

European Bioinformatics Institute · Cambridge

Scientific Report 2017

Contents

Foreword	3
Major achievements in 2017	4
Services	12
Research	84
Training	116
Industry and innovation	122
Technical Services	128
Administration and External Relations	144
Facts and figures	150
Publications, Scientific Advisory Boards and major database collaborations	158
Organisation of EMBL-EBI leadership in 2017	176

On the cover

The illustrations in this publication, including the cover artwork and the chapter images, were created by Spencer Phillips in EMBL-EBI’s External Relations team, with scientific input from Matthew Conroy in the Protein Data Bank in Europe team. The artwork is inspired by the rise in single-cell sequencing, and the institute’s involvement in the Human Cell Atlas project. The initiative aims to create comprehensive reference maps of all human cells to understand human health and disease.

EMBL-EBI, together with the Broad Institute and the UCSC Genomics Institute, and supported by the Chan Zuckerberg Initiative (CZI), is building a cloud-based data coordination platform for the Human Cell Atlas.

© 2018 European Molecular Biology Laboratory

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

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



Foreword

EMBL-EBI is one of the six sites of the European Molecular Biology Laboratory (EMBL), an intergovernmental organisation that sits at the heart of the life science community. We are located on the Wellcome Genome Campus, near Cambridge in the UK, and provide sustainable infrastructure and tools for data-driven science. Like all EMBL sites, we place great value on world-class research in the biomedical, agricultural, environmental sciences and beyond.

In 2017 we swept into a new era of scientific discovery, which saw us tackle even greater volumes of biological data submissions using innovative technologies and data strategies. To accelerate science, we applied robust and scalable ways to create common data repositories and references, enable creative exploration and further open data.

Common references

Our value and utility in building infrastructures that allow for purposeful data aggregation came to fruition in 2017 with the launch of the Human Cell Atlas. This unprecedented collaboration aims to generate comprehensive maps of all human cells in order to understand health and disease. Within the project, we embarked on building the first data coordination platform for single cell profile data. This, as well as our participation in other landmark initiatives such as the Global Alliance for Genomics and Health, and our knowledge management of growing data types such as metagenomics and bioimaging, reimpress the unparalleled value of collecting, analysing and disseminating data for reuse in research.

Enabling creative exploration

The increasing availability of heterogeneous biological data in large volumes naturally lends itself to diverse research questions. To meet these needs, EMBL-EBI developed numerous methods and interfaces that make data more discoverable and give scientists the tools to handle multiple data types. Our discovery solutions transcend scientific boundaries, bringing together biologists, clinicians, physicists, mathematicians and software engineers. By connecting diverse skills and expertise, we are enabling the creative exploration of complex questions in biology today.

Furthering open data

Open data spurs the advancement of new science and facilitates coordinated research. Data from projects such as UK Biobank are helping us understand human health at the molecular level. This science is undertaken in collaboration with global partners and is an essential part of who we are. We enhance bioinformatics capacity through research, training, knowledge exchange and consensus building. In doing so, we hope to make the scientific community stronger, and ensure the long-term security of research results for future generations.

Sincerely,

Rolf Apweiler, Joint Director

Ewan Birney, Joint Director

Major achievements in 2017

2017 was a year of rapid growth on all levels, including staff numbers, data storage, data usage and user communities. This sustained increase demonstrates the growing data needs of the life science community around the world, and the new ways in which researchers are using bioinformatics to accelerate science. This Scientific Report serves as a snapshot of how EMBL-EBI contributed towards accelerating science and technology in 2017.

In 2017, EMBL-EBI staff numbers grew by approximately 10%, reaching 604 full-time equivalents (FTEs). We welcomed two new research group leaders, Evangelia Petsalaki, whose group focuses on human cell signalling in healthy and disease conditions, and Zamin Iqbal, who works on understanding genetic variation in microbes, and develops methods for exploring outbreak surveillance and diagnostics to tackle antimicrobial resistance.

Our data resources also welcomed several new team leaders. The Protein Families team, led by Rob Finn, has reshaped itself as the Sequence Families team, to reflect its continuing focus on both protein and non-coding RNA. Claire O'Donovan moved into the role of Team Leader for Metabolomics, Sandra Orchard is now a Team Leader for Protein Function Content, and Ardan Patwardhan is the Team Leader for Cellular Structure and 3D Bioimaging.

Our impact

Several major projects and collaborations dominated the year, all of which reflect the growing needs of life science researchers for common references, creative exploration and open data.

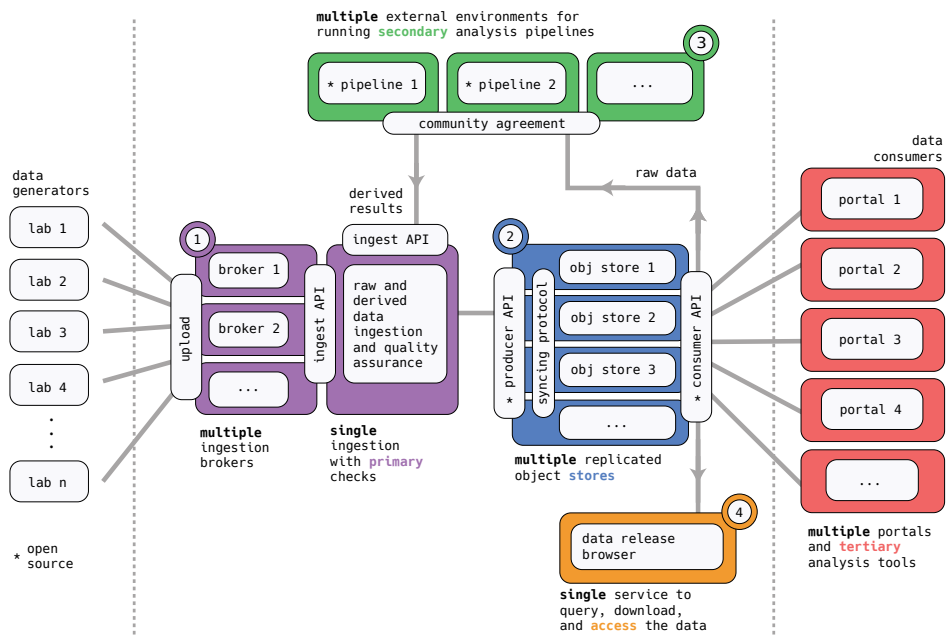
We welcomed the launch of the Human Cell Atlas (HCA), a collaborative data-intensive endeavour to chart the specific genetic properties of all human cells, and build a new reference map of the healthy human body. Funded by the Chan Zuckerberg Initiative, the project aims to make the resulting data open and easily accessible to all researchers, enabling the scientific community to innovate rapidly, without barriers. EMBL-EBI will work alongside the Broad Institute and the University of California Santa Cruz Genomics Institute to set up an open, cloud-based Data Coordination Platform (DCP) to curate, analyse and share the terabytes of data generated by hundreds of labs around the world for this project.

In 2017, the first version of the Human Cell Atlas DCP was deployed for the user community to submit, analyse and access data. The platform is scheduled for production release in October 2018 and will contain community datasets located on a cloud-hosted data store and analysed using standardised pipelines.

Another unprecedented data-sharing project featured UK Biobank, which now holds health information on over 500 000 individuals. In 2017, UK Biobank shared its genetic data via the European Genome-phenome Archive (EGA), a joint resource developed by EMBL-EBI and the Centre for Genomic Regulation. Access to this data offers endless possibilities and substantial efficiency savings for biomedical research and for understanding the causes of disease. In its first few weeks of activity, more than 300 researchers across 139 institutes requested access to genetic data from the UK Biobank.

Partnering with UK Biobank leverages EMBL-EBI's infrastructure investment and our work on international standards for genomic medicine through the Global Alliance for Genomics and Health (GA4GH), an international, nonprofit alliance that enables genomic data sharing to advance human health. In 2017, GA4GH was reorganised to focus on 15 driver projects and eight work streams. EMBL-EBI is involved in three of these driver projects through its Molecular Archives teams (ENA, EGA and EVA), membership in ELIXIR and collaboration in the Human Cell Atlas DCP.

Diagram of key components of the open-source HCA data coordination platform



Data resources

At the close of 2017, EMBL-EBI reached 155 petabytes (PB) of data storage capacity, up from 101 PB at the end of 2016. The daily web requests to our data resources averaged just under 38 million per day, an increase from 27 million in 2016.

In an effort to anticipate the research community's data needs, EMBL-EBI continued to grow its capacity for existing data types and initiated several projects to accommodate new data types, which show increased research and analysis potential for the life science community.



Molecular Archives

Our Molecular Archives teams support several core data resources, including the European Nucleotide Archive (ENA), one of the first universal data repositories in molecular biology. In 2017, the Molecular Archives teams increasingly supported high-throughput phenomics projects for human and model species, such as the International Mouse Phenotyping Consortium (IMPC) and Virtual Fly Brain (VFB). In 2017, Molecular Archives delivered data management, coordination and analysis for a wide range of projects from bacterial species to human, including the Human Cell Atlas project.

The assignment of long-term stable identifiers to genetic variants (rsIDs) is essential for the scientific community because it allows researchers to share known variants. From September 2017, EMBL-EBI's European Variation Archive (EVA) will maintain reliable accessions for non-human genetic variation data, enabling a more rapid turnaround for data sharing in this burgeoning field.

Ontologies help researchers everywhere to find the data they need, but different organisations use different ontologies, which can be confusing and inefficient. Our new Ontology Xref Service (OxO), helps users map different ontologies to each other, allowing for more creative exploration of data. In 2017, we also developed a new mapping prediction algorithm that can map private vocabularies from industry to public ontologies in the areas of disease and phenotype.

As data submissions to our resources continue to grow, the Molecular Archives Cluster is leading a project to unify data submissions through a new portal and API — the Unified Submissions Interface (USI). The USI will improve user experience by reducing submission complexity and will result in faster submission and submission processing.

Genes, Genomes and Variation

In 2017, Ensembl welcomed the arrival of multiple vertebrate genome assemblies using newly updated approaches for genome annotation and comparative genomics. For example, the teams used a new streamlined genome annotation pipeline for 18 primate species and 15 rodent species.

The Non-vertebrate Genomics team updated the genome assemblies of several crucial species including barley, where EMBL-EBI was part of the consortium that generated the new data, and the yellow fever mosquito. The team continues to collect and integrate RNA-seq data; there are now data from over 2 500 studies in plant, vector and microbial species.

The Ensembl user base is growing significantly and its use cases are diversifying. Researchers in clinical settings, for example, often need to run their own analysis locally, to avoid sending potentially sensitive data to our servers. To help such users, the Ensembl team launched a deployable REST service, which allows users easy access to the service. In response to the growing demand from the community, in 2017, Ensembl colleagues trained over 2 000 people in 17 countries.

The GWAS Catalog introduced summary statistics, with 68 studies already featuring them, and the number is expected to grow significantly in 2018.



The yellow fever mosquito can spread dengue fever, chikungunya, Zika fever, Mayaro, and other disease agents.

Molecular Atlas

The Molecular Atlas teams laid the foundations of two new services in 2017, both of which will accommodate increasing volumes of new data types. The Single Cell Expression Atlas will be dedicated to single-cell RNA sequencing data, while the Bioimage Archive pilot project will store biological image reference datasets. Bioimaging has seen a significant increase in recent years thanks to advances in cryo electron microscopy and the development of new imaging techniques. EMBL-EBI hopes to contribute to the hosting and management of bioimaging archives, and by working alongside the community, to improve the integration of bioimaging data with other bioinformatics data.

In other news, as a result of the Molecular Atlas' participation in the Pan-Cancer Analysis of Whole Genomes (PCAWG) visualisation working group, the Expression Atlas has been designated as one of the PCAWG data portals.

Another key focus in 2017 was the development of open and reproducible proteomics data analysis pipelines, and their deployment in cloud environments. In that context, the team started the continuous integration of proteomics datasets coming from PRIDE into the Expression Atlas.

Protein and Protein Families

UniProt continues to scale in tandem with the growth of sequence data. In 2017 UniProt doubled the number of Reference Proteomes to over 10 000.

Capitalising on the importance of data visualisation, the Protein Function Development team improved UniProt's graphical visualisations, which now also include protein-protein interactions and protein subcellular location. The improvements enable users to rapidly scan the components of very complex biological systems.

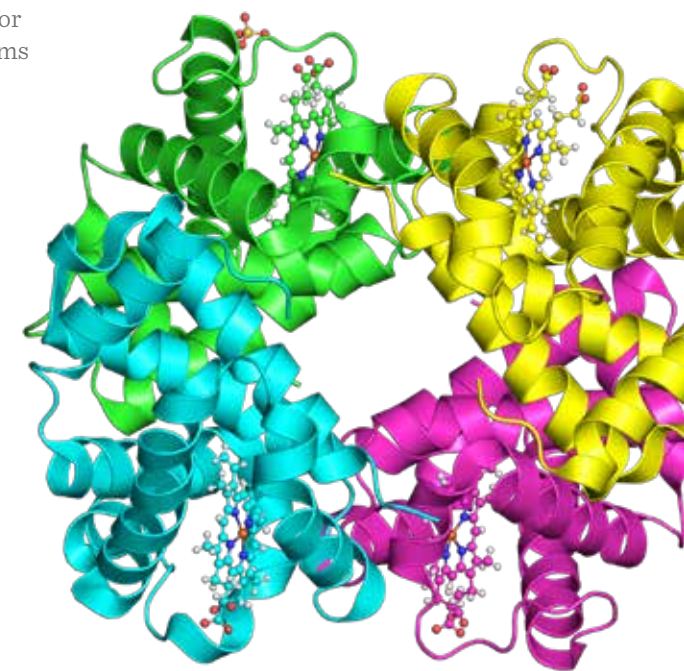
UniProt also established an exchange pipeline with ClinVar for sharing variants of clinical interest, with the aim of deciphering the functional genome and facilitating biomedical research. The team also designed a platform for interpretation of variants based on the role and functional mechanisms of proteins in disease.

One of EMBL-EBI's fastest growing data resources of the recent years, EBI Metagenomics, has also seen an increase in users. To accommodate this growth, the metagenomics analysis pipeline underwent a substantial update during which the entire taxonomic profiling section was refactored. As a result of this upgrade, the pipeline now offers clearer prokaryotic and eukaryotic taxonomic classification.

