it comes with excellent administrative support."

Bioinformatics Training Programme

EMBL-EBI's training programme in bioinformatics is designed to help professionals exploit data from new and emerging technologies, which can be challenging in a time of rapid technological change and scientific advancement. As part of EMBL's International Centre for Advanced Training, the programme is complemented by a broad range of courses in molecular biology held at different sites in Europe.

Our courses go well beyond guiding researchers to data portals, often featuring leading-edge bioinformaticians providing hands-on instruction in the use of advanced, sophisticated tools for multi-omics data analysis.

The reach of our Training Programme extends throughout the world, and impacts research and development across multiple disciplines and sectors. In 2016 our staff participated in over 300 face-to-face training, outreach and knowledge-exchange events, connecting with over 13 000 people.

People at over 350 000 unique IP addresses accessed our online training resource, demonstrating the growing need for on-demand training in bioinformatics.

In addition our Train online resource, we launched a new training portal for the Ensembl project that features course books, presentations, worked examples and training event details.

Training in 2016

- Participated in 350 training and outreach events, supporting biomedical and life-science professionals throughout the world
- Extended the reach of 'Train online', which was accessed by 358 905 unique IP (compare with 214 470 in 2015)*
- Expanded our training offerings, adding 12 new courses and 27 webinars
- Co-created the curriculum for five ELIXIR 'train the trainer' courses, which prepared 40 new bioinformatics instructors in Europe
- Trained eight new bioinformatics instructors in Australia

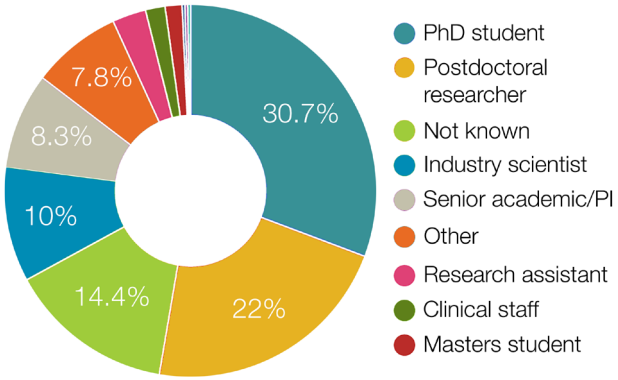
**A unique IP address may represent any number of users at a single institute.*

Webinars and the regular addition of new online courses serve to reinforce the professional development activities of our Training Programme.

Courses and workshops

Delegates to our courses in 2016 covered a wide range of career levels. The figures below reflect career levels and host countries of delegates to the following courses:

Advanced RNA-Seq Plant and Pathogens, Cancer Genomics, Structural Bioinformatics, Metagenomics, Bioinformatics for Protein Biology, MTP, Exploring Genetic Variation, Exploring Biological Sequencing Data, EMBO Practical Course on Metabolomics, Bioinformatics for Life Scientists, Introduction to Next Generation Sequencing, Introduction to Omics data integration, Networks and Pathways, Bioinformatics for Principal Investigators, Next Generation Sequencing, Micro and long non-coding RNAs, and Exploring Genomic Variation with High-Throughput Sequencing.



People at many different career levels, working in many different life-science sectors, travel from all over the world to attend our courses. Most of them disseminate what they learned to others - sometimes individuals, sometimes an entire class.

Impact: Training Programme in 2016

End-of-course surveys

- 93.7% of respondents rate the content of our courses "good" (50%) or "excellent" (43%)
- Growing proportion of delegates from industry and healthcare sectors

Six-month post-course surveys

- Over 92% of course participants who responded to a survey sent six months after the event indicated they would recommend the course to others
- 84% had disseminated their learning to others
- 69% still use data-analysis methods covered in their course
- 23% established new collaborations with others during the course

Professional qualifications: The Health Education England-led Master's in Genomic Medicine programme teaches healthcare professionals how to work with genomic data. EMBL-EBI, together with the Sanger Institute and the Wellcome Genome Campus Advanced Courses, is a partner on the courses for the programme at University of Cambridge. EMBL-EBI staff also contribute to programme courses at the University of Southampton, University of Exeter and Queen Mary University of London.

To ensure publicly funded research infrastructures are equipped with the right expertise, we are also co-developing training for technical operators (CORBEL project) and an executive Master's programme for managers and leaders of research infrastructures (RItrain).

Blended learning: Our new Bioinformatics for Discovery programme serves biologists working in pharmaceutical, agri-food and consumer goods companies, and is based on real-life research problems. Funded by the Biotechnology and Biological Sciences Research Council (BBSRC), the programme offers face-to-face workshops complimented by webinars and a forum for peer-to-peer and instructor interactions.