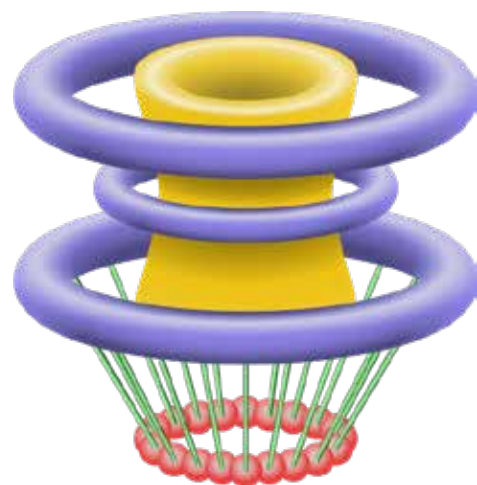
Molecular and Cellular Structure

Working alongside the other Worldwide Protein Data Bank (wwPDB) partners, the Protein Data Bank in Europe (PDBe) continued the development and management of the OneDep system for deposition, validation and curation of new molecular structure submissions. Through a BBSRC-funded project, PDBe developed a OneDep module, which serves as a prototype for supporting a federation of archives in structural biology.

PDBe also kicked off FunPDBe, a new BBSRC-funded collaborative project



The structure of the human haemoglobin molecule, showing the four subunits, as featured in PDBe



An icon from the Reactome icon library depicting nuclear pores, large protein complexes that cross the nuclear envelope, which is the double membrane surrounding the eukaryotic cell nucleus.

that will enrich structure data in PDB with manually curated and automatically predicted functional annotations, which are currently held across a large number of smaller niche resources.

The Cellular Structure and 3D Bioimaging team received funding to continue the development of EMPIAR, which will become an important component of EMBL-EBI’s upcoming BioImage Archive. To enable semantic segmentation of cellular structure data the team developed the EMDB Segmentation File Format (EMDB-SFF) in consultation with community experts.

Molecular Systems

The IntAct database of molecular interactions released a major new dataset in 2017, providing more than 15 000 instances where mutations have been experimentally shown to affect a protein-protein interaction. The dataset, curated in the context of the International Molecular Exchange consortium (IMEx), is a valuable resource in the analysis of functional effects of mutations, and in the development and testing of prediction software in the domain.

The Complex Portal has been completely redeveloped, focusing on integration with other resources, including the Expression Atlas, Reactome and PDBe.

The team also redeveloped the Reactome user interface, which now provides a multi-scale, high performance platform for the visualisation and analysis of biomolecular pathway data. Researchers can now download editable Reactome pathway diagrams to use in their publications and presentations.

The Omics Discovery Index (OmicsDI) is a brand new resource providing dataset discovery across heterogeneous transcriptomics, genomics, proteomics and metabolomics data resources spanning more than 100 000 datasets from 15 repositories in four continents. OmicsDI provides harmonised metadata across its partner repositories, allows users to “claim” datasets they have contributed to, and provides extensive links to related datasets.

Chemistry Services

Our Chemistry data resources saw significant data growth across the board in 2017. One continued area of focus was the therapeutic target and indication annotation of marketed drugs, withdrawn drugs and compounds in clinical development. Drug discovery data is increasingly complex and EMBL-EBI has undertaken significant work to extend the ChEMBL database schema to capture information from, for example, phenotypic screens, *in vivo* toxicity assays or pharmacokinetic end points in a more structured format.

The contributions to the Open Targets collaboration continued with a focus on clinical pipelines. The team is exploring methods for target tractability assessment and delivery of a practical tractability decision-making workflow for use in drug discovery.

The number of novel chemical entities annotated in the SureChEMBL patent resource stood at approximately 19 million, extracted from 18 million patents. In a pilot project for the NIH-funded Illuminating the Druggable Genome project, the team explored the scope of patent data as a source of bioactivity data and information on “underexplored” targets. The team also designed a semi-automated workflow to identify patents of potential interest for subsequent manual inspection.

As the metabolomics field continues to grow, MetaboLights surpassed 500 datasets and become a reference database for metabolomics studies and individual metabolites. In 2017, EMBL-EBI also coordinated the launch of the PhenoMeNal project, an international endeavor that aims to use data generated by metabolomics applications to improve our understanding of the causes and mechanisms underlying health, ageing and disease.

Literature Services

Our literature discovery service, Europe PMC, continued its integration with ORCID. An author search now brings up a “suggested authors” box linking to matching researchers that have an ORCID. It also displays the two most prolific researchers and links to their author profile pages.

To allow users to discover the content in new ways, the team expanded the Europe PMC programmatic tools with the Annotations API. This provides access to targeted information, text-mined from millions of biomedical abstracts and full-text articles. It allows users to retrieve, for example, all articles that discuss the involvement of a specific gene or protein in their disease of interest.

Data integration is a unique feature of Europe PMC. To highlight the link between studies and their data, the service has integrated with the BioStudies database. This way, each full article in Europe PMC has an associated BioStudies record containing supplemental data files or mentioning data identified by text-mining accession numbers for over 20 major data resources.

Research

EMBL-EBI has been a world leader in computational biology research since its inception in 1994, with work spanning fundamental methods in sequence analysis, multidimensional statistical analysis and data-driven biological discovery. Our research programmes focus on scientific discovery, as well as method development to help other researchers accelerate their science and discoveries.

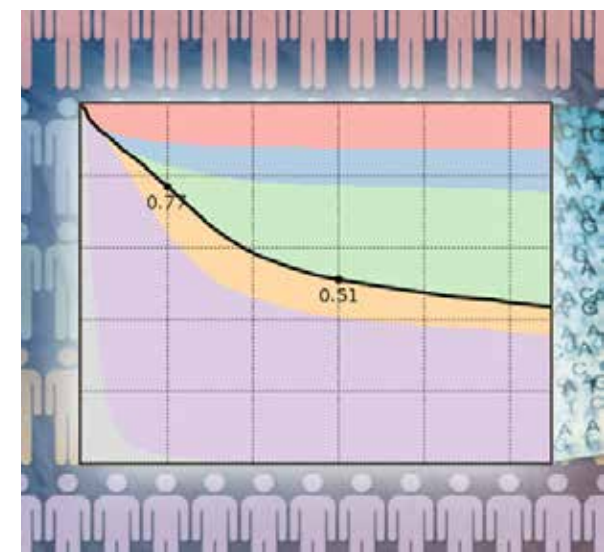
Our highly collaborative groups publish high-impact works on sequence and structural alignment, genome analysis, basic biological breakthroughs, algorithms and methods of widespread importance.

Another strong focus in 2017 was understanding health and disease development on a molecular level. An international research consortium led by EMBL-EBI group leader Moritz Gerstung has shown proof of concept that personalised therapy will be possible in the future for people with cancer (Gerstung M et al, 2017). The study, published in *Nature Genetics*, described how a knowledge bank could be used to find the best treatment option for people with acute myeloid leukaemia (AML).

In consonance with the rise of single-cell sequencing, our researchers developed a number of methods for analysing this increasingly popular type of data. In 2017, researchers in the Marioni group and collaborators shed light on a long-standing debate about why the immune system weakens with age (Martinez-Jimenez CP et al, 2017). Their findings, published in *Science*, show that immune cells in older tissues lack coordination and exhibit much more variability in gene expression compared with their younger counterparts.

The Iqbal group and collaborators at the University of Oxford developed a method for extracting *Mycobacterium tuberculosis* DNA from sputum and sequencing it directly. This reduced the sequencing time from around two weeks, most of which was spent waiting for bacteria to grow in culture, to less than 24 hours, demonstrating the potential for a portable, handheld point of care test for tuberculosis. The study was published in the *Journal of Clinical Microbiology* (Votintseva AA et al, 2017).

The Gerstung group described how a knowledge bank could be used to find the best treatment option for people with acute myeloid leukaemia (AML).





The CABANA project will focus on three challenge areas – communicable disease, sustainable food production and protection of biodiversity.

Researchers in the Stegle group have shown that the health of individual mice is influenced by the genetic makeup of their partners. The unexpected results, published in *PLoS Genetics*, found that the genetics of social partners were found to affect wound healing and body weight as well as behaviour (Baud A et al, 2017). The methods used to detect ‘social genetic effects’ encourage future research into the mechanisms whereby one individual influences another.

In 2017, eight of EMBL-EBI’s 24 PhD students obtained their degree, with the successful theses focusing on topics such as genetic analysis of molecular traits in skeletal muscle, deep neural networks, and statistical models for studying single-cell DNA methylation.

Training

EMBL-EBI delivers a comprehensive range of bioinformatics training to help the global research community keep pace with rapid technological development. In 2017, EMBL-EBI’s Training Programme celebrated its tenth anniversary. During the course of the year, over 185 staff participated in training and scientific outreach, delivering 340 training and engagement events, to support biomedical and life-science professionals around the world. This allowed us to reach more than 18 000 people face-to-face and many more online.

In support of our goal to provide training for research infrastructure staff across Europe and beyond, EMBL-EBI launched the RItrain Executive Masters in Management of Research Infrastructures and a new webinar series for operators of research infrastructures as part of the CORBEL project.

Looking beyond Europe, in 2017 we also launched the CABANA project, a bioinformatics capacity-strengthening programme in Latin America. Funded by the Research Councils UK, this project is a collaboration with nine research institutes in Latin America.

Industry, innovation and translation

Our Industry Programme continued to grow in 2017, with the addition of Celgene as a new member and the delivery of 13 workshops on up-and-coming topics in the life sciences, including predictive modelling for biomarkers, the human microbiome, single-cell RNA-seq, ontologies in agricultural research and many more.

Open Targets, the pre-competitive, public-private partnership welcomed a new member, Takeda. Open Targets uses human genetics and genomics data for systematic drug discovery identification and prioritisation. EMBL-EBI continued to play a central role in the design, development and implementation of the Open Targets Platform.

We also worked with the Pistoia Alliance, a not-for-profit alliance of life science companies, to design a free User Experience for Life Sciences (UXLS) toolkit. The toolkit, set to launch in 2018, will enable companies to benefit from user experience (UX) to design better digital products for the life sciences industries.

European coordination

EMBL-EBI is an active partner in the ELIXIR infrastructure, which aims to coordinate life-science resources throughout Europe so that they form a single infrastructure. In July 2017, ELIXIR announced the selection of the first set of ELIXIR Core Data Resources.

These recommended life science data resources are of fundamental importance to the worldwide life-science community and to the long-term preservation of biological data.

EMBL-EBI actively participated in the development of the criteria for selecting ELIXIR Core Resources, and 13 EMBL-EBI resources were included in the initial set of ELIXIR resources. ELIXIR also compiled a list of archival resources that are recommended for deposition of experimental data. Such archival resources are vital to the scientific community for ensuring that experimental data are available for re-use by the worldwide community. Twelve EMBL-EBI resources were named as ELIXIR Deposition Databases.

ELIXIR Core Data Resources

⊙ <i>Array Express</i>	⊙ <i>IntAct</i>
⊙ <i>ChEBI</i>	⊙ <i>InterPro</i>
⊙ <i>ChEMBL</i>	⊙ <i>PDBe</i>
⊙ <i>EGA</i>	⊙ <i>PRIDE</i>
⊙ <i>ENA</i>	⊙ <i>UniProt</i>
⊙ <i>Ensembl</i>	
⊙ <i>Ensembl Genomes</i>	
⊙ <i>Europe PMC</i>	

ELIXIR Deposition Databases

⊙ <i>ArrayExpress</i>	⊙ <i>IntAct</i>
⊙ <i>BioModels</i>	⊙ <i>MetaboLights</i>
⊙ <i>BioSamples</i>	⊙ <i>PDBe</i>
⊙ <i>BioStudies</i>	⊙ <i>PRIDE</i>
⊙ <i>EGA</i>	
⊙ <i>EMDB</i>	
⊙ <i>ENA</i>	
⊙ <i>EVA</i>	

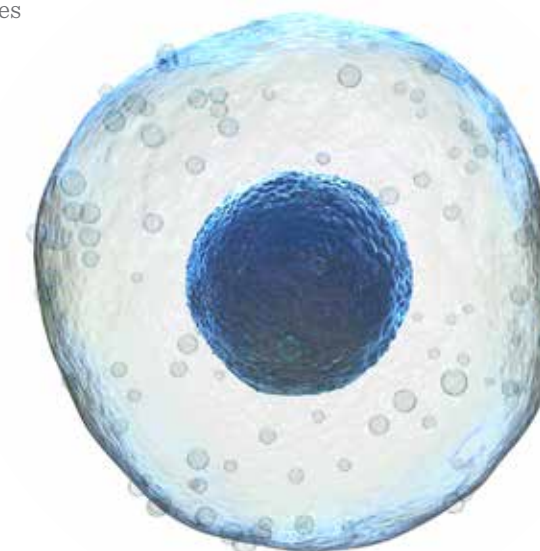
Looking ahead

In 2018, we aim to focus more on molecular and cellular function and drawing knowledge from data. We aim to enable researchers to creatively explore our open data resources to ask essential questions in the life sciences arena. We are constantly developing new tools for data analysis and interpretation, to accelerate the way we do science.

To do so, we are increasing the scope and reach of our training programme and our industry collaborations in a range of bioscience areas. We are also creating sustainable open access resources for new data types such as imaging and single-cell sequencing data. As we integrate existing and new resources, we are also creating sophisticated models and tools for interpreting complex multi-omics and phenotype data, which could pave the way for a whole new approach to bioinformatics.

Large-scale collaborations mapping biology, such as the Human Cell Atlas and the Pan-Cancer Analysis of Whole Genomes are increasingly looking to EMBL-EBI for expertise in data coordination. Such projects could become the backbone on which today’s scientists build their research and accelerate scientific discovery.

By contributing our experience in handling large, complex data sets and creating the next generation of analysis tools, we are hoping to shape the way discovery happens not just in the lab, but also by a computer. Data play a central role in tomorrow’s life sciences, so making sure we have robust, open and easy-to-use data resources and tools is more important than ever before.



Services



Cardiomyocyte

The cardiomyocyte is a specialised type of muscle cell, that can be found within the heart. It is packed full of mitochondria to ensure that it never tires as it has to work continuously. Cardiomyocytes beat spontaneously, even in a petri dish, but in the heart they all beat together, controlled by pacemaker cells.

Molecular Archives

The Molecular Archives Cluster (MAC) includes the groups of Guy Cochrane, Thomas Keane and Helen Parkinson, which collectively deliver EMBL-EBI's highly utilised data deposition databases: the European Nucleotide Archive (ENA), the European Variation Archive (EVA), the European Genome-phenome Archive (EGA) and the BioSamples Database (BioSD).



These resources provide international archiving and data access for millions of DNA sequences, genetic variants, and biological sample records for the international scientific community. MAC activities are supported by 85 personnel and 38 externally funded grants.

In 2017 MAC received the largest number of online data requests across EMBL-EBI and received more than 300 000 data submissions. In addition to serving data directly to users, the cluster's archival data are processed and presented to added-value databases at EMBL-EBI. For example, ENA genome sequence assemblies are processed by Ensembl via UniProt and the Metagenomics Portal.

MAC provides semantics as a service to many data resources at EMBL-EBI and worldwide. This includes tools such as the Ontology Lookup Service and the Ontology Cross Reference Service, which provide access to ontologies and ontology cross-references in support of the FAIR principles (Findable, Accessible, Interoperable and Reusable) for both academic and industry use.

MAC increasingly supports high-throughput phenomics projects for human and model organisms such as the International Mouse Phenotyping Consortium (IMPC) and the Virtual Fly Brain (VFB). MAC provides data management and data coordination for numerous projects, ranging from bacterial species (e.g. the COLlaborative Management Platform for detection and Analyses of (Re-)emerging and foodborne outbreaks in Europe - COMPARE project) to human (e.g. Human Cell Atlas). MAC also develops standards, new systems and architecture underlying data access (e.g. in the Global Alliance for Genomics and Health - GA4GH) including cloud-based data access.

We faced two major challenges in 2017. Firstly, a rising volume and complexity of data, and the need to develop strategies for engagement with data brokers and national initiatives to ensure smooth data flow and access at petabyte scale. Our work with UK Biobank to distribute array-based data via EGA and the microbial research communities represents significant progress in this regard. Secondly, we are working to verify our data are 'FAIR' for our core resources and grant funded projects. Our work with the user and data generators for Human Cell Atlas, Pistoia Alliance, GA4GH and ELIXIR addresses this challenge in development of tools, standards and APIs for the user community.

In 2018/19 we will collaborate with ELIXIR colleagues to assess the applicability of the FAIR principles to ELIXIR data resources funded by an ELIXIR Implementation Study.

Image by Rich Gemmell illustrating an artist's interpretation of the Human Cell Atlas. Designed for the 2018 Pint of Science festival.

Major achievements

Human Cell Atlas

Data Coordination Platform

The Human Cell Atlas is organising and standardising terabytes of data for billions of cells, across multiple modalities, generated by hundreds of labs around the world. We want to make data open and easily accessible to all researchers, enabling the scientific community to innovate rapidly without barriers to data access.

We also want to make it easy for computational researchers to develop and share new analysis approaches. To do this, we are developing a cloud-based Data Coordination Platform in collaboration with the University of California Santa Cruz, the Broad Institute, and the Chan Zuckerberg Initiative (CZI). The project is funded by CZI.

Unified Submission Interface (USI)

Data submissions to the Molecular Archives Cluster resources continue to grow and increasingly involve parallel data submissions to multiple repositories from single and multiexperiment volumes. The cluster is therefore leading a project to unify data submissions through a new portal and API – the Unified Submissions Interface (USI). The USI will improve user experience by reducing submission complexity and will result in faster data processing.

We have already consolidated the archival database help desks to a single system and have implemented single sign on, using technology developed within ELIXIR and in collaboration with EMBL-EBI's Technical Services Cluster. To enable cross-archive submission we have developed a canonical data model and produced a system to receive user submissions and broker them in to the archive databases. We are in the process of developing a web user interface (UI) and extending our data validation infrastructure to support flexible metadata standards. We currently support data submission to BioSamples and will extend coverage to support other archives through 2018 and 2019.

Industrial collaboration

In 2017, ENA renewed its agreement with Oxford Nanopore Technologies to provide technology development around data formats for storing and representing nanopore sequencing data. Through this collaboration, ENA is working with the GA4GH and the wider genomics community to standardise formats and tooling for nanopore sequencing technology, enabling improved data storage and access.



Helen Parkinson

Head of Molecular Archival Resources. Team Leader - Samples, Phenotypes and Ontologies

*PhD Genetics, 1997.
Research Associate in Genetics,
University of Leicester 1997-2000.
At EMBL-EBI since 2000.*

The cluster provides both data and semantic services to the Open Targets platform. The EVA provides a genetic variation data source applicable to target validation, which uses ClinVar as the primary data source and provides added value by filtering, quality control, and a synergistic human and automated curation process to provide a normalised and ontological representation of trait names. ClinVar was selected as the primary data source because it is one of the largest open databases of variants that includes allele level variant and disease populations information. This is complemented by a data feed from the GWAS Catalog, which together with EVA data is used by Open Targets as a source of genetic evidence for target validation.

To make it easier for researchers everywhere to find the information they need, data scientists use ontologies to define and classify terms, concepts and relationships. Ontologies also make it possible to link relevant datasets, even if they were produced in entirely different fields. Our new Ontology Xref Service (OxO) aims to help users map ontologies to one another. In collaboration with the Pistoia Alliance, a not-for-profit alliance of life science companies and academics, we are improving OxO to provide better mappings between human disease ontologies used by industry and academic projects.

UK Biobank

Genomic datasets from the 500 000 individuals participating in the UK Biobank initiative are now distributed via the European Genome-phenome Archive (EGA), a MAC resource developed jointly by EMBL-EBI and the Centre for Genomic Regulation (CRG). UK Biobank provides extremely detailed, high-quality datasets on individuals. It is an unprecedented collection that offers endless possibilities and substantial efficiency savings for biomedical research and understanding the causes of disease.

Distribution of the data via the EGA ensures long-term data security, accessibility and sustainability, which will help researchers to better understand human disease.

In its first few weeks of activity, more than 300 researchers across 139 institutes requested access to the genetic data from UK Biobank. Half a petabyte of data was transferred in the first two weeks alone and demand for the data continues to grow. Partnering with UK Biobank leverages EMBL-EBI’s infrastructure investment and our work on international standards for personalised medicines through the Global Alliance for Genomics and Health (GA4GH).

The Global Alliance for Genomics and Health (GA4GH)

During 2017, the Global Alliance for Genomics and Health (GA4GH) was reorganised around a matrix management model of 15 driver projects and eight work streams. The following EMBL-EBI resources were selected as single driver projects in recognition of their importance and interconnected requirements: the European Genome-phenome Archive (EGA), the European Variation Archive (EVA) and the European Nucleotide Archive (ENA).

MAC provides leadership for the Large Scale Genomics work stream (co-chair, Thomas Keane) and participates in almost all GA4GH work streams. In 2017, we were involved in developing htsget, a specification to provide secure streaming access to sequencing read data. ENA and EGA released a pilot implementation in 2017 alongside four other providers including the Wellcome Sanger Institute, DNAnexus, Google Cloud Platform, and Oxford University. EGA has also demonstrated secure streaming between two htsget services (EMBL-EBI and ELIXIR-Finland) and now supports htsget as a standard access method.

A challenge for users accessing data in resources such as EGA is to determine the consented uses for the data (e.g. is industry use permitted). We have addressed this by developing a Data Use Ontology

(DUO), within the GA4GH Data Use and Researcher ID work stream, to formally represent data usage requirements. DUO is available via the Molecular Archives’ Ontology Lookup Service. DUO semantically tags datasets with restrictions about their usage, making them automatically discoverable based on the user authorisation level or intended usage.

Pathogen data services

Uptake of whole genome sequencing in the public health and food safety domains continues to bring new data to the cluster’s resources. Through ongoing and new collaborations, a number of cluster resources and services have been extended and enhanced to support such uses as pathogen surveillance, outbreak investigation and the exploration of the biology of drug-resistant infections.

ENA Data Hubs provide endpoints for the structured sharing of sequence data, associated metadata and derived analysis data. Our SELECTA cloud compute offering provides comprehensive analysis workflows, such as assembly, serotyping and resistance gene calling. Pathogen Portal, a web and associated programmatic interface provides secure search and exploration of pre-publication data shared amongst collaborating groups and comprehensive public data with onward links to collaborative data visualisation services, such as Notebooks.

To help users working in antimicrobial resistance, we support a new data type across ENA and BioSamples, the “antibiogram”. This data type provides a ground truth, laboratory-measured profile of sensitivities to antimicrobials that is informing users working on methods for resistance prediction. Extensive support for users of pathogen data services through our data coordination activities includes the development of appropriate in-project and broader standards

(such as in collaboration with the Global Microbial Identifier project).

Data coordination

We continue to provide data coordination services to diverse user communities. To help leverage and improve our data resources and tools, we provide valuable support for researchers sharing, analysing and interpreting biomolecular data. Covering domains such as agricultural livestock functional genomics, marine biology, eukaryotic diversity, pluripotent stem cell research, rare disease, cancer, and pathogen genomics, our services include support for existing standards and extensions to these, data submission services, release preparation, web data portals, programmatic interfaces and tools, which operate across cluster resources such as ENA, EGA, OLS, BioSamples and Webin.

A marine “Configurator”

The archival data resources of the cluster see use not only in the late stages of scientific initiatives, where users are looking to provide permanent discoverability and reusability to their data, but also as platforms for data management of live data as they are shared amongst collaborators and subject to analysis processes and downstream interpretation.

Similarly, our standards, tools and ontologies provide technical solutions for those operating projects with ‘omics components. Through our partnership with research infrastructures, as part of ELIXIR, in the bioprospecting, biological collections, marine biological resource centres and aquaculture domains in the EMBRIC project, we have launched the Configurator service. Intended for marine scientists embarking on biomolecular investigations, the service provides expert advice on the set up of appropriate configurations of infrastructure, such as data resources, tools, standards and expert groups, from the array of offerings within the cluster and across the broader ELIXIR tools and resources portfolio. With serviced cases across a number of domains, including aquaculture, algal biotechnology and finfish genomics, we see this service not only as an offering to marine scientists, but a model for a future broader gateway into biomolecular data infrastructure across many areas of science.

Genetic variation

The data volumes from genetic variation studies continue to expand dramatically as third generation sequencing technologies enable the creation of reference genomes for non-model species. The assignment and maintenance of long-term stable identifiers to genetic variants (rsIDs), is essential for the scientific community because it allows researchers to share and communicate known variants. In 2017, the EVA

and dbSNP (NCBI) published an agreement that the EVA will maintain these long-term locus accessions for all non-human species, and dbSNP will maintain the human identifier space. Furthermore, dbSNP will no longer accept submissions of genetic variation in non-human species. As a result of this agreement, the EVA has already seen an increase in submissions, which is expected to continue in coming years.

Future plans

Ensuring the Human Cell Atlas datasets are easy to access and well described remains a significant challenge facing MAC. Over the course of 2018, expected achievements include the first release of the HCA metadata specification and several HCA data releases, providing rules for describing human single cell RNAseq data and best practice on use of ontologies in this metadata. Looking further forward, this specification will be expanded to cover imaging assays, such as spatial transcriptomics. The platform itself will enter production phase in July 2018 with a minimal dataset and we expect a much larger release of data in the final quarter of 2018.

We will continue our work on the Unified Submission Interface through the deployment of ENA and BioStudies submission support.

In 2018, we are also planning to maintain our industry collaborations, which will include a release of the Pistoia Alliance mappings and supporting publication, with the aim of improving cross ontology mappings. We will also renew Oxford Nanopore Technology funding to define standards formats for ONT data.

We are working with the UK Biobank to prepare for the next phases of the project including exome data, and leveraging EGA’s scalable infrastructure for the benefit of the user community.

The next step of our collaboration with GA4GH will be the deployment of standards for the ENA/EGA/EVA driver projects, which will enable scaled dataflow. We will also release the htsget standard providing configurable data access for genomic data and deploy the DUO data use ontology for EGA enabling users to code and query consent.

Finally, to further facilitate data coordination, we aim to develop user-friendly access points, and contribute our expertise to building capacity across the ELIXIR community. We will also pilot Data Hubs for large-scale production initiatives, such as the prospective surveillance programmes of the European Centre for Disease Prevention and Control (ECDC) enabling simpler access to data.

Image: Pathogen Portal, showing (clockwise from top left) home page with key functions, search filter selection, configuration of search output and search results.



Molecular Archives

Data resources

BioSamples

BioSD stores and supplies descriptions and metadata about biological samples used in research and development across all sectors. These are either ‘reference’ samples (e.g. from 1000 Genomes, HipSci, FAANG) or have been used in an assay database such as the ENA.

www.ebi.ac.uk/biosamples/

Experimental Factor Ontology (EFO)

The Experimental Factor Ontology is an application ontology that combines elements of open ontologies into a single application ontology which is used to annotate, query, visualise and analyse data across EMBL-EBI and external resources. EFO is the ontology for the Open Targets platform and provides a comprehensive classification of disease and phenotype.

www.ebi.ac.uk/efo/

European Nucleotide Archive (ENA)

The ENA provides globally comprehensive primary data repositories for nucleotide sequencing information. ENA content spans raw sequence reads, assembly and alignment information and functional annotation of assembled sequences and genomes. ENA data and services form a core foundation upon which scientific understanding of biological systems has been assembled.

www.ebi.ac.uk/ena

European Genome-phenome Archive (EGA)

The EGA, co-developed with the Centre for Genomic Regulation (CRG) in Spain, permanently archives and shares all types of personally identifiable human genetic and phenotypic data resulting from biomedical research projects.