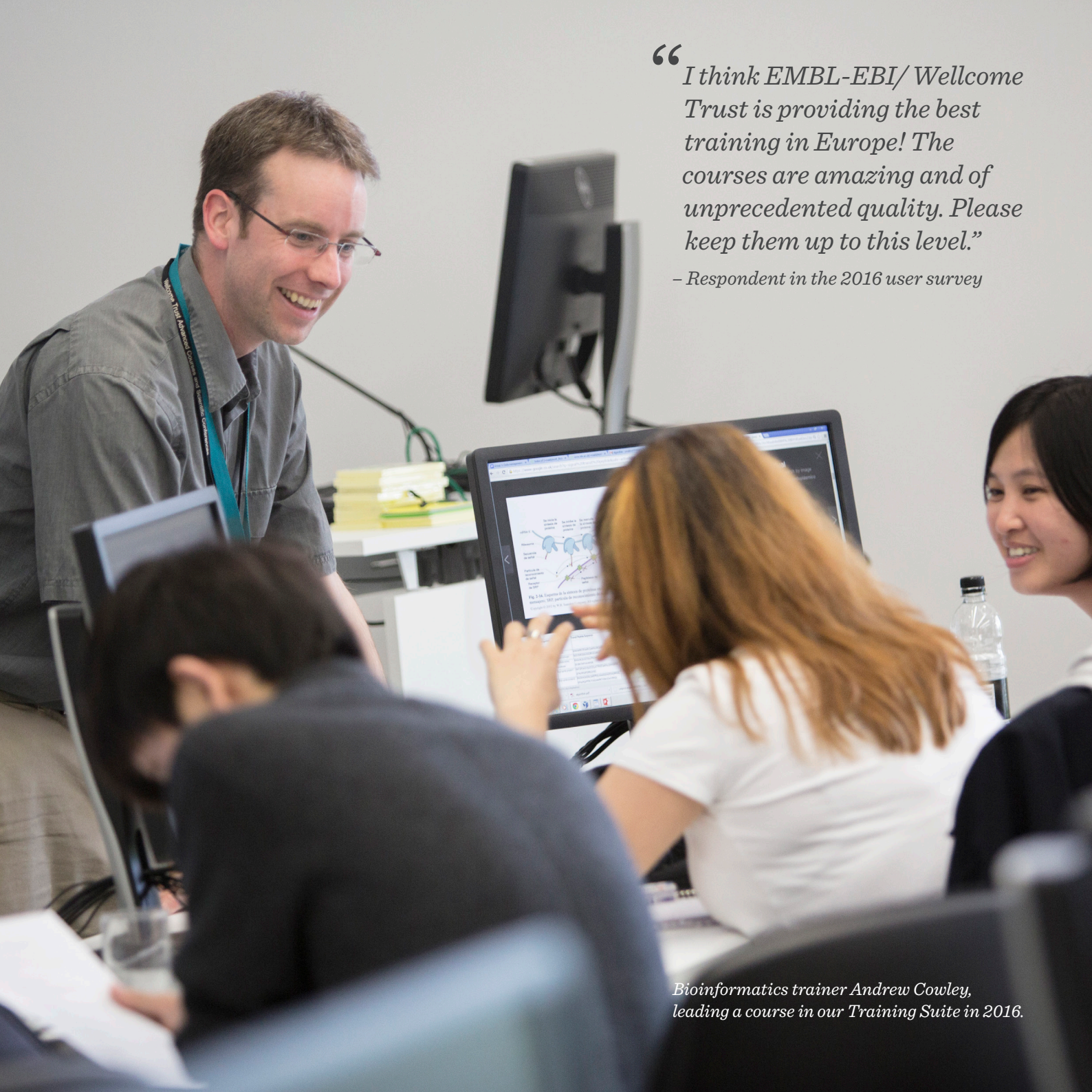
Training new instructors: We build bio-informatics capacity by training new instructors, who in turn carry out bioinformatics training in other countries. In 2016 our curriculum was used to train new bioinformatics instructors in Australia and throughout Europe, both through our own programme and in partnership with ELIXIR.

Training needs: Our Training Programme consults closely with the community to identify gaps in training provision, and seeks to address them across sectors. We have developed a competency-based approach to identifying training needs and have applied these in several contexts, including research infrastructure management and operation and use of high-end computer for biomolecular research (BioExcel).

Public engagement: The Protein Data Bank in Europe (PDBe) team set up a collaborative project with art departments at secondary schools in Cambridge, UK. This project introduces pupils to the wealth of information available in the PDB, challenging and inspiring them to create artwork that depicts health-related concepts.

Our annual **Science and Society** event, run under the auspices of the EMBL Science and Society Programme, presents exciting, relevant research topics to a general audience in central Cambridge. our 2016 event focused on pandemics, welcoming over 250 attendees at the Cambridge Union and generating significant interest in a wide audience on social media.

Our staff participated in Big Biology Day in Lincolnshire, UK, training pupils to lead activities that engage young people with bioinformatics. Staff throughout the institute participate in a range of events led by the Campus Public Engagement Team, including the Cambridge Science Festival, presentations to school children and many others.



> *I think EMBL-EBI/ Wellcome Trust is providing the best training in Europe! The courses are amazing and of unprecedented quality. Please keep them up to this level."*
>
> *– Respondent in the 2016 user survey*

Bioinformatics trainer Andrew Cowley, leading a course in our Training Suite in 2016.

Meeting the needs of industry

The EMBL-EBI Industry Programme

Our Industry Programme is unique in the world. It is a forum for interaction and knowledge exchange for those working at the forefront of applied bioinformatics in 22 major companies with global R&D activities. The programme focuses on pre-competitive collaboration, open-source software and informatics standards, which have become essential to improving efficiency and reducing costs for the world's bioindustries.

Three life-science companies with global R&D operations joined the EMBL-EBI Industry Programme in 2016: AbbVie, Merck Sharp & Dohme (MSD) and Takeda, through Takeda Oncology, Boston, US. The addition of these companies allows us to further extend the international reach of our membership, and extend our activities in the US where there is an unmet need in this area.

Member-driven workshops

Our Industry Programme members specifically request workshops designed to help them develop data and tools around emerging technologies and shared areas of interest. Resolving practical challenges in the uptake of new technologies removes roadblocks to discovery, paving the way to identify new biological targets for treating disease, and novel mechanisms that optimise drug-discovery pipelines.

In 2016 we delivered 11 Industry Workshops covering topics prioritised by the members, ranging from cryo-electron microscopy/tomography to predictive toxicology and antibody developability.

Reducing redundancy in cancer research

Our Industry Programme members provided invaluable input to a new, NIH-funded project to centralise all data related to patient-derived xenograph (PDX) mouse models, which are important tools for cancer research. The new resource, PDX Integrator, will be a public resource to facilitate data access, reduce redundancy, and enable data comparison and integration in PDX-related cancer research.

Innovative Medicines Initiative

Our Industry Programme is a starting point for EMBL-EBI collaborations in the IMI programme.

In 2016 our active IMI collaborations included:

- **EBiSC**: European Bank for induced pluripotent Stem Cells
- **EU-AIMS**: Bringing together academic research, commercial R&D and patient organisations to develop and assess novel treatment approaches for autism
- **EMIF**: A common framework for patient-level data to facilitate access to diverse medical and research data sources

Industry workshops in 2016

Opportunities in data:

- Predictive toxicology
- Biomedical opportunities in deep-sequencing of immunological repertoires
- Informatics, genetics and genomics for target selection and validation
- Computational immuno-oncology
- Cryo-Electron Microscopy/Tomography

Data standards:

- Practical use of biomedical ontologies
- Embedding clinical standards in research: development, implementation and curation

Tool development:

- Molecular informatics open-source software
- Success in life science R&D through User Experience Design
- Antibody developability: aggregation, stability and viscosity
- Deep learning: applications in pharma and agri-food

Revathi Nathaniel and Nikiforos Karamanis, EMBL-EBI UX designers, with delegates at the 'Success in life science R&D through User Experience Design' workshop in 2016, jointly organised with the Pistoia Alliance.

Open Targets

In 2016 the US-based pharmaceutical company Biogen joined Open Targets, a partnership that also includes EMBL-EBI, GSK and the Sanger Institute. Open Targets helps scientists discover and prioritise relationships between biological targets and diseases. With the addition of Biogen, the partners started to pursue neurodegeneration as a new therapeutic area.

Open Targets published three updates of its freely available Target Validation digital platform, including new visualisations, new ways to search, and additional data sources. The platform is designed based on extensive user-experience research from our Web Development team.