<https://ega-archive.org/>

European Variation Archive (EVA)

The EVA is an open-access database of all types of genetic variation data from all species. From September 2017, support for non-human variant data archival and accessioning transitioned from NCBI dbSNP to EVA.

www.ebi.ac.uk/eva

Mouse Resources

Mouse informatics consists of the following projects: the data portal and core data archive of the International Mouse Phenotyping Consortium, and the Infrafrontier database of mutant mice.

www.mousephenotype.org
www.infrafrontier.eu/search

Ontology Lookup Service (OLS)

The Ontology Lookup Service provides web-based and API search for approximately 200 highly-used public biomedical ontologies. It is a critical service infrastructure for building interoperable data in the life sciences. OLS has multiple components, including the ontology cross reference service (OxO), which allows users to translate between ontologies based on domain specific knowledge.

www.ebi.ac.uk/ols/index



Team achievements



Helen Parkinson

Samples, Phenotypes and Ontologies team

NHGRI-EBI GWAS Catalog

- ⦿ *Delivering a comprehensive extraction and curation of GWAS from the literature*
- ⦿ *Access to full p-value summary statistics, and development of a database for integrated storage and access to these statistics*
- ⦿ *New RESTful API to support high throughput programmatic access to GWAS data enabling a variety of downstream analysis scenarios, including drug target prediction*

BioSamples

- ⦿ *New production release deployed using a microservices model*
- ⦿ *BioSamples services are now provided as Docker images for ease of development and deployment by external users*
- ⦿ *New data submission processes implemented via the Unified Submissions Interface using ELIXIR AAI to manage access and authorisation*
- ⦿ *Automated curation pipeline implemented to add ontology values*

Mouse Informatics

- ⦿ *Published four novel papers identifying novel gene-disease associations*
- ⦿ *Formed key collaborations with international consortia including PDXnet, Patient Derived Model Repository and EuroPDX*
- ⦿ *Published community-driven PDX-MI Standard*

EFO and Semantic Tools

- ⦿ *Transition of EFO to an automated build pipeline*
- ⦿ *Deployment of a new lightweight EFO website driven from the OLS API*
- ⦿ *Installations of OLS and toolkit in a number of pharmaceutical companies*



Guy Cochrane

European Nucleotide Archive team

- ⦿ *Addition of 60 million new assembled/annotated sequences and 1.5x10¹⁵ base pairs of read data, a contribution to INSDC representing 30% of global data*
- ⦿ *Tens of thousands of monthly users of the ENA Browser with tens of millions of monthly hits, often through mirrors and secondary resources*
- ⦿ *15 pathogen-related Data Hubs and launch of the Pathogen Portal*
- ⦿ *New ENA Discovery API and toolkit for rapid deployment of data presentation websites, and launch of a cloud-based computational analysis environment*
- ⦿ *12 data coordination projects and a new service, the “EMBRIC Configurator”*
- ⦿ *Operation of 14 externally-funded projects*



Thomas Keane

EGA, EVA and Archive Infrastructure team

- ⦿ *EGA successfully distributed UK Biobank data to over 300 users totalling over 0.5 PB of data*
- ⦿ *EGA distributed over 3.9PB of data via the streaming API*
- ⦿ *Upgraded the EGA download API to support the GA4GH htsget protocol*

Samples, Phenotypes and Ontologies

The team delivers collaborative community resources such as Virtual FlyBrain, the GWAS Catalog, the International Mouse Phenotype Consortium (IMPC) Database and the Human Cell Atlas. All of these projects focus on integrating complex phenotypic data with high throughput omics data, and increasingly use cloud technologies and containerisation approaches to provide portable solutions for FAIR data access and analysis. These approaches are aligned with our delivery of the ELIXIR core data deposition database: BioSamples, an archival database for sample information at EMBL-EBI.

Major achievements

Our data resources continue to grow both in data volume, complexity and usage. For example, at the end of 2017, the GWAS Catalog contained 4993 curated studies from 3,308 publications and 58 438 SNP-trait associations with users in over 160 countries. Open Targets and NHGRI have funded an extension of remit of the Catalog to store full p-value summary statistics for studies and we encourage authors to deposit these with us. New visualisations for the Catalog include trait pages, for visualising all known single variant associations for a particular disease or phenotypic trait, and variant specific pages, for exploring the phenotypic associations from an individual locus. A new API for the Catalog supports richer data access allowing phenotype specific or pleiotropic meta-analysis, functional analysis, drug target prediction and fine mapping.

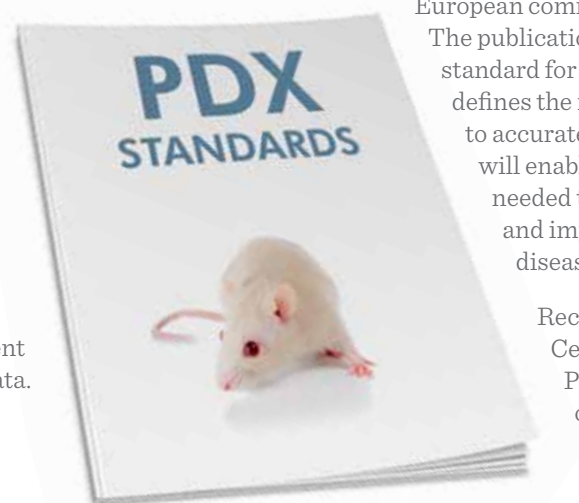
In 2017, we successfully concluded the Innovative Medicines Initiative (IMI) EBISC project, developing sustainable links between the hPSCreg project, a community induced Pluripotent Stem Cell database (iPSC) to ensure long-term data access using existing databases. The use of BioSD identifiers allows these to be tracked throughout the literature in future and linked to future publications and sequence datasets in open and controlled access databases. We have invested in a new containerised version of BioSD, allowing controlled access resources such as the EGA to deploy BioSD's ontology integration tools, metadata standards and GA4GH metadata APIs. We also now use the ELIXIR Access, Authentication and Identification infrastructure. This allows users to login to multiple resources without the need to use different user credentials for different services enabling faster access to data.

Four International Mouse Phenotyping Consortium (IMPC) publications demonstrated the prevalence of sexual dimorphism across phenotypes, identified hundreds of new models for rare genetic diseases, and discovered dozens of new links between genes and hearing or metabolic disorders. To better understand the translational impact of IMPC resources, we also worked with Literature Services to identify 1500 publications using IMPC mice or data across a wide spectrum of research fields including behaviour, sensory, development and immunology. This work is now used by the INFRAFRONTIER project to track use of mouse resources in Europe, informing data storage and visualisation. IMPC efforts have been acknowledged by the G7, where the IMPC is one of five recognised global research infrastructures.

In 2017, we launched a new mouse informatics resource in collaboration with the Jackson Laboratory – PDX Finder – a comprehensive open global catalogue of Patient Derived Xenograft models and their associated datasets. PDX mouse models have emerged as an important oncology research platform to study tumor evolution, drug response and for tailoring chemotherapeutic approaches to individual patients. The PDX Finder project formed new collaborations with international consortia including the NCI-funded PDXnet and Patient Derived Model Repository as well as European commission-funded EuroPDX.

The publication of minimal information standard for PDX models (PDX-MI) that defines the minimal attributes needed to accurately describe a PDX model will enable users to access the data needed to deliver new therapeutics and improve the understanding of disease progression.

Recent funding for the Human Cell Atlas Data Coordination Platform extends our ontology work and uses our



tooling to deliver an ontology for cell types, processes and tissues. The new assay types for single-cell sequences revealed by the HCA have encouraged us to review how we capture sequence types and metadata at submission time. As data becomes available in the HCA, we will apply the Experimental Factory Ontology (EFO) process to extend ontological representation of cell types, which will be based on transcriptomic profiles. This representation will be used to drive queries and analysis in a number of portals including the Single Cell Expression Atlas being developed at EMBL-EBI.

We have also continued to develop our semantics as a service toolkit (OLS, Zooma, OxO and Webulous) and work with a number of projects to develop requirements and new features. These range from supporting targeted application ontology building for Open Targets and the HCA, to automated large-scale metadata annotation within BioSamples. The new OxO Ontology Mapping service supports interoperability of ontology standards within the CORBEL project and received additional funding from pharmaceutical companies via the Pistoia Alliance. We have developed a new mapping prediction algorithm that can map private vocabularies from industry to public ontologies in the areas of disease and phenotype. The user-base of these tools has doubled since 2016, and we have seen an increasing demand for the deployment of these services in external settings, including from our industry partners.

Future plans

The GWAS Catalog aims to support increasing data flow and study complexity, including use of whole exome sequencing. We will automate the process for deposition of summary statistics in a new programmatically accessible database, while also working with the community to define standard formats for these summary statistics. In 2018, we will deploy an improved literature search process for identifying publications, expand the Catalog's reach to include targeted array studies, increase the use of ontologies, and improve the search interface.

BioSamples will see improved linkage with EMBL-EBI resources and support for project-specific views enabling sustainable mini portal deployment, which will be lightweight and customisable. We will also introduce sample-based validation criteria for metadata supporting the FAIR principles and metrics funded by an ELIXIR implementation study. Lastly, we will implement the GA4GH metadata model for genomics data sharing via a new API.

IMPC will launch a new portal that focuses on translatability and succinctly presents the broad array of data we are capturing. We will update the analysis pipeline to incorporate phenotype data



Helen Parkinson

Head of Molecular Archival Resources. Team Leader – Samples, Phenotypes and Ontologies

*PhD Genetics, 1997.
Research Associate in Genetics,
University of Leicester 1997-2000.
At EMBL-EBI since 2000.*

from ageing studies and use machine learning in our analysis pipelines to help process image data to better understand disease mechanisms. In 2018 we will also launch and expand PDX Finder.

Our metadata and standards work continues and we will deliver an improved classification and terminology for experimental methods, including sequencing methods, that will provide a common metadata checklist for data submitted to archival resources at EMBL-EBI and projects such as the Human Cell Atlas. We will also improve the EFO disease representation for neurodegenerative disease, inflammatory bowel disease and rare disease. We will expand the EFO disease classification through integration with Monarch Disease Ontology (MONDO) to provide a single disease hierarchy across all major public disease ontologies. To improve accessibility, we will develop a new user interface that integrates our semantic tools and delivers cloud-ready installations of our toolkit.

Selected publications

MacArthur J, et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* doi: 10.1093/nar/gkw1133

Morales J, et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology.* doi:10.1186/s13059-018-1396-2

Streeter I, et al. (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* doi:10.1093/nar/gkw928

Meehan TF, et al. (2017). PDX-MI: Minimal Information for Patient-Derived Tumor Xenograft Models. *Cancer Res.* 2017 Nov 1;77(21):e62-e66. doi: 10.1158/0008-5472.CAN-17-0582

Karp NA, et al. (2017). Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat. Commun.* doi: 10.1038/ncomms15475

European Nucleotide Archive

The Data Coordination and Archiving team focuses on nucleic acid sequence data, while also extending into multi-omics. The team's work spans the creation and maintenance of data resources, the operation of Data Coordination Centres for a portfolio of collaborative projects, standards development and technology development for data sharing and coordination. They also build and maintain the European Nucleotide Archive (ENA).

Our flagship data resource, the ENA, continues to provide a record number of sequences and associated contextual data, spanning the spectrum from raw data, through alignments and assemblies to functional annotation. Now in its fourth decade of service, the ENA provides a critical feed of primary data into a host of EMBL-EBI data resources and beyond.

ENA also provides a platform for the management and coordination of data. This allows researchers to publish and archive data, as well as manage data flows within the live phase of a project.

Our portfolio of Data Coordination Centres included 14 collaborative projects during 2017, spanning marine science; biodiversity and environmental sciences; public health; food safety and infectious disease molecular epidemiology; pluripotent stem cell applications; livestock species functional genomics; and emerging sequencing platforms.

Major achievements

In 2017, we saw ENA content grow significantly in size (e.g. with the addition of 60 million new assembled/annotated sequences and 1.5×10^{15} base pairs of read data) and in breadth (e.g. with the addition of new data types, identification tables and pathogen analysis summaries). We have made extensions and improvements to many of our user-facing services, spanning submissions and data presentation. We also delivered numerous enhancements to our back-end data archiving systems.

New tools

We have launched the Pathogen Portal, a website and associated web services that provide a focal point for users involved in pathogen genomics and sequence-based surveillance. Supported by the COMPARE project, the portal provides access management tools that allow users to control the sharing of their pre-publication data to different Data Hubs. It also boasts a powerful search across comprehensive

public data and authorised pre-publication data, plus a browser for reference pathogen genome assemblies.

Through the EMBRIC marine biotechnology-based Research Infrastructure collaboration, we have launched the EMBRIC Configurator service. The service assists marine scientists planning projects that include omics methods to explore and select appropriate elements of ELIXIR infrastructure, such as data resources, tools, standards and expertise into project-specific configurations.

ENA

Driven by new data types and UX testing, we have enhanced and extended presentation services for ENA content. With the addition of new views and groupings of data in the ENA Browser, we have delivered simpler user workflows to access, for example, sets of contigs that contribute to a genome assembly and higher level assembly information.

A major new tool is the RESTful ENA Discovery API, which provides powerful text, dictionary, taxonomic, numeric, date and geographical search across all ENA content and options for the retrieval of metadata. The new ENA FTP Downloader allows users to access data files at scale across FTP, providing a powerful foundation for access to ENA's large datasets. Further tools include the ENA browser scripts; these downloadable Python scripts provide dataset retrieval services by accession (including Run, Project and Sample) and taxonomy, to support the simple retrieval of datasets required for onward study or analysis.

ENA introduced submissions support for a number of new data types and launched a major new user workflow into the Webin system. New data types include the “Antibiogram”, representing structured laboratory measures of minimal inhibitory concentrations of antimicrobial agents of value to those building and testing sequence-based antimicrobial resistance prediction; third party metagenomics assemblies, used so far by the EMBL-EBI Metagenomics group;

“identification” data, representing the outputs of environmental genomics taxonomic and functional diversity studies; and “pathogen analysis” summary tables, representing high-level summaries of key pathogen sequence analyses, such as taxonomic identification and typing.

Finally, our major new submissions workflow supports the submission of annotated sequences that do not form part of a genome, transcriptome or metagenome assembly. This brings greater consistency between records, a significant speed improvement for submitters and an opportunity for the team's staff to focus on submissions support and improvement of the validation rules and standards that are applied to data as they flow into ENA.

Data coordination

Our involvement in agricultural livestock functional genomics initiatives sees us provide the Data Coordination Centre for the FAANG Consortium, driving advances in metadata standards, representations in BioSamples and ENA, and improvements in our Ontology portfolio. Our engagement with pluripotent stem cell researchers has brought rich data integration across BioSamples, EGA and ENA through projects such as EBISC and HipSci. In our marine microbial collaborations, such as Tara Oceans, we continue to provide data coordination and services on the environmental metadata across ENA and BioSamples.

Work in biodiversity discovery includes the UniEuk initiative, in which we coordinate data aspects of a global effort to understand evolutionary relationships between eukaryotic species, with a particular focus on the protists. We continue to support the early adopter community around the Oxford Nanopore Technologies sequencing platform. Finally, our pathogen data coordination activities now extend to 15 Data Hubs, around which specific user groups collaborate on different aspects of detection and characterisation, covering, for example, viral metagenomics, antimicrobial resistance prediction through machine learning and global urban sewage system-based tracking of pathogenic species and resistance genes.

Future plans

We will grow our data brokering activities through the development of an international network of data brokers. Data brokers are trusted groups with expertise in a particular area of biology, that field and support data submissions on behalf of ENA.



Guy Cochrane

Team Leader – European Nucleotide Archive

*PhD University of East Anglia, 1999.
Team Leader since 2009.
At EMBL-EBI since 2002.*

We will build support for the addition of new data types of relevance to ENA. Using our “Analysis” records, we will formalise a process for the addition of such new data types. Our first example will likely be reference datasets for taxonomic classification in environmental omics applications.

The new ENA website will offer a simpler interface for users, and will offer more powerful search than is currently available, with a more logical layout of documentation.

In its next phase, the Pathogen Portal will integrate data visualisation and interpretation systems, likely through the iPython Notebook system, already trialled in a number of data hubs. This system can support both users with bioinformatics scripting skills and, through pre-configured Notebooks, users requiring direct interactive access.

Selected publications

Silvester N, et al. (2018). The European Nucleotide Archive in 2017. *Nucleic Acids Res.* doi:10.1093/nar/gkx1125

Cook CE, et al. (2018). The European Bioinformatics Institute in 2017: data coordination and integration. *Nucleic Acids Res.* doi:10.1093/nar/gkx1154

Karsch-Mizrachi I, et al. (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Res.* doi:10.1093/nar/gkx1097

Ten Hoopen P, et al. (2017). The metagenomic data life-cycle: standards and best practices. *Gigascience.* doi:10.1093/gigascience/gix047

Mitchell AL, et al. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* doi:10.1093/nar/gkx967

EGA, EVA and Archive Infrastructure

The core mission of the European Genome-phenome Archive (EGA), European Variation Archive (EVA), and European Nucleotide Archive (ENA) is to provide the basic infrastructure to enable global, public and secure sharing of genetic data. The team delivers the underlying archival and submission infrastructure for the ENA, which captures and presents information relating to experimental workflows that are based around nucleotide sequencing.

The main underlying driving technology for nucleotide data generation is genome sequencing, which has undergone a fundamental technological shift in the past ten years, reflected in the vast increases in nucleotide data submitted to the EGA and EVA.

Major achievements

In 2017, the Global Alliance for Genomics and Health appointed a set of key global resources to become Driver Projects for responsible genomic data sharing. These are real-world genomic data initiatives that help guide the GA4GH development efforts and pilot the standards and tools. The ENA/EVA/EGA was appointed as a driver project, an acknowledgment of our ongoing commitment to align our services with global standards to ensure interoperability.

European Genome-phenome Archive

Human data submissions to the EGA have continued to increase. In 2017 we witnessed a 12% increase in the number of studies submitted and a 29.5% increase in total archive size, bringing it to 5.85 petabytes (PB). We continued to improve the EGA data access service so that it now supports secure, real-time streaming protocols for slicing genetic data by genomic region via the GA4GH htsget protocol. In ELIXIR, we partnered with the RD-Connect platform in an Implementation Study to develop the EGA streaming API for real-time streaming of aligned sequence data into the RD-Connect genome browser.

In 2017, we collaborated with the UK Biobank initiative to provide the archiving and distribution services for the first release of genotyping data for all 500 000 UK Biobank participants. Between July and December, the EGA has received over 300 applications from more than 130 institutes to access this dataset. Over half a petabyte of data was transferred within the first three weeks alone and demand for the data continues to grow. Partnering with UK Biobank leverages EMBL-EBI's infrastructure investment and our work on international standards for personalised medicines through the GA4GH. On the basis of this initial collaboration with UK Biobank, the EGA is expected to start receiving whole-exome sequencing for UK Biobank samples in 2018.

The EGA collaboration with the Foundation for the National Institutes of Health (FNIH) Accelerating Medicines Partnership has resulted in the creation of a European federated node for the AMP T2D Knowledge Portal, which enables browsing, searching, and analysis of human genetic information linked to type 2 diabetes and related traits, while protecting the integrity and confidentiality of the underlying data.

European Variation Archive

The data volumes from genetic variation studies continue to expand dramatically as third-generation sequencing technologies enable the creation of reference genomes for non-model species. In 2017 the number of species represented in the EVA increased to 22 and 577 million fully-browsable variants, representing a 17% growth in variation data. Our API continues to support third-party community portals such as the International Wheat Information System, Ensembl Genomes, and the International Sheep Genomics Consortium.

The assignment and maintenance of long-term stable identifiers to genetic variants (rs IDs) is essential for the scientific community because it allows researchers to share and communicate known variants. In 2017, the EVA and dbSNP (NCBI) published an agreement that the EVA will maintain these long-term locus accessions for all non-human species, while dbSNP will maintain the human identifier space. Because of this agreement, the EVA has already seen an increase in submissions by 59% in 2017, which is expected to continue in coming years.

Future plans

The EGA will continue its collaboration with the UK Biobank as we enter the next phase of the project: the generation of exome sequencing (provided by Regeneron Pharmaceuticals). We anticipate the arrival of the first batches of sequencing data in mid-2018 with completion by the end of 2019. The EGA will provide the distribution for both sequencing and genetic variation data, leveraging the EGA's scalable data distribution infrastructure. In our collaboration with the FNIH AMP project, we will continue to enrich the catalogue of type 2 diabetes relevant studies available in the federated node at EMBL-EBI, to power the online analysis available via the AMP-T2D knowledge portal.

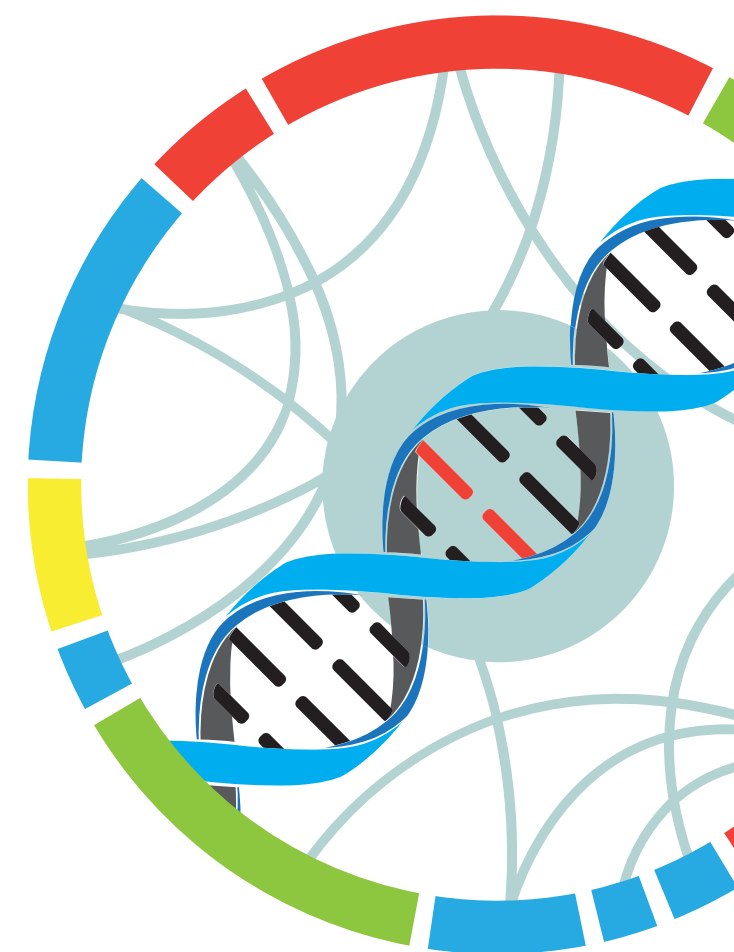
In 2018, the EVA will complete the transfer of all non-human variation data from dbSNP. We are aiming to make the first official rs ID release from the EVA by the end of the year.

The next step of our collaboration with GA4GH will be the deployment of standards for the ENA/EGA/EVA driver projects, which will enable scaled dataflow. We will also release the htsget standard providing configurable data access for genomic data and deploy the DUO data use ontology for EGA enabling users to code and query consent.



Thomas Keane

Team Leader - EGA, EVA and Archive Infrastructure
Wellcome Trust Sanger Institute,
2008-2016. PhD Biology, 2006.
At EMBL-EBI since November 2016.



Genes, Genomes and Variation

Genes, genomes and variation data resources include core EMBL-EBI services such as Ensembl, Ensembl Genomes and the GWAS Catalog. Teams working in this area play a vital role in data coordination for large-scale projects such as GENCODE and the Global Alliance for Genomics and Health (GA4GH).

Major achievements

In 2017, the Ensembl data resource expanded the number of supported species, developed new tools for genome interpretation and improved its interactive data distribution platform. Ensembl also released four comprehensive updates, in addition to a special update of resources supporting GRCh37, the previous version of the human genome assembly.

Ensembl welcomed the arrival of multiple vertebrate genome assemblies using newly updated approaches for genome annotation and comparative genomics. For example, in 2017 we used our new streamlined genome annotation pipeline for 18 primate species and 15 rodent species, while also increasing our ability to deliver comparative genomics resources for a large collection of species.

In 2017, the leadership of the GENCODE project transferred from the Wellcome Sanger Institute to EMBL-EBI. GENCODE is the major activity of Ensembl HAVANA — the Human and Vertebrate Annotation and Analysis — and is becoming progressively integrated with the Ensembl infrastructure and workflows. GENCODE aims to provide globally recognised reference annotation for the human and mouse genomes by combining computational and manually curated gene annotation, as well as experimental data.

Our Non-vertebrate Genomics team updated the genome assemblies of several important species including barley (where we were part of the consortium that generated the new data) and the yellow fever mosquito. Specifically, Ensembl Plants now contains the full dataset from the 1001 Arabidopsis project, and over 20 million variant loci in bread wheat. We have also continued to collect and integrate RNA-seq data; there are now data from over 2500 studies in plant, vector and microbial species.

To support new use cases and more flexible ways of interacting with the Ensembl resources, we designed an automated tool that installs ready-to-deploy REST services on a standard virtual machine. The growth in users was also reflected by our growing focus on training; in 2017, Ensembl taught 2000 people across 17 countries.

A major new development for WormBase was the launch of the Genome Decoders Project, a collaboration with the Institute for Science in Schools, which aims to introduce school students to scientific work. We have trained over 1000 students in community curation tools, so they can contribute to a re-annotation of the whipworm genome.

The GWAS Catalog also saw significant growth in 2017, with a total of 454 studies and 6766 SNP-trait associations from 273 publications. In addition, the

team released structured sample ancestry and location of recruitment information. In 2017, 68 GWAS Catalog studies had summary statistics and the number is expected to grow significantly in the future.

Future plans

Over the coming year, the resources of the GGV cluster will broaden the EMBL-EBI portfolio by incorporating and annotating a growing list of sequenced genomes. Specifically, in 2018 we anticipate the annotation of more than 100 vertebrate genomes, the vast majority of which will not have been supported in Ensembl before. A highlight will be the annotation and release via Ensembl of the 25 genomes being sequenced as part of the Wellcome Sanger Institute's 25th anniversary. We are also looking forward to extensive involvement in other emerging projects aimed at sequencing species across the tree of life.

GENCODE's goal of a complete first pass manual annotation of the mouse genome is likely to be realised in 2018, as are significant improvements to our process for creating and updating the human gene set. Finally, in 2018, we plan to significantly increase the number of plant genomes in Ensembl Plants. We will upgrade the wheat genome assembly to the new International Wheat Genome Sequencing Consortium (IWGSC) reference sequence, and build a new wheat sub-site offering access to genomes of additional cultivars as they become available.

Data Resources

Ensembl

Ensembl is a browser for chordate genomic data that supports research in comparative, population and functional genomics. We provide data for over 100 genomes from vertebrates and model organisms.

www.ensembl.org

Ensembl Genomes

Ensembl Genomes covers all non-vertebrate cellular life, including bacteria and archaea, fungi, protists, plants and non-vertebrate metazoa. Genome sequence, gene models and sequence analysis are available for each species.

www.ensemblgenomes.org



Paul Flicek

Head of Genes, Genomes and Variation.
Team leader – Vertebrate Genomics. Senior Scientist.

*DSc Washington University, 2004.
Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008.
Team Leader at EMBL-EBI since 2007,
Senior Scientist since 2011.
At EMBL-EBI since 2005.*

GWAS Catalog

The NHGRI-EBI GWAS Catalog is a quality-controlled, manually-curated, literature-derived collection of all published genome-wide association studies. It provides starting points for studies to identify causal variants, understand disease mechanisms, and establish targets for novel therapies.

www.ebi.ac.uk/gwas

PhytoPath

PhytoPath integrates genome-scale data from important plant pathogen species with literature-curated information about the phenotypes of host infection. Using the Ensembl Genomes browser, it provides access to complete genome assembly and gene models of priority crop and model-fungal, oomycete and bacterial phytopathogens.

<http://www.phytopathdb.org/>

VectorBase

VectorBase is a resource by the National Institute of Allergy and Infectious Diseases (NIAID) Bioinformatics Resource Center (BRC) providing genomic, phenotypic and population-centric data to the scientific community for invertebrate vectors of human pathogens. The VectorBase resource is a collaboration between University of Notre Dame, USA, Imperial College London, UK, and EMBL-EBI.

www.vectorbase.org/

WormBase

WormBase is a comprehensively curated database of the genetics, genomics and biology of *C. elegans* that also contains curated genomes for key parasitic nematodes. WormBase is a collaboration between EMBL-EBI, the Wellcome Sanger Institute, the Ontario Institute for Cancer Research and Caltech.

<https://parasite.wormbase.org/index.html>

An artist's interpretation of the diversity of life.