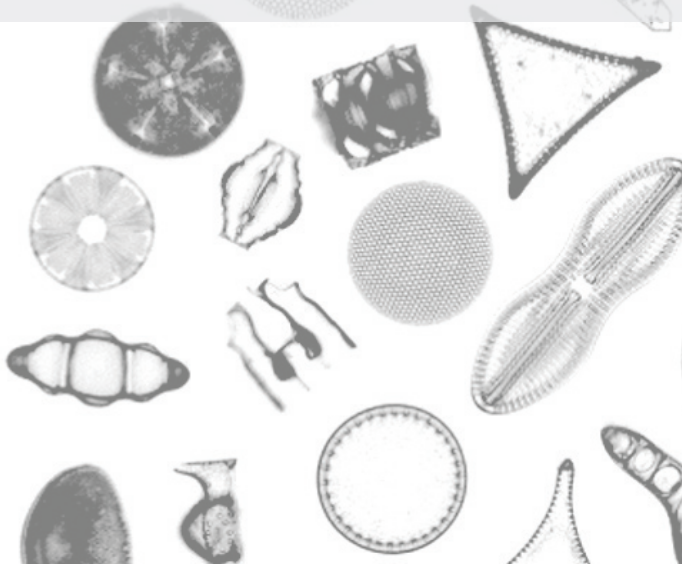
EMBL-EBI's role in the development and implementation of the Target Validation platform cannot be understated. For example, we enhanced and integrated ontology tools and content to support disease annotations, which makes it possible for Open Targets to link rare and common diseases according to shared phenotypes.

To help researchers establish high-confidence links between targets, compounds and diseases, our ChEMBL team annotated marketed drugs, withdrawn drugs and compounds in clinical development with therapeutic-target and 'indication' information. This work, published in *Nature Reviews Drug Discovery*, has been of great utility in the context of both Open Targets and Illuminating the Druggable Genome.

In all, 60 EMBL-EBI staff worked on some 40 Open Targets projects in 2016, including computational pipelines, oncology, induced-pluripotent stem cells, single-cell genomics and an inventory of epigenetic complexes.

Open Targets in 2016

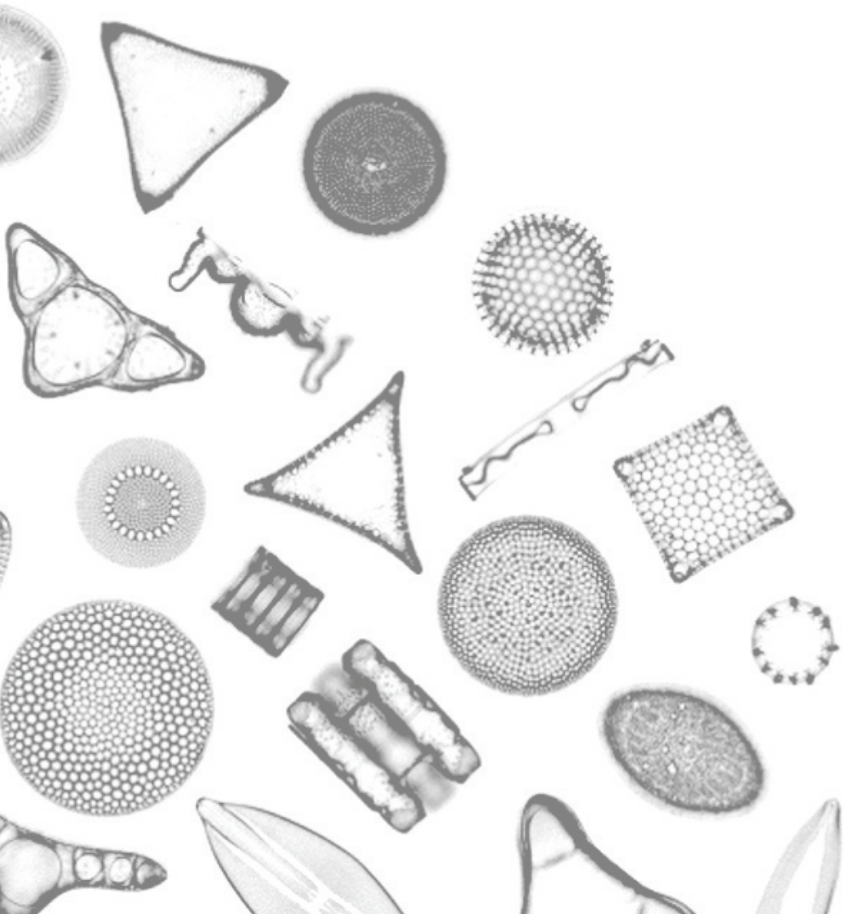
- New partner: Biogen
- Target Validation Platform: three releases
- Freely available integrated data: More than 31 000 therapeutic targets, 8600 diseases and phenotypes and 2.5 million target-disease associations
- Usage: 900 unique IP addresses access the platform every week
- New search mechanisms: search with user list of targets, drug names and phenotypes



Small and medium-sized enterprises

EMBL-EBI supports SMEs primarily by providing free data, tools and infrastructure.

We also organise networking events in collaboration with organisations such as OneNucleus, the UK Trade and Investment agency (UKTI), the InnovateUK Bioinformatics knowledge-transfer network and the ELIXIR SME and Innovation Forum.



Highlight: SME collaboration

'Mining' the latest scientific knowledge for commercial application demands robust analytical tools. But creating them takes highly specialised knowledge and costly infrastructure, which can be a barrier for product discovery and development. Large companies make use of external companies to provide these services, but such outsourcing is out of reach for most smaller businesses.

An Innovate UK-funded collaboration with Cardiff-based company Biocatalysts Ltd. is facilitating access to high-end analytical tools, allowing companies of all sizes to compete for first advantage and scale up their solutions. MetXtra, the project's novel software platform, enables enzyme discovery through the analysis of large, open-access environmental sample datasets hosted in EMBL-EBI's Metagenoimcs service. The platform has enabled the commercialisation of new enzymes, and will boost innovation in the production of fine chemicals, flavours, fragrances, and pharmaceutical products.

Infrastructure

In 2016, demands for EMBL-EBI's data resources continued to grow. Our technical service teams installed 72 petabytes of raw storage, and our main compute farms now have 34,000 cores (27 000 high-throughput and 7000 high-performance, specially tuned for large-scale data access) to support the analysis efforts of our users. We were able to continue to scale our infrastructure and meet demand thanks in part to support from the UK's Large Facilities Capital Fund (LFCF).