Genes, Genomes and Variation

Team achievements



Paul Flicek

Vertebrate Genomics team

- Four major releases of Ensembl and Ensembl Genomes, including updates to other highly used resources such as human (now assembly GRCh38.p11) and mouse (now GRC38m.p5) genomes
- Updated Ensembl resources for the previous human genome assembly GRCh37
- Four updates to the mouse GENCODE gene set and two updates to the human GENCODE gene set
- Collaboration with the Human Genome Structural Variation Consortium to incorporate valuable new fully open genomic data into IGSR
- Leadership roles in the Vertebrate Genomes Project and Genome 10K Project to help ensure that the newly sequenced genomes from these efforts have high-quality annotation and are widely available to the scientific community
- Participation in the planning, launch and realisation of GA4GH Connect



Daniel Zerbino

Genome Analysis team

- Active support of international consortia (GTEx, Open Targets, MuG, IGSR, IHEC, GA4GH, BLUEPRINT and GREEKC)
- Creation of the Ensembl Metadata Registry to track Ensembl databases both inside and outside EMBL-EBI
- Deployable REST service
- Archival of REST data servers
- Streamlined microarray mapping process
- Enriched annotation of miRNAs



Bronwen Aken

Vertebrate Annotation team

- Annotation of 18 primates and 15 rodent genomes using an improved genome assembly annotation pipeline
- New annotation for the cat genome, which also includes 28 tissue-specific gene tracks made using RNA-seq data
- Collaboration with Eagle Genomics/Horizon Discovery produced and openly released an improved annotation of the Chinese hamster ovary cell line assembly CHOK1GS_HDv1
- Tuatara genome project
- Collaboration with the Norwegian and Swedish ELIXIR nodes to help them set up and run our genome annotation pipelines



Fiona Cunningham

Variation Annotation team

- Correct handling of RefSeq transcript variant annotation in the Variant Effect Predictor (VEP) tool
- REST endpoint that returns all currently known identifiers for a given variant name
- Normalisation for insertion and deletion variants in Ensembl
- New VEP plugins support more detailed descriptions of variants located near splice sites and loss of function intolerant scores for genes from ExAC
- Ontology mappings for traits in mouse and for a number of livestock species to the Mammalian and Clinical Measurement ontologies respectively, and associated REST endpoint
- Over 3200 GWAS Catalog studies curated
- Summary statistics generated for a first 68 GWAS Catalog studies



Paul Kersey

Non-vertebrate Genomics team

- Issued three public releases of Ensembl Genomes
- Added the full datasets from the 1000 Anopheles and 1001 Arabidopsis population genomics projects
- Increased the number of unicellular eukaryotic genomes to almost 1000
- Genome assembly updates for numerous species including barley and the yellow fever mosquito
- Collection of community-curated data from numerous species and launch of the Genome Decoders project, in which over 1000 school students have been trained to help annotate the whipworm genome
- Major participation in the launch of the new AGR (Alliance for Genome Resources) website, a major new resource for model organism species
- Produced and made available RNA alignment data from 1642 plant, 376 vector and 657 microbial eukaryote experiments



Andy Yates

Genomics Technology Infrastructure team

- EMBL-EBI now hosts Ensembl from our London data centre
- Yearly archives of Ensembl Genomes sites are now available
- Development of a portal to integrate managed access EBiSC data from EGA with Ensembl genome annotation
- Integration of Reactome's pathway viewer into Ensembl Genomes
- Two new tools to convert Ensembl flat file data into other formats and to calculate LD across 1000 Genomes populations
- Developed a new library for accessing UCSC binary formats from Perl
- Trained over 2000 researchers across 17 countries
- Over 10 000 Ensembl Twitter followers
- Hosted three Google Summer of Code students

Vertebrate Genomics

The Vertebrate Genomics team works with the Vertebrate Annotation, Genome Analysis, Variation Annotation and Genomics Technology Infrastructure teams to create and deliver Ensembl resources. Other Vertebrate Genomics projects include the International Genome Sample Resource, the world's largest open repository of human genomic data, which incorporates the 1000 Genomes Project data; and the Genome Reference Consortium, which is responsible for distributing and updating the genome assemblies for human, mouse, zebrafish and chicken.

Vertebrate Genomics is responsible for the scientific leadership of the Ensembl project and is the largest component of the GENCODE consortium, which is based at EMBL-EBI and seeks to create foundational reference annotation for the human and mouse genomes. We are also actively involved in the Global Alliance for Genomics and Health.

Major achievements

Ensembl

We released four comprehensive updates to Ensembl in addition to a special update of our resources supporting GRCh37, the previous version of the human genome assembly. There are three major areas that drive our plans, goals and priorities: distributing data and developing tools that facilitate genome interpretation, creating high quality informatics resources for all species, and enabling researcher-driven analysis. Each of these areas has seen significant progress this year.

Ensembl is a flagship genome annotation resource receiving more than 5 million user sessions annually from nearly 800 000 unique locations around the world. Tens of thousands of researchers published results in 2017 based on, or otherwise using, Ensembl annotations and tools. Ensembl is also extensively involved in major international genomics consortia. For example, this year saw the culmination of our participation in the Genotype Tissue Expression (GTEx) project, as well as extensive involvement with the Vertebrate Genomes Project (VGP) and Genome 10K.

International Genome Sample Resource (IGSR)

In 2017, we published a Data Note describing the alignment of the complete set of 1000 Genomes sequencing reads to the GRCh38 human reference assembly. We have been working extensively to call the variants from these alignment files using methods that are comparable to those used in the 1000 Genomes Project phase 3. We anticipate that these will be completed in the second half of 2018 and will be released on the IGSR website.

We also added a significant amount of new data from the Human Genome Structural Variation Consortium to the IGSR website. The data include a number of newly-available or newly-mature technologies that are especially relevant for discovery of structural variation.

Genome Reference Consortium

This year, the GRC released two patch updates of the GRCh38 human genome assembly, one patch update for the GRCm38 mouse assembly and the comprehensively updated GRCz11 zebrafish assembly. We also completed and published an extensive analysis of the GRCh38 human reference genome assembly.

GENCODE

In 2017, the leadership of the GENCODE project transferred from the Wellcome Sanger Institute to EMBL-EBI, with the start of the third funding cycle for the project. GENCODE is the major activity of Ensembl HAVANA – the Human and Vertebrate Annotation and Analysis – a project, which moved to EMBL-EBI in June 2017 and is becoming progressively integrated with the Ensembl infrastructure and workflows.

Updated GENCODE gene sets are currently released twice annually for human and four times annually for mouse. A unique feature of GENCODE is the incorporation of manual genome annotation using structured procedures to ensure accuracy and consistency across the genome.

Global Alliance for Genomics and Health

The Global Alliance for Genomics and Health (GA4GH) launched “GA4GH Connect” at its fifth plenary meeting in October, 2017. This new phase followed several months of reorganisation and detailed planning to shift the project focus firmly on framing policy and developing standards to enable real-world genomic data sharing by 2022.

A number of leadership roles in GA4GH are held across Ensembl, Vertebrate Genomics and EMBL-EBI, including membership in the GA4GH steering committee and co-chair of the Data Security Foundational Workstream. In this latter role, we published guidance for mitigation of re-identification risk associated with the GA4GH Beacon project.

Future plans

In 2018 Ensembl will make significant progress in each of our three major areas of focus. For example, we are working closely with the RefSeq project at NCBI to converge on a single human transcript set that can be used for genome interpretation in tools such as the Ensembl Variant Effect Predictor (VEP).

We plan to annotate more than 100 vertebrate genomes, the vast majority of which will not have been supported in Ensembl before. This effort includes the annotation and incorporation into Ensembl of the 25 genomes being sequenced as part of the Wellcome Sanger Institute's



Paul Flicek

Head of Genes, Genomes and Variation. Team leader – Vertebrate Genomics.

*DSc Washington University, 2004.
Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008.
Team Leader at EMBL-EBI since 2007,
Senior Scientist since 2011.
At EMBL-EBI since 2005.*

25th anniversary. We will also continue extensive behind-the-scenes developments to support the needs of Ensembl's users.

We are working with several groups to incorporate new data into IGSR and will see the first GRC release of a reference chicken genome assembly.

GENCODE's goal of a complete first pass manual annotation of the mouse genome is likely to be realised in 2018, as are significant improvements to our process for creating and updating the human gene set. Finally, we are looking forward to extensive involvement in emerging projects aimed at sequencing species across the tree of life.

Selected publications

Aken BL, et al. (2017). Ensembl 2017. *Nucleic Acids Res.* doi: 10.1093/nar/gkw1104

Clarke L, et al. (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* doi: 10.1093/nar/gkw829

Raisaro JL, et al. (2017). Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *J Am Med Inform Assoc* 24. doi: 10.1093/jamia/ocw167

Schneider VA, et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* doi: 10.1101/gr.213611.116

Zheng-Bradley X, et al. (2017). Alignment of 1000 Genomes Project Reads to Reference Assembly GRCh38. *Gigascience.* doi: 10.1093/gigascience/gix038

In 2017 Ensembl released a special update supporting the GRCh37 human genome assembly.

Genome Analysis

The Genome Analysis team develops software infrastructure for the storage and high-throughput analysis of genome-wide datasets within the Ensembl ecosystem.

Because of its stability, both within the cell and across evolution, genomic sequence offers a broad canvas on which to project an array of sequence markers, tracing a complex cascade of molecular events from sequence evolution to phenotypic change. As our experimental toolkit expands, more shades are added to this painting. The Ensembl databases contain records of sequence evolution, sequence conservation, genetic diversity, transcription, DNA marks, chromatin marks, phenotypic association etc. The Genome Analysis group ensures that all these aspects of molecular biology can be stored and studied jointly.

The diversity of genomic datasets is particularly marked in the field of epigenomics. Large international consortia, such as ENCODE, Blueprint, GTEx, MuG or the Horizon2020-funded MultipleMS project are currently surveying an ever-expanding space of DNA modifications, histone marks, chromatin states, chromatin conformation, protein-DNA binding, transcriptional activity, translational activity, and cis-regulatory interactions. We engage directly with these consortia and help coordinate them through our involvement in the International Human Epigenome Consortium (IHEC). These datasets are then integrated into the Ensembl Regulatory Build, our evidence-based annotation of regulatory elements of the genome.

To make the data available to researchers around the world, we maintain a number of web services, in particular the Ensembl REST API. This programmatic interface allows developers to directly access the

Ensembl databases using the programming language of their choice, without having to install any specialised software. Around it, a number of specialised services track external datasets, such as the TrackHub Registry and the Epigenome Reference Registry (EpiRR).

Major achievements

Collaborations

The Genome Analysis group collaborates with a large number of consortia, in particular GTEx, Open Targets, MuG, IGSR, IHEC, GA4GH, Blueprint and GREEKC, for which we provided bioinformatics support and advice.

To provide highly valuable informatics resources even as the amount of genome sequence data accelerates, we developed a consistent system to track assemblies across Ensembl and partner instances. The Ensembl Metadata Registry will be a single point of entry to locate the databases describing any genome assembly. In the future, it will allow us to integrate the annotation efforts of external specialists, multiplying the reach of the Ensembl infrastructure.

Ensembl REST

As the popularity of the Ensembl REST service grows, so do its use cases. In particular, researchers in clinical settings expressed the need to run their own instance locally, rather than sending potentially sensitive data

to our servers. To make it easier for external users to deploy new services, we now provide an automated tool that will install ready-to-deploy REST services on a standard virtual machine. We are successfully using this functionality in our production cycle. Similarly, researchers expressed the need to access a stable copy of Ensembl annotations, without the risk of annotations changing subtly and breaking the reproducibility of their analyses. For this and other purposes, we started providing archives of the REST services.

Mapping and annotation

One service which remains popular despite the continuous emergence of new technologies is microarray probe mapping onto the genome, genes and transcripts. We have considerably streamlined our mapping process so that it now runs in a much more automated and reproducible manner and the infrastructure is now shared with Ensembl Genomes.

Finally, Ensembl has provided GO terms for its annotation for many years. However, these have typically been imported from UniProt, meaning they are only available for genes with protein-coding transcripts. In collaboration of RNACentral and UCL, we are now able to provide annotations for non-coding transcripts.

Future plans

The research community is poised to assemble the genomes of most vertebrate species over the next several years. Ensembl is directly involved in the Vertebrate Genomes Project, which brings together these initiatives. The Genome Analysis team will support this exponential acceleration by removing all bottlenecks in Ensembl's storage system and maintaining the eHive workflow manager.

As a key member of the International Human Epigenome Consortium, Ensembl is working to ensure that all datasets registered by the consortium will flow directly into Ensembl's Regulatory Build.

Beyond existing techniques, an array of functional assays is shedding new light on the rich biomolecular activity of DNA. We will map SELEX binding motifs onto the genome and integrate DNA modification results into our annotations. We are creating a comprehensive database of genome editing experiments and plan to create a database of molecular QTL datasets. Further, we are keeping an active watch on novel assays that may warrant dedicated databases, such as chromatin conformation experiments like Hi-C, functional validation experiments, including STARR-Seq and MPRA.



Daniel Zerbino
Team Leader – Ensembl
Genome Analysis
*MSc in Biotechnology, Ecole Nationale Supérieure des Mines de Paris, 2005.
PhD in Bioinformatics, EMBL and University of Cambridge, 2009.
At EMBL-EBI since 2013.*

As genotyping becomes routine in clinical settings to diagnose rare diseases or cancer, Ensembl resources are increasingly used by clinical geneticists. Supporting their use cases requires some adjustment to our services. We are currently developing the Transcript Archive (Tark) in collaboration with the Variant Annotation group to allow the definition and sharing of disease-relevant transcript sets. In collaboration with Open Targets, we are in the process of testing an integrated post-GWAS analysis pipeline, POSTGAP, which infers causal genes from GWAS summary statistics, intersecting a wide array of public genomic databases .

Finally, Ensembl is undergoing significant technical consolidation in an effort to better integrate a number of EMBL-EBI resources. Genome Analysis will bring together the infrastructure underpinning Ensembl, Ensembl Genomes and Ensembl HAVANA. Further, UniProt proteins will be connected to Ensembl genes via a new intermediary database, GIFTS, which we have started developing.

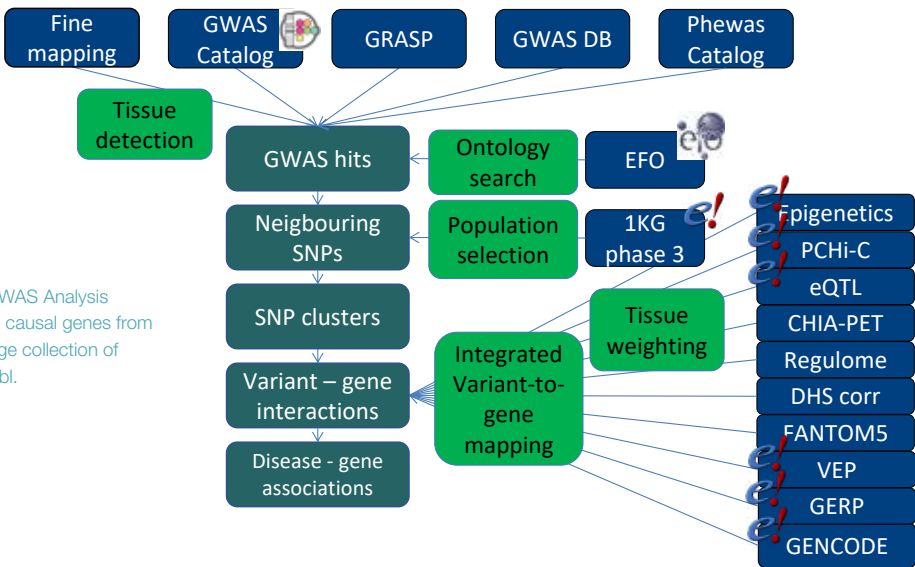
Selected publications

Zerbino DR, et al. (2017). Ensembl 2018. *Nucleic acids res.* doi: 10.1093/nar/gkx1098

Ruffier M, et al. (2017). Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database.* doi: 10.1093/database/bax020

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature.* doi: 10.1038/nature24277

Schematic diagram of the POSTGAP (Post-GWAS Analysis Platform), currently under testing, which infers causal genes from GWAS results, through the integration of a large collection of bioinformatics databases, in particular Ensembl.



Vertebrate Annotation

The Vertebrate Annotation team creates comprehensive gene annotation and comparative genomics resources for genome assemblies supported by Ensembl. Its goal is to produce high-quality reference datasets that enable research in biology, evolution and the mechanisms of disease.

Our resources link diverse species at the DNA and gene level, and are the primary annotations used in the analysis of many international genome projects and by communities interested in research, clinical and agricultural applications. We have annotated reference gene sets for over 100 vertebrate species, including key model organisms and livestock. In addition, we enable data discovery and integration by annotating gene families, gene orthologues and conserved genomic regions across these species.

Major achievements

New genome assemblies

In 2017 we supported the arrival of multiple vertebrate genome assemblies, produced by initiatives such as the Genome 10K project. The Genome 10K project aims to sequence the genome of at least one individual from each of the approximately 10 000 vertebrate genera. To ensure that Ensembl can meet the needs of these communities, we have been collaborating closely with the Vertebrate Genomes Project (VGP), a component of the Genome 10K project that is co-led by the Wellcome Sanger Institute.

We also released new or updated genome assemblies and annotations in 2017 for several important and highly accessed species including pig, cat and the Chinese hamster ovary (CHO) cell line.

Annotation and comparative genomics

The arrival of multiple new genome assemblies for different species requires new thinking about how Ensembl's computational approach and pipelines work for both genome annotation and comparative genomics. The first outcomes of these developments started to appear in earnest in 2017: the release of several genomes annotated using the new genome annotation pipeline, including 18 primate species and 15 rodent species.

In parallel we have also made significant headway in preparing our comparative genomics pipelines so that we are able to complete genomic alignments for the many new species that we anticipate annotating during each Ensembl release. For example, in the past year, we carried out three times as many multiple species alignments than in 2016.

One aspect of improving our comparative pipelines, namely migrating our Hidden Markov Model (HMM) search to HMMER 3, is already completed. HMMER 3 is faster and uses both CPU and disk more efficiently. Another aspect of this work, the integration of a new multiple alignment algorithm, is underway.

Enabling others

We have also taken steps to support groups outside Ensembl who want to use the Ensembl annotation pipeline, including collaborations with the Norwegian and Swedish ELIXIR nodes, to help them set up and run our pipelines. In parallel, we have also developed tools to enable us to present comparative genomics information from genomes annotated by groups other than Ensembl. This has been enabled by a collaboration with the Tuatara Genome Project. The tuatara is a reptile endogenous to New Zealand that can be considered a 'living fossil'.

Future plans

Over the next year, we will continue to develop our annotation and comparative genomics infrastructure, with the aim of increasing the quality and the automation of the annotation. We will also use these tools to rapidly increase the number of vertebrate genomes annotated and presented in Ensembl. In particular, we plan to annotate approximately 50 fish genomes as well as a large number of bird genomes.

Beyond these efforts, we hope to complete our new pseudogene and small ncRNA pipelines in order to improve our annotation in both of these areas. We will continue to upgrade our annotation mapping code to support larger evolutionary distances and to include the mapping of non-coding genes/transcripts.

Other planned improvements to the annotation pipeline include integration of *ab initio* methods for higher quality annotation for distant species where no suitable reference yet exists.

From an infrastructure perspective, we have started the development of a pipeline (with an associated database backend) to find new publicly available genome assemblies and assign stable identifier prefixes and ID spaces. We plan to tie this into our previous work on establishing a genome assembly quality control pipeline to identify, classify and prepare assemblies for annotation.



Bronwen Aken

Team Leader – Ensembl
Vertebrate Annotation

*BSc in Molecular and Cell Biology,
University of Cape Town, South
Africa. MSc in Bioinformatics and
Computational Biology, Rhodes
University, South Africa.
Ensembl team member since 2005.
At EMBL-EBI since 2014.*

We will continue to update our RNA-seq based annotation methods, with a particular focus on automating the retrieval, classification and use of samples and data from the European Nucleotide Archive (ENA). This ties in with our efforts to improve metadata standards across various projects and we hope in the future to be able to separate samples into the appropriate tissue/development stages using the metadata information alone.

We further plan to incorporate decision-making logic directly into our pipeline to determine, for example, when we have enough reads per tissue to stop searching for and aligning further samples. We also hope to improve our support for human annotation, with a focus on providing new evidence tracks for use by both the Ensembl HAVANA manual annotators and our users.

Selected publications

Pujar S, et al. (2018). Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1031

Jasinska AJ, et al. (2017). Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. *Nat Genet.* doi: 10.1038/ng.3959

Aken BL, et al. (2017). Ensembl 2017. *Nucleic Acids Res.* doi:10.1093/nar/gkw1104

In 2017 the Ensembl team released several genomes annotated using the new genome annotation pipeline, including 18 primate species.



Variation Annotation

The Variation Annotation team focuses on building large-scale systems for understanding genomic variation. This is fundamental for progress in biology, from basic research to translational genomics applications. The team creates novel workflows and databases to integrate data for Ensembl, resulting in one of the largest catalogues of integrated and annotated variant, phenotype, trait and disease data.

To aid the interpretation of genetic variants in their evolutionary and disease context, we develop Ensembl’s Variant Effect Predictor (VEP), a tool for *in silico* annotation of variants using the data in Ensembl.

The team also focuses on data curation for the GWAS Catalog, which summarises key findings from all eligible published genome wide association studies (GWAS). Furthermore, to improve annotation of genes associated with disease, we review the annotation to provide a more robust framework for clinical reporting of variants (Locus Reference Genomic), and provide data on disease-causing variants in a structured manner for high throughput access (Gene2Phenotype). This can be used to filter results from genome and exome sequencing studies.

Major achievements

Variation data in Ensembl

In 2017 the number of variants and structural variants across Ensembl databases increased to 741 million, from 562 million in 2016, and we expect a greater increase in 2018. This was driven in part by the doubling of the number of short variants available for humans. A number of large-scale sequencing projects, such as the Genome Aggregation Database, UK10K and NHLBI Trans-Omics for Precision Medicine projects, have also made allele frequency data available and accessible in a standardised manner via the Ensembl infrastructure. We have extended our API to perform allele equivalence checking for insertion and deletion variants to normalise more data and promote data discovery.

Variant Effect Predictor (VEP)

This year we significantly updated the VEP code to improve its robustness and functionality. We also improved our ability to annotate variants using RefSeq human transcripts. Additionally, the VEP now reports the impact of missense variants on the protein function of RefSeq transcripts using SIFT and PolyPhen2.

Phenotypes and ontologies

We import phenotype and disease associations from many different sources into Ensembl. Often, we encounter the same disease or trait in different databases under different labels (e.g. Type 2 diabetes and ‘diabetes, type II’). Further to our ontology mapping work on human data last year, we now hold mappings for traits in mouse and for a number of livestock species to the Mammalian and Clinical Measurement ontologies respectively. This improves the ability to query results aggregated across many sources. It also improves the utility of our association tables, which can be viewed grouped by ontology term. New REST endpoints have been created to allow programmatic access to these mapped results.

GWAS Catalog

In 2017 a total of 454 studies and 6766 SNP-trait associations from 273 publications were manually curated for the GWAS Catalog. In addition, we released structured sample ancestry and location of recruitment information.

The Catalog includes over 5000 curated studies from over 3200 publications and more than 58 000

SNP-trait associations. It provides starting points for studies to identify causal variants, understand disease mechanisms, and establish targets for novel therapies. We now have summary statistics for 68 GWAS Catalog studies. Their availability, combined with an increase in public engagement via email and Twitter, has resulted in regular requests to host full p-value summary statistics prior to publication and we expect this number to increase.

Locus Reference Genomic (LRG) sequences

An LRG record contains stable reference sequences that are used for reporting sequence variants with clinical implications. These are created manually with community input and consultation. We are collaborating with NCBI to have one agreed stable transcript between Ensembl (EMBL-EBI) and RefSeq (NCBI). In 2017 we have achieved UTR to UTR consensus with RefSeq for 33 of 130 (25%) transcripts and ensured CDS consensus and discussion on UTRs with RefSeq for 30 of 130 (23%) transcripts. We also reviewed our automated transcript choosing pipeline results for 117 transcripts and found that we agreed with the automated selection three quarters of the time.

CRISPR

In collaboration with the Genome Analysis team, we have been involved in starting a small project to create a curated archive of CRISPR experiments, called the ‘Genome Editing Catalogue’. The Genome Editing Catalogue will contain links to CRISPR screen papers, targeted mutagenesis experiments and summary statistics.

Future plans

Within Ensembl, our goal is to continue delivering variation data in a standardised manner for an increasing number of species and increasing volumes of data. With this in mind, we will further extend our Perl API to reduce the dependency on databases and instead retrieve more data directly from the EVA. This will allow us to support a greater range of species at minimal cost.

The linkage disequilibrium statistics we provide via our REST server have proved popular, so we will release a tool that will provide support for calculations over larger



Fiona Cunningham

Team Leader – Variation Annotation

BA Natural Sciences, University of Cambridge, 2000. MSc Bioinformatics, 2001, and PhD in Bioinformatics, University of Cambridge, 2014. At EMBL-EBI since 2008.

regions. To improve our ability to use protein-based annotations for interpretation of variation data, we are also collaborating with colleagues in the PDBe team to create interactive views of variants located on protein structures. We plan for existing prototypes to be in production in the first half of 2018.

Within the GWAS team, we will develop and implement an improved literature search process for identification of publications to ensure all eligible publications are considered. Meanwhile we are focusing on expanding the number of studies with summary statistics captured in the GWAS Catalog and are collaborating with Open Targets to deliver a summary statistics database.

For our collaboration with RefSeq at NCBI, we aim to have one primary transcript for each gene, perfectly matching at the coding sequence level, for 97% of genes, by October 2018.

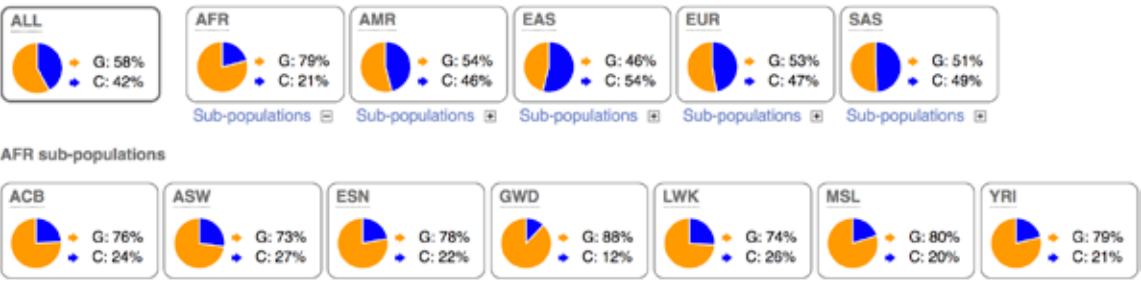
Selected publications

Aken BL, et al. (2017). Ensembl 2017. *Nucleic Acids Res.* doi: 10.1093/nar/gkw1104

MacArthur J, et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucl. Acids Res.* doi: 10.1093/nar/gkw1133

McLaren, et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biology.* doi: 10.1186/s13059-016-0974-4

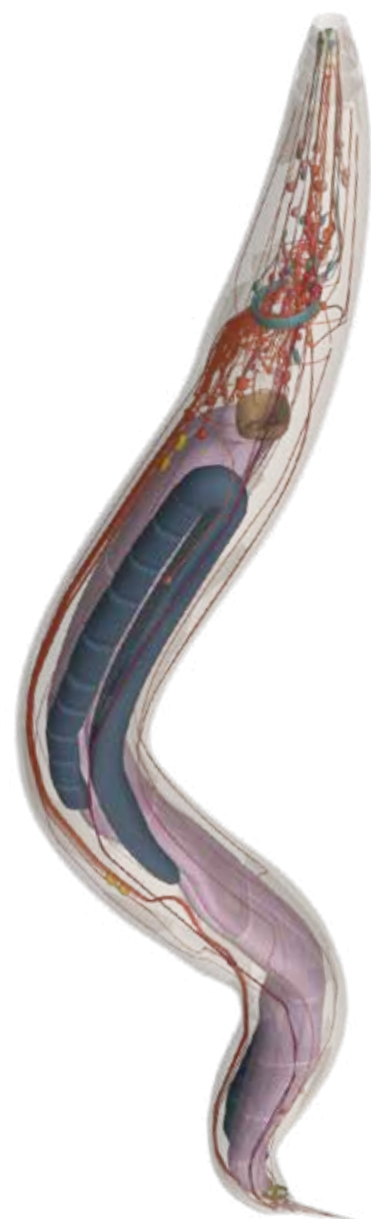
1000 Genomes Project Phase 3 allele frequencies



Opposite: The frequency spectrum of variant rs1333049 in the 1000 Genomes project panels. The allele frequencies in the continental populations and African sub-populations are displayed here.

Non-vertebrate Genomics

High-throughput sequencing is transforming both understanding and application of the biology of many organisms. The Non-vertebrate Genomics team integrates, analyses and disseminates such data for scientists working in domains as diverse as agriculture, pathogen-mediated disease and the study of model organisms.



WormBase provides information on the genetics, genomics and biology of *C. elegans* and related nematodes. WormBase is part of the Alliance of Genome Resources, which launched its web portal in 2017. Image credit: openworm.org

We run services for bacterial, protist, fungal, plant and invertebrate metazoan genomes, mostly using the power of the Ensembl software suite and usually in partnership with interested communities. In such collaborations, we contribute to the development of many resources, including VectorBase for invertebrate vectors of human disease, WormBase for nematode biology, and PhytoPath for plant pathogens. In the plant domain, we collaborate closely with Gramene in the US and with a range of European groups in the ELIXIR-EXCELERATE project.