Variety, capacity and resilience

To ensure our teams can deliver data services that support diverse areas of science, our technical teams employ a broad range of high-performing technologies including Oracle, MySQL, MangoDb, Hadoop and Vertica. To meet the growing demand for all our resources, we have been implementing a database virtualisation programme with enterprise software company Delphix since 2015.

Our geo-dispersed Object Store fulfills the greater part of our storage needs. It offers efficiencies in data resilience, reducing our data-centre footprint and overall costs and improving our environmental standing. Other storage solutions range from general-purpose scale-out such as Netapp and Isilion (which grow to multi petabyte) to high-performance systems like Lustre, Spectrum Scale and our bespoke object tape archive. This gives us maximum flexibility and required value for money in data archiving.

Our internal network is provided by JANET, and moves data at 20 gigabits a second (gb/s). Because of the increasing demand for molecular data, in 2016 we laid the groundwork for an upgrade to 40 gb/s in 2017.

Collaboration in the cloud

In 2016 we developed a cloud portal to support EMBL-EBI and ELIXIR projects. This portal allows participating researchers to select a cloud service provider (including our own internal OpenStack provision), retrieve the relevant software applications and deploy virtual machines or containers. We use GitHub as an application definition and sharing platform, which makes it easier for our cloud engineering teams to collaborate and streamlines versioning.

Embassy Cloud, our infrastructure-as-a-service, is based on the OpenStack cloud platform. It features private, secure 'tenancies' hosted within EMBL-EBI's data centres, providing users with direct access to datasets and services and negating the need to download large data resources before undertaking analyses. Embassy Cloud is currently used by external groups collaborating with EMBL-EBI teams. It includes 6000 cores, 40 terabytes of RAM, 50 terabytes of SSD fast scratch space, two petabytes of NFS and two petabytes of 'object store' storage.

We are partners in several large infrastructure projects, for example the Helix Nebula Science Cloud (HNSciCloud), which aims to establish a European hybrid cloud platform to support the deployment of high-performance computing and big-data capabilities for scientific research. Helix

partners include CERN, the European Space Agency, the European Grid Infrastructure and other institutes from across Europe. Launched in 2016, the technical specifications for this pre-commercial procurement tender were defined according to use cases: Pan-Cancer Analysis of Whole Genomes, EuroBioImaging and ELIXIR. The designs will be prototyped in 2017 before large-scale pilots in 2018.

Cancer research: feasibility study

We conduct feasibility studies to support efforts to establish national research infrastructure. In 2016 we evaluated the technical and economic feasibility of executing bioinformatics workloads in a 'multi-cloud' environment to support big-data analysis for international, pan-cancer collaborations. This work inclued a benchmarking suite that works across multiple providers. We extended our Authentication, Authorisation and Profile service to provide a single repository of identities, groupings and attributes, which could be used to unify access across many datasets and services.

Chain of tools

EMBL-EBI offers programmatic and web access to about 150 computational tools. To help our users extract greater benefit from a wider range of EMBL-EBI services, we are making it easier for users to bring their own data and store their results for future reference. We are also making it easier for users to 'chain' tools so that we can provide workflows, rather than single tools.

Data coordination services

EMBl-EBI infrastructure is applied increasingly in the area of data coordination. We ensure the smooth handling of complex data flows in research projects of all sizes. This makes it easier to share, analyse and publish experimental data during and beyond the lifetime of a project.

We build collaborative tools such as data hubs on top of our archival infrastructure. Our data portal platform provides access to structured, pre-publication, shared data. Features include:

- Quick set-up of web entry points for data coordination projects
- Discovery tools for searching shared data across the 'pre-publication' and 'published' divide
- The Analysis Archive: an agile, extensible data-management system that allows rapid configuration of new structured data types relating to sequencing

Combined with elements of our established services, such as the Webin data submission system, we provide our collaborators with advice on setting up appropriate elements of infrastructure, leadership of in-project data standards development, data coordination and curation, and user support.

European coordination

EMBL-EBI hosts the Hub of ELIXIR, the pan-European research infrastructure that aims to promote scientific progress by making it easier for scientists to find and share data, exchange expertise, and agree on best practices. ELIXIR's 21 members and 180 participating organisations make it is the largest of Europe's public research infrastructures. In its 2016 Roadmap, The European Strategy Forum on Research Infrastructures (ESFRI) recognised ELIXIR as a 'landmark' infrastructure.

ELIXIR projects

Interoperability

EMBL-EBI played a pivotal role in seven ELIXIR implementation studies in 2016, including data identification and interoperability, ELIXIR Beacons, mining the proteome, a microbial metabolism resource for Systems Biology and data resource implementations for the GA4GH.

Federated access

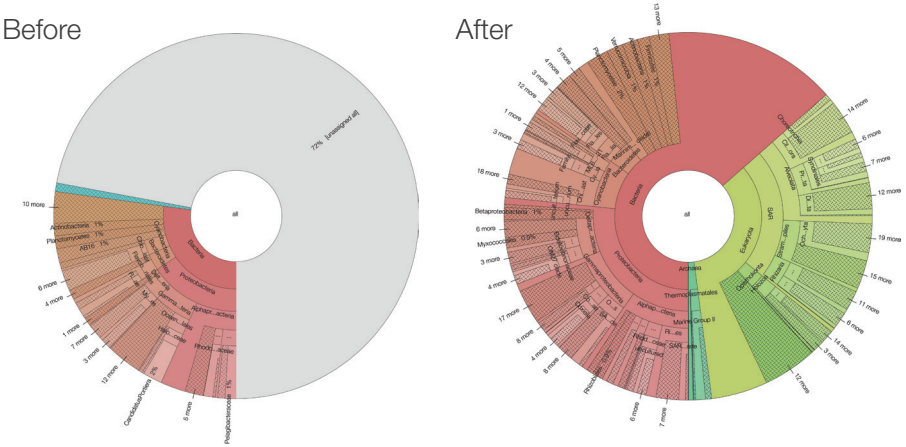
In collaboration with the Centre for Genomic Research (CRG) in Spain and ELIXIR-Excelerate, we delivered a prototype of the 'Federated access' model for the EGA. Importantly, we implemented a systematic way to record data use conditions based on consent permissions in the datasets.

In 2016 we also made EGA data more discoverable and accessible by implementing a pipeline for exporting public-access data in the EGA to the Omics Discovery Index.

A sea of metagenomics data

To enrich analysis output and improve performance, we deployed our Metagenomics analysis pipeline within Embassy Cloud, Google Cloud and Amazon Web Services. The distributed backend for the service was developed in collaboration with ELIXIR's 'Compute Platform' as a proof of concept, and will be scaled up in external cloud environments in 2017.

Benchmarking of marine metagenomics capability shows a step change in understanding, thanks to updated taxonomy reference databases and refined tools for analysis. The data shown in this figure is from an environmental sample from the Red Sea, Gulf of Aqaba (ENA project number GCA_900157355).



Communication channels

A new 'ELIXIR Reports' channel on *F1000Research* features perspectives on the selection process for ELIXIR core data resources and recommended metrics for tracking life-science research software. It also provides information about a new tool for downloading sensitive data securely from the EGA into a Galaxy server for analysis and sharing.

Knowledge exchange

We extended our annotation of human macromolecular complexes, expanding the offerings in our Complex Portal thanks to a knowledge exchange supported by ELIXIR. The collaboration involved curators in our service teams and Masters students from the Central European Institute of Technology and Masaryk University in the Czech Republic.

Standardising plant data

In 2016 we intensified efforts to develop standard metadata requirements for Ensembl Plants, engaging with the community through collaboration in the context of ELIXIR.

ELIXIR Excelerate: Compute Platform

We deployed two GridFTP servers to support the ELIXIR community. These servers are being used to understand better the 'heartbeat' of file-transfer performance across the network. At regular intervals, a file transfer is initiated between GridFTP servers at participating ELIXIR Nodes to build a mesh of interactions.

This service is hosted within the Embassy Cloud. Real-time analytics are based on direct input of data from the Heartbeat application.

ELIXIR implementation studies at EMBL-EBI in 2016

- Identifiers.org: Data Identification and Interoperability
- Data Resource Implementations for the GA4GH Data Schema
- ELIXIR Beacons 2017
- Towards a Distributed Ensembl
- Mining the Proteome: Enabling Automated Processing and Analysis of Large-Scale Proteomics Data
- A Microbial Metabolism Resource for Systems Biology
- Bioschemas: making data in the BioSamples database (annotated using the Ontology Lookup Service) more easily searchable and indexable by search engines such as Google

Staff and alumni

All of the contributions to science and technology in this report were possible because we are a hub for bioinformatics professionals all over the world.

Because we recruit internationally, our culture represents a broad diversity of perspectives and working practices. In 2016 our 547 members of staff represented 64 countries. We welcomed 108 new members of staff and fellows, many of whom relocated to work with us. We also benefitted from the contributions of visitors who worked with us on specific projects (total: 141 visitors from 34 countries).

EMBL's built-in turnover scheme aims specifically to build capacity beyond our buildings. Its long-standing, 'opt-in' Alumni Programme helps former staff connect with one another and develop new collaborations. Of the 66 professionals who left EMBL-EBI in 2016, 45 joined the EMBL Alumni Programme and so remain a valued part of our community.

New leadership

In 2016 we welcomed three new team leaders to support and guide the development of our public data resources.

Andrew Leach joined EMBL-EBI from to lead our chemogenomics resources, which are unique in the public domain and heavily used in commercial research and development. Previously, Andrew was Global Head of Biomolecular Sciences at GSK Research and Development.

EMBL-EBI
Staff in 2016

We are proud of our internationality. In 2016 our 547 members of staff hailed from 64 nations. Of these, 395 were from EMBL member states.

Thomas Keane joined EMBL-EBI from the Sanger Institute's Sequence Variation Infrastructure Group. In his new role, he will lead EMBL-EBI's development of infrastructure for the European Genome-phenome Archive (EGA), which is a joint project with the Centre for Genomic Research (CRG) in Barcelona, Spain.

Jonathan Hickford joined us from RedGate Software to lead the Web Development team, which promotes and coordinates user-experience research and design to ensure our offerings are accessible to scientists working in diverse settings.

New research group leaders **Zamin Iqbal** and **Evangelia Petsalaki**, recruited in 2016, began their work at EMBL-EBI in 2017 (see page 26).

Alumni Profile: Catalina Vallejos, from postdoc to group leader

As a postdoctoral researcher at EMBL-EBI and the MRC Biostatistics Unit, Catalina developed statistical methodologies for capturing technical noise in single-cell sequencing. After finishing her postdoc in October 2016, Catalina joined the Alan Turing Institute, the national institute for data science, as a Research Group Leader.

"I joined EMBL-EBI after following a traditional degree in statistics," explains Catalina. "All of a sudden, I was working with people from very different backgrounds and I could see how my methodologies fit into the bigger picture. That's when it all clicked, and I started thinking about how my work could be used to solve real-life challenges.

"Now I'm taking this approach even further. My group will use data analysis for predictive maintenance – this is all about being able to anticipate when something is likely to break, and fixing it before it does. We envisage building methodologies that could be used for a range of applications, from gas turbines to personalised medicine.

"My EMBL-EBI postdoc was an incredible opportunity to work with interesting people who are top experts in their field. It opened my mind to new possibilities. My job is now in London, but I still live in Cambridge and collaborate with my two postdoc groups as a mentor. It's amazing to see people building on your work to make new discoveries."

Funding facts and figures

We are grateful for the continued support of our member states and other funding bodies, which in 2016 helped us retain staff, maintain our core public resources and, thanks to additional support from the UK Government, absorb the doubling of the data we store in our archives.

This figure shows sources of EMBL-EBI funding in 2016. EMBL-EBI is part of EMBL and is funded by its member states. Sources of external funding include the Wellcome Trust, UK Research Councils, the European Commission, the US National Institutes of Health, and our Industry Programme (€0.6 million). We receive smaller grants from various other sources. A special contribution is made by the UK Government's Large Facilities Capital Fund (LFCF).*

Sources of funding

Our funding in 2016 was €73.2 million, of which 5.1 million was passed through to grant subcontractors.

Funding for EMBL-EBI comes primarily from EMBL member states, which in 2016 was €40.3 million (59%; see figure on opposite page).

We receive external funding in the form of grants (€27.8 million in 2016, excluding sums for subcontractors). Our major sources of external funding include the Wellcome Trust, UK Research Councils, the European Commission, the US National Institutes of Health, and our Industry Programme. We also benefit from a large number of grants from other sources.

Special contributions

We receive additional contributions from the UK Government through its Large Facilities Capital Fund (LFCF). This has provided for the EMBL-EBI South Building, which houses the ELIXIR technical hub and an Innovation and Translation Suite. It also provides for the on-going use of Tier-IV data centres and the equipment necessary to enable data service provision. In 2016, these additional contributions to EMBL-EBI were €9.6 million.

Expenditure

Our largest expenditure in 2016 was staff (69%), followed by operating costs (27%) and capital expenditure and depreciation (4%).

With thanks to our funders

Our funders make it possible for us to deliver open data resources, perform excellent research and provide professional training for the global scientific community. In 2016 our funders included:

- Biotechnology and Biological Sciences Research Council
- British Heart Foundation
- Cancer Research UK
- European Commission
- European Molecular Biology Organisation
- European Research Council
- Foundation of the National Institutes of Health
- Bill & Melinda Gates Foundation
- Human Frontier Science Program
- Innovate UK
- International Policy Governance Organization
- Medical Research Council
- National Institutes of Health
- National Science Foundation
- Parkinson's UK
- Research Councils UK
- Wellcome Trust

Looking ahead

When our users can make the best use of the tremendous variety of resources we offer, everyone wins. To serve our diverse user community, we are focusing on gaining efficiencies by unifying our data services, providing better tools for data deposition, and increasing access to and re-use of biological research data.

We will create innovative portals that get researchers straight to the data they need, whether it's through programmatic access or user-experience-designed web interfaces. We will continue our tradition of building sophisticated tools for analysis, and providing robust infrastructure to support collaboration across borders.

We will empower the next generation of scientists by expanding our formal training, both through career development in world-class computational biology research and in our growing bioinformatics Training Programme.

Building on our relationships with bioindustries, we will pursue new opportunities for collaboration with companies in fields spanning healthcare, biotech and agri-food.

Our ultimate goal is to help researchers in all disciplines, across sectors, turn data into knowledge, and knowledge into sustainable solutions for society's most pressing challenges.

Serving our user communities

Molecular data opens up huge potential to create clinically useful treatments and diagnostics, to the point where basic and translational research now overlap. In the coming years, we will build bridges between biological information and clinical data to develop medically relevant data infrastructures.

We will engage with medical communities in member states and beyond to exchange knowledge and best practice, enabling the translation of multi-omics data and ensuring our provision of data, information and training are fit for purpose for these new users.

Securing the global food supply in the era of climate change demands a deeper understanding of genetic diversity in plants, life within the soil, and the species that endanger our crops. The open data and collaboration we champion are critical to timely success in these fields, in which genomic data is being produced at a far greater scale and rate than ever before.

We aim to connect agri-food researchers directly to the most important, relevant information for their research, share it swiftly and analyse it appropriately as new technologies emerge and mature. We also aim to make existing resources, such as seed banks and phenotyping centres, increasingly data rich and accessible. By doing so, we will facilitate the design of more adaptable staple foods and effective disease interventions.

Open data, effective ontologies and well-designed data pipelines – all hallmarks of a healthy bioinformatics ecosystem – offer massive efficiencies for biotechnology. We will continue to engage with researchers seeking to harness biological processes to produce antibiotics, alternative fuels and other useful products, ensuring our public bioinformatics services can support advances in this innovative sector.

Expecting the unexpected

Over the next few years, we expect to see exponential growth in the generation of human, model-organism and microbial data from high-throughput genomics and high-dimensional imaging studies. Improvements in the performance and portability of these technologies, and their increasing affordability, open doors to molecular exploration and analysis well beyond the confines of the laboratory.

With burgeoning applications in biomedicine, agriculture, marine environments and biotechnology, sequencing is already transforming healthcare, food production and climate research. Combining these invaluable molecular data with multidimensional bioimaging has the potential to bring about further radical tranformation in our understanding of health and disease.

The bottleneck will be, as ever, data analysis.

We will continue to create, in collaboration with academic and commercial researchers, bespoke bioinformatics algorithms and analysis pipelines to facilitate discovery and development.

We will foster collaboration between basic, curiosity-driven research and clinical, healthcare and commercial R&D in all life-science domains. By building on the strengths of many approaches and facilitating knowledge exchange globally, we will enable the translation of novel ideas to solutions that benefit humankind.

EMBL-EBI Leadership in 2016

Pedro Beltrao group

Ewan Birney group

Paul Flicek group

Nick Goldman group

John Marioni group

Julio Saez-Rodriguez group*

Christoph Steinbeck group*

Janet Thornton group

Paul Bertone group*

Anton Enright group

Moritz Gerstrung group

Zamin Iqbal group (2017)

Evangelina Petsalaki group (2017)

Oliver Stegle group

Sarah Teichmann group*

Molecular Archival Resources
Helen Parkinson

European Nucleotide Archive
Guy Cochrane

EGA & Archive Infrastructure
Thomas Keane

RESEARCH GROUPS
**Visiting group in 2016*

SERVICE TEAMS

EXTERNAL-FACING ACTIVITIES

EMBL-EBI Leadership in 2016

Pedro Beltrao group

Ewan Birney group

Paul Flicek group

Nick Goldman group

John Marioni group

Julio Saez-Rodriguez group*

Christoph Steinbeck group*

Janet Thornton group

Paul Bertone group*

Anton Enright group

Moritz Gerstrung group

Zamin Iqbal group (2017)

Evangelina Petsalaki group (2017)

Oliver Stegle group

Sarah Teichmann group*

Molecular Archival Resources
Helen Parkinson

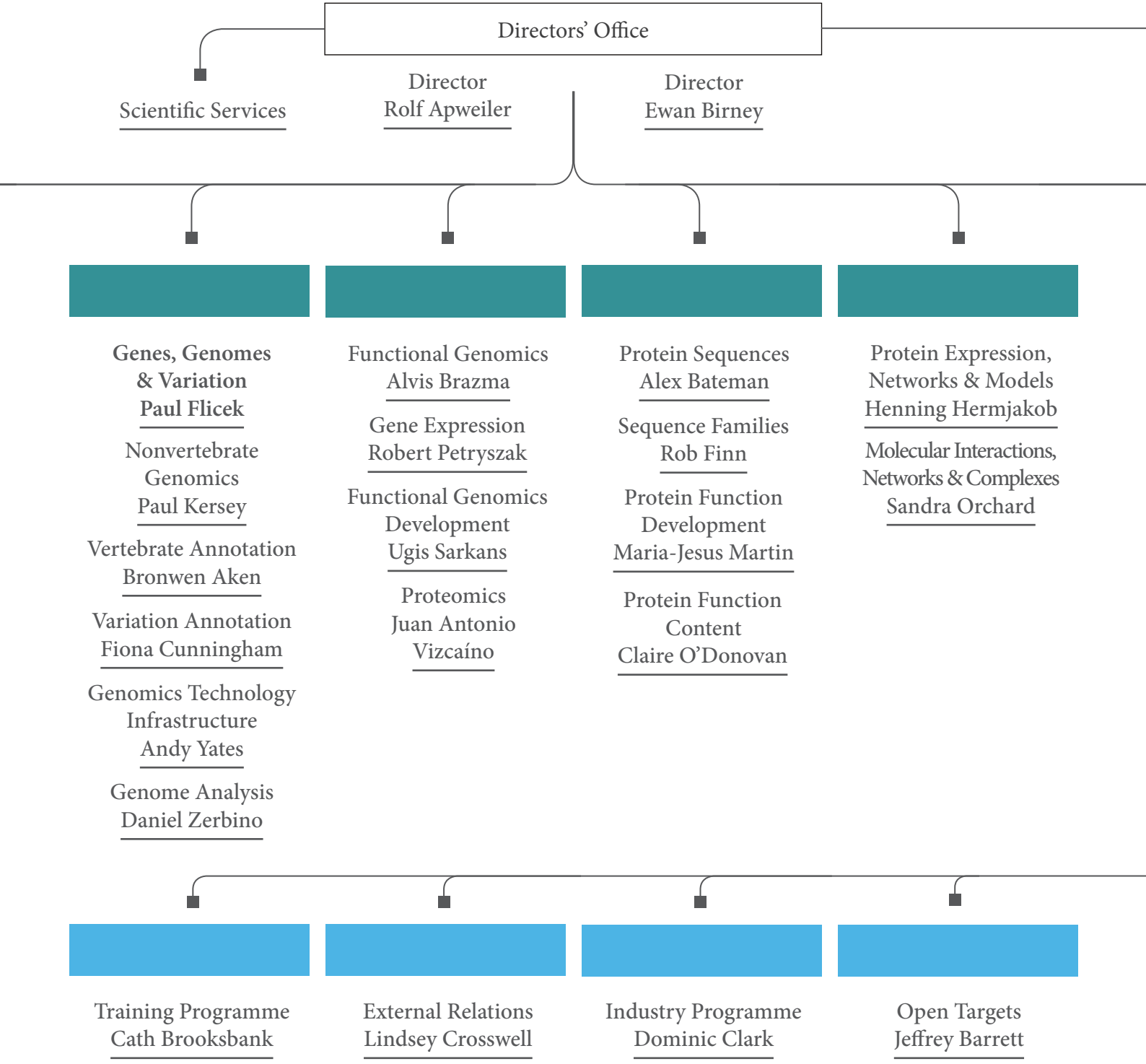
European Nucleotide Archive
Guy Cochrane

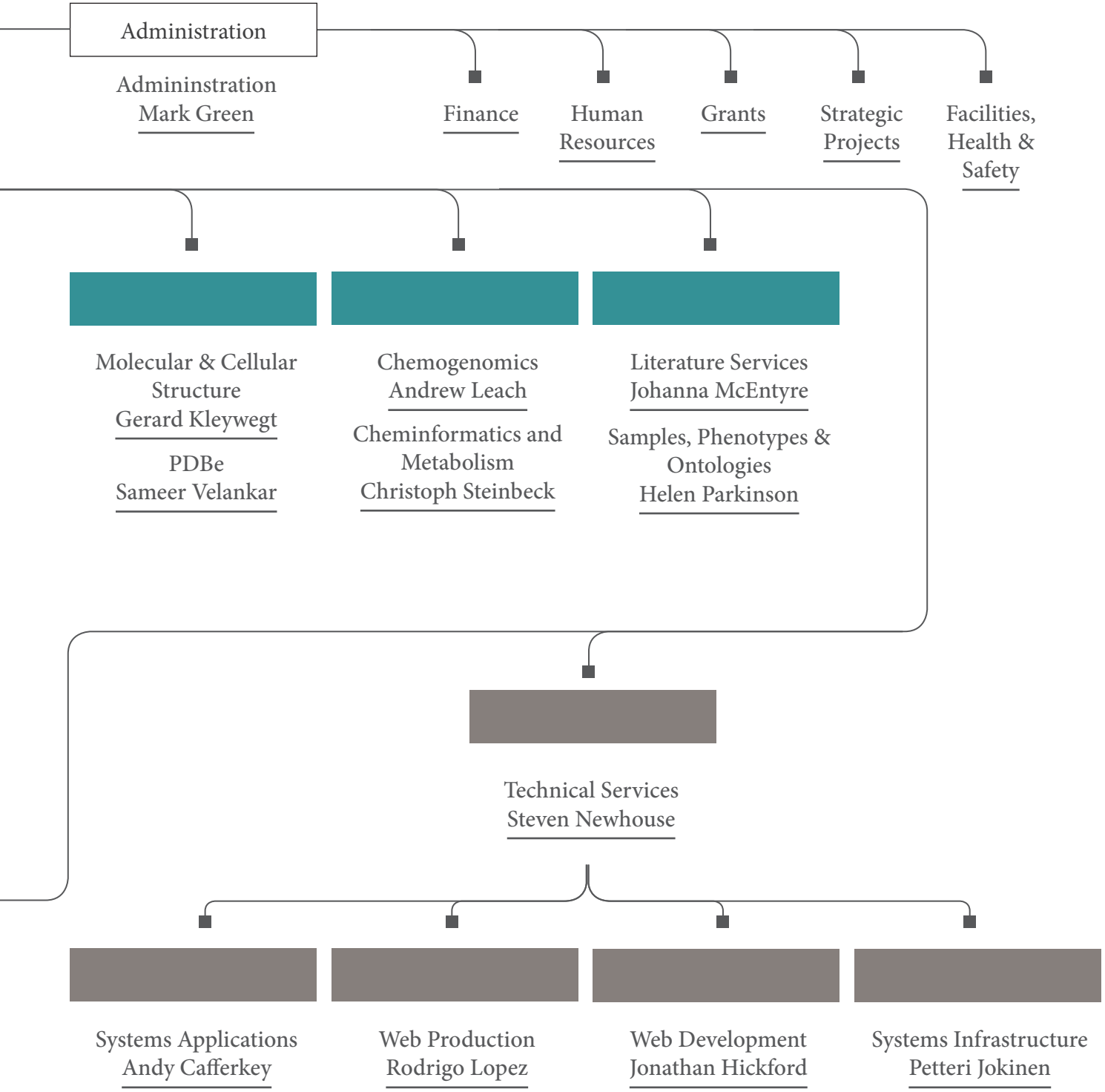
EGA & Archive Infrastructure
Thomas Keane

RESEARCH GROUPS
**Visiting group in 2016*

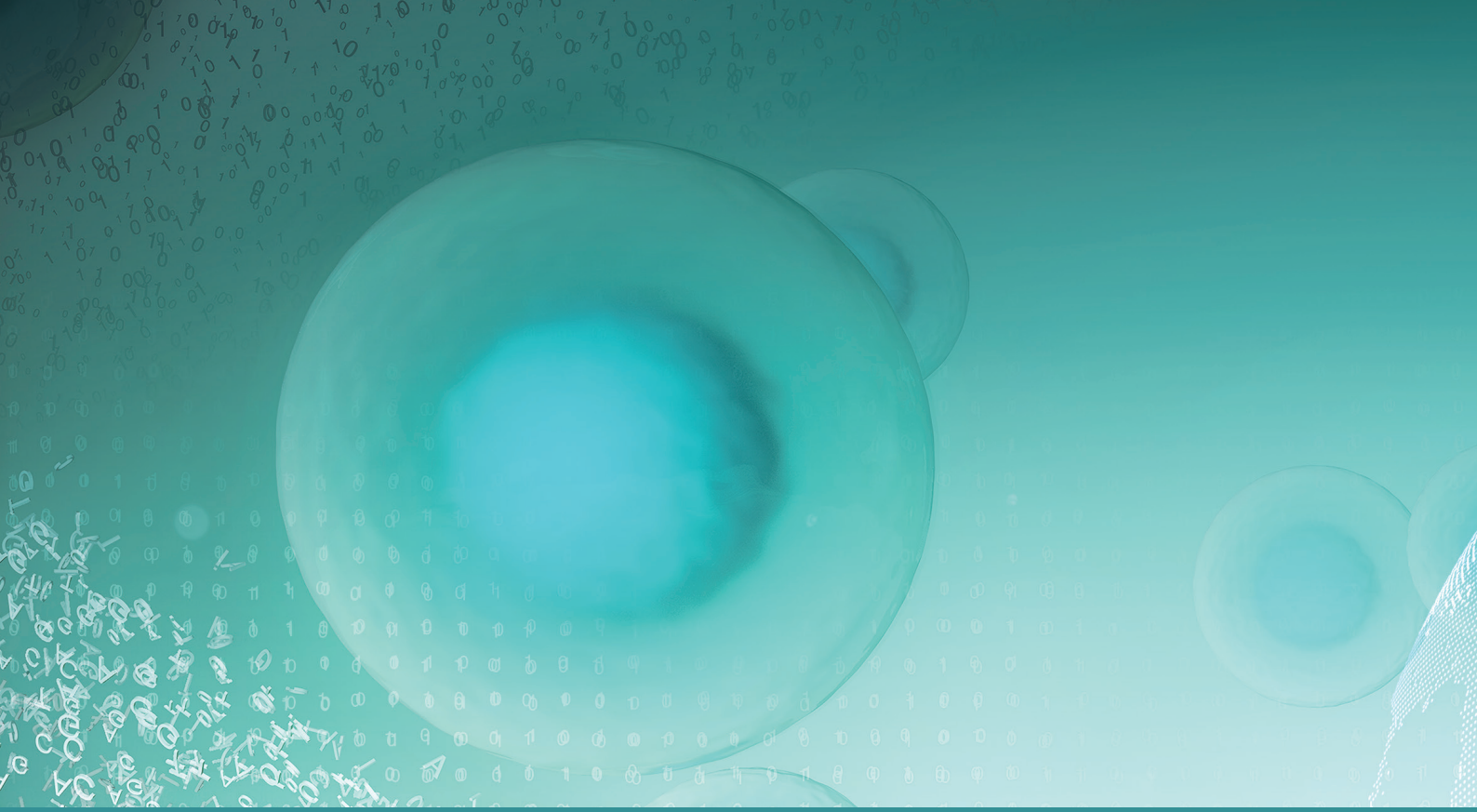
SERVICE TEAMS

EXTERNAL-FACING ACTIVITIES


Directors' Office
Scientific Services
Director
Rolf Apweiler
Director
Ewan Birney
Genes, Genomes & Variation
Paul Flicek
Nonvertebrate Genomics
Paul Kersey
Vertebrate Annotation
Bronwen Aken
Variation Annotation
Fiona Cunningham
Genomics Technology Infrastructure
Andy Yates
Genome Analysis
Daniel Zerbino
Functional Genomics
Alvis Brazma
Gene Expression
Robert Petryszak
Functional Genomics Development
Ugis Sarkans
Proteomics
Juan Antonio Vizcaíno
Protein Sequences
Alex Bateman
Sequence Families
Rob Finn
Protein Function Development
Maria-Jesus Martin
Protein Function Content
Claire O'Donovan
Protein Expression, Networks & Models
Henning Hermjakob
Molecular Interactions, Networks & Complexes
Sandra Orchard
Training Programme
Cath Brooksbank
External Relations
Lindsey Crosswell
Industry Programme
Dominic Clark
Open Targets
Jeffrey Barrett


Administration
Admininstration
Mark Green
Finance
Human Resources
Grants
Strategic Projects
Facilities, Health & Safety
Molecular & Cellular Structure
Gerard Kleywegt
PDBe
Sameer Velankar
Chemogenomics
Andrew Leach
Cheminformatics and Metabolism
Christoph Steinbeck
Literature Services
Johanna McEntyre
Samples, Phenotypes & Ontologies
Helen Parkinson
Technical Services
Steven Newhouse
Systems Applications
Andy Cafferkey
Web Production
Rodrigo Lopez
Web Development
Jonathan Hickford
Systems Infrastructure
Petteri Jokinen


MANCHESTER
University of Manchester
A14
Cambridge
Research Park
Waterbeach
Station
CAMBRIDGE
OXFORD
Wellcome Trust Centre for Human Genetics
Translational Neuroscience & Dementia Research
NORWICH
Earlham Institute
Cambridge
Science Park
St John's Innovation Centre
University of Cambridge
School of Clinical Medicine
Autism Research Centre
Alan Turing Instutute
Cambridge Cardiovascular
Department of Zoology
A14
Academic partners
Other Science parks
Cambridge
Station
ARM
Cambridge Biomedical Campus
Cambridge University Hospitals
(Addenbrookes, Papworth, Rosie)
MRC Laboratory of Molecular Biology (LMB)
Cancer Research UK - Cambridge Institute
NIHR Cambridge Biomedical Research Centre
- AstraZeneca
- GlaxoSmithKline Clinical Unit
- e-hospital Project
M11
Shelford
Station
Babraham Institute
Foxton
Station
A505
Granta Park
Shepreth
Station
Whittlesford
Parkway
Station
A11
A1301
WELLCOME
GENOME
CAMPUS
Base of
sanger
EMBL-EBI
Meldreth
Station
STEVENAGE
GlaxoSmithKline R&D
Bioscience Catalyst
LONDON
Francis Crick Institute, Imperial College,
University College London, Wellcome Trust,
LSHTM, Kings College London,
Institute of Cancer research
Royston
Station
Great
Chesterford
Station
Chesterford
Research Park
(Illumina)



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
EMBLmedia

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

EMBL-EBI is part of the European Molecular Biology Laboratory
www.ebi.ac.uk/about