By collaborating with EMBL-EBI and re-using our established toolset, small communities with little informatics infrastructure can perform and interpret highly complex and data-generative experiments. We also work on large, complex genomes like hexaploid bread wheat, establishing informatics frameworks for the analysis of species for which genomic datasets are only now gaining traction as technologies improve.

Our major activities include genome annotation, broad-range comparative genomics and the visualisation and interpretation of genomic variation, which is studied increasingly in species throughout the taxonomy.

Major achievements

Ensembl Genomes

In 2017, we made substantial improvements to our main public services: Ensembl Genomes, VectorBase and WormBase. These included three public releases of Ensembl Genomes, increasing the number of unicellular eukaryotic genomes in the resource to almost 1000. The team also made upgrades to the genome assemblies of several important species including barley (where we were part of the consortium that generated the new data) and the yellow fever mosquito. We have also added many significant new datasets to the resource, especially variation data.

Ensembl Plants

Ensembl Plants now contains the full dataset from the 1001 Arabidopsis project, and over 20 million variant loci in bread wheat. We have also continued to collect and integrate RNA-seq data; there are now data from over 2500 studies in plant, vector and microbial species.

We work with scientific communities to collect additional information on gene models. We integrated data from the grey mould causing *Botrytis cinerea* (which was completely re-annotated by the community), and incremental updates from 33 vector species. We are currently working with the wheat pathogen *Zymoseptoria tritici* community.

WormBase

A major new development was our collaboration with the Institute for Science in Schools – which aims to introduce school students to scientific work – to launch the Genome Decoders Project. We have trained over 1000 students in community curation tools, so they can contribute to a re-annotation of the whipworm genome.

WormBase is part of a larger consortium, the Alliance for Genome Resources (AGR), which aims to provide integrated access to model organism data. We are contributing significantly to the conceptual and technical development of the AGR; the AGR's web portal went live in 2017.



EMBL-EBI collaborated with the Institute for Science in Schools to launch the Genome Decoders Project, which trained over 1000 students in the use of curation tools. The students will use the knowledge to contribute to the reannotation of the whipworm genome.

New projects

We have worked on the development of standards for data representation and exchange in the context of two European projects: ELIXIR (in which we are concentrating on plant data) and in a new project, Infravec2, which is focused on vectors of infectious diseases. Another new project, which started in 2017, is 'Designing Future Wheat'. The project brings together eight UK institutes and universities working on this species, increasingly using molecular techniques to prepare the way for future improvement of the crop.

Future plans

In 2018, we plan to significantly increase the number of plant genomes in Ensembl Plants. We will upgrade the wheat genome assembly to the new IWGSC reference sequence and build a new wheat sub-site offering access to genomes of additional cultivars as they become available.



Paul Kersey

Team Leader – Non-Vertebrate Genomics

PhD University of Edinburgh, 1992.
Postdoctoral work at University of Edinburgh and MRC Human Genetics Unit, Edinburgh.
At EMBL-EBI since 1999.

We will also be participating in the second phase of the Arabidopsis 1001 genomes project, which will produce 50-100 high quality assemblies and explore the use of graph-based structures for data representation and visualisation of population-wide data. In addition, we will complete the first phase of the IRIS project, and will be looking at how we can repeat the exercise in future years.

As always, we are looking at ways of increasing our capacity to handle large quantities of data. One focus is on automation of the processing of variant data through tighter integration with the European Variation Archive. We are expecting to see increasing numbers of datasets linking genotype to phenotype (for example, from vector species and fungal pathogens), and will be seeking to deliver the data in a useful manner to users. Finally, we are expecting to roll out a new comparative genomics framework, better able to link protein families that span the taxonomic space.

Selected publications

Kersey PJ, et al. (2018). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1011

Raymond Y N Lee, et al. (2018). WormBase 2017: molting into a new stage. *Nucleic Acids Res.* doi: 10.1093/nar/gkx998

Tello-Ruiz MK, et al. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.* doi: 10.1093/nar/gkx998

Clavijo BJ, et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* doi: 10.1101/gr.217117.116

Mascher M, et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature.* doi: 10.1038/nature22043

Genomics Technology Infrastructure

The Genomics Technology Infrastructure team provides web, production and outreach support for the Ensembl and Ensembl Genomes projects, including the development of data-mining tools. Fast, clear and reliable access to genomic annotation is fundamental to the continuing growth of this field. Ensembl and Ensembl Genomes websites receive over 5 million visits per year. The team provides interfaces to visualise annotation within the context of the genome and gene evolution.

Complementary to this is our online tools platform. We provide access to sequence search via BLAST and Blat, variation consequences via the Ensembl Variant Effect Predictor (VEP), and data format conversion. Consistent interfaces, tools and visualisations across all Ensembl hosted species help to lower the barrier to performing genomic data analysis across the taxonomic spread.

In addition, we offer worldwide on-site and remote training in genomics, provide Ensembl interfaces and tools, and operate the Ensembl helpdesk.

Major achievements

Infrastructure improvements

This year saw the culmination of a two-year project to migrate Ensembl’s infrastructure from the Wellcome Sanger Institute to EMBL-EBI. The switch over in March 2017 to having all UK public services (Ensembl, FTP, REST, public MySQL, archives) served by our data centres occurred with minimal external impact.

The Ensembl and Ensembl Genomes websites continue to be responsive to new use cases. For example, we have integrated Reactome’s pathway widget into the Ensembl Plants website with pathways available for a selection of genes from *Arabidopsis thaliana*.

Tools

Our tools offering continues to grow with the release of our Linkage Disequilibrium (LD) Calculator capable of calculating LD across genomic regions, a set of variants or across a defined window for a variant based on the 1000 Genomes population data. We have also expanded our phenotype support providing an ontological aware search via the Ensembl search interface.

Our work on a data mining platform that will eventually replace our BioMart service continued within the context of the EBiSC (European Bank for induced pluripotent Stem Cells) project. We are in the final stages of developing an “Allelic Query Service”, offering a way to select lines of importance based on genotype and consequence. This service uses Ensembl data and integrates with EGA managed access VCF files, retrieved using the GA4GH developed htsget protocol.

Training and outreach

Ensembl has taught over 2000 people in 2017 across 17 countries. Our social media presence continues to grow with Ensembl’s Twitter account reaching over 10 000 followers regularly engaging with our “Gene of the Week”, a platform where other EMBL-EBI services promote their service offerings through the context of a single gene.

Finally, Genomics Technology Infrastructure hosted three Google Summer of Code students to develop a catalogue of Ensembl genomic annotation, a searchable database and interface over our FTP sites, and a JavaScript library to visualise BigWig files within the Genovise genome browser.

Future plans

In 2018, a focus will be the ongoing redevelopment of the Ensembl web platform. Initiatives such as Genome10K and the Vertebrate Genomes Project provide an unprecedented opportunity to expand our resources for vertebrates and require new ways to annotate, interpret and visualise these genomes.

Our website redesign is firmly rooted in user-driven tasks. We are identifying known and future workflows, which are termed user journeys, that a genome browser must solve to be successful, and have hired a designer to help lead this effort. These user journeys will be plotted into a journey map allowing the identification of key components and capabilities that should be our first focus in the redesign. In addition, we are developing a number of paper and code prototypes to validate our design ideas. Our BioMart replacement service will be trialed through this process, with the intention of serving as a benchmark of our ongoing design approach. We plan for the complete Ensembl redesign process to take in excess of two years. Our aim is to make regular releases available to key external collaborators for feedback and design iteration.

While we are redeveloping our platform, the need to update our existing infrastructure to serve the needs of the community continues. We are developing three new interfaces during 2018. The first, an integrated 3D viewer of variants on protein structures, is the result of a collaboration between Ensembl and PDBe. The second, an improved interface to our regulation data sets, is in



Andy Yates
Team Leader - Genomics Technology and Infrastructure
*MSc in Bioinformatics, University of Manchester, UK.
At EMBL-EBI since 2006.*

response to the increasing number of epigenomes now available from Epigenomes Reference Registry (EpiRR). We are also developing a new transcript variation viewer (TVV) widget to help visualise variation transcript consequences from tools such as VEP. TVV is an initial design approach to highlight the pertinent data through a process of decluttering and simplification of our existing interfaces without losing context or meaning. Like many other Ensembl widgets, it is intended to be run either in standalone mode or as an integrated visualisation component, allowing the integration of TVV into future interfaces. In addition, our EBiSC “Allelic Query Service” will be released during 2018.

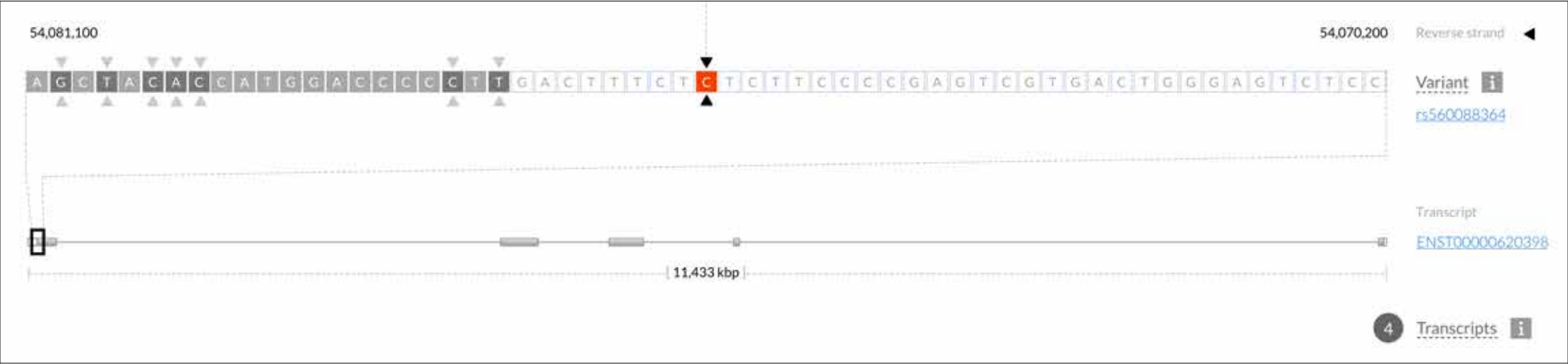
Finally, we are embarking on a wider adoption of container technology within the team. Containers, and their orchestration frameworks, provide an attractive way to manage microservices (web services with intended limited functionality). They also provide a useful mechanism to distribute our services to external partners. We will continue to develop our container portfolio and offer more containers to run Ensembl services within a wider range of environments.

Selected publications

Zerbino DR, et al. (2017). Ensembl 2018. *Nucleic Acids Res.* 46:D754-D761. doi: 10.1093/nar/gkx1098

Kersey PJ, et al. (2017). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* 46:D802-D808. doi: 10.1093/nar/gkx1011

A design for our new TVV interface to be included in the upcoming website redesign. TVV reduces the complexity of comparing multiple genomic variations against multiple transcripts to a single variant against a single transcript. The design shows a variant in a genomic, transcript, protein and protein domain context whilst flagging neighbouring variations through the use of triangles on the transcript sequence..



Molecular Atlas

Our RNA and protein expression service teams are working to create a comprehensive, integrated, scalable atlas of gene and protein expression. The cluster manages ArrayExpress, the Expression Atlas and PRIDE data resources, and works very closely with the Literature cluster to develop the BioStudies database, which will supersede ArrayExpress and will serve as a platform to archive bioimage data.

Major Achievements

During 2017 our cluster focused on several key projects. We developed the components to enable the representation of the single cell RNAseq data in the Atlas and prototype the single-cell RNA Atlas. The team also developed metadata templates and guidelines for single-cell RNAseq data representation, which are being used to select suitable experiments for re-analysis.

We started working on the development of open and reproducible proteomics data analysis pipelines, and their deployment in cloud environments. In that context, we started the continuous integration of proteomics datasets coming from PRIDE in the Expression Atlas.

The team laid the groundwork for dealing with multi-omics data and wrapping the ArrayExpress archive into the BioStudies database. We made significant improvements to the Expression Atlas user interface and expanded its content with new high-profile datasets. The Expression Atlas now contains 3166 datasets, from which 134 are reporting baseline expression across 32 different organisms and 3035 datasets concern differential expression.

As a result of our participation in the Pan-Cancer Analysis of Whole Genomes (PCAWG) visualisation working group, the Expression Atlas is one of the PCAWG data portals.

The PRIDE database, a key ProteomeXchange partner, has continued to grow, strengthening its role as the world-leading proteomics repository. It received a record number of submitted datasets (approximately 200 per month). The volume of data downloads from PRIDE was the largest to date in a single year, reaching 295 TB. In 2017, PRIDE focused on quality control and reprocessing of submitted datasets.

Enabling translational and biomedical research, we delivered data-management solutions for the EU-AIMS project on autism spectrum disorder and built a cloud environment for multi-omics data for the European Medical Information Framework (EMIF) IMI project, with the aim of reusing patient health records in clinical research.

The number of individual study records in the BioStudies database exceeded one million. To enable exploration of this information, we developed better means of data filtering, as well as an API. We started using BioStudies for data capture and sharing in EU-ToxRisk – a large European collaborative project. We worked on improving the scalability of the system to prepare it for multi-terabyte datasets, in particular, for imaging data.

Jointly with our collaborators at Dundee University, we have started building the BioImage Archive to store biological image reference datasets.

Future plans

One of the major focuses of the cluster will be building a portal for single-cell gene expression data. This will be done in collaboration with the Wellcome Sanger Institute. The data from the Human Cell Atlas will be a significant contributor to EMBL-EBI's Single Cell Expression Atlas, but we will also be curating and loading high-quality datasets available through ArrayExpress, Gene Expression Omnibus (GEO) and other sources, across species and conditions. In addition, we will be continuing our data support for the Open Targets initiative and establishing a regular data provision for single-cell RNAseq in baseline, as well as differential, disease contexts.

In 2018 we will continue to work on the development of open analysis pipelines for mass spectrometry proteomics data. Particularly, we will start developing Data Independent Acquisition workflows (e.g. SWATH-MS). Our overall strategy involves using these and other additional data pipelines to improve the dissemination of proteomics data from PRIDE into other bioinformatics resources (Ensembl, UniProt and the Expression Atlas).

We will also continue working on better representation of image data, as well as easy to use image data deposition. We will release a new version of the BioStudies submission tool, providing flexible metadata templates, which will enable control over the structure of datasets submitted to individual projects within BioStudies.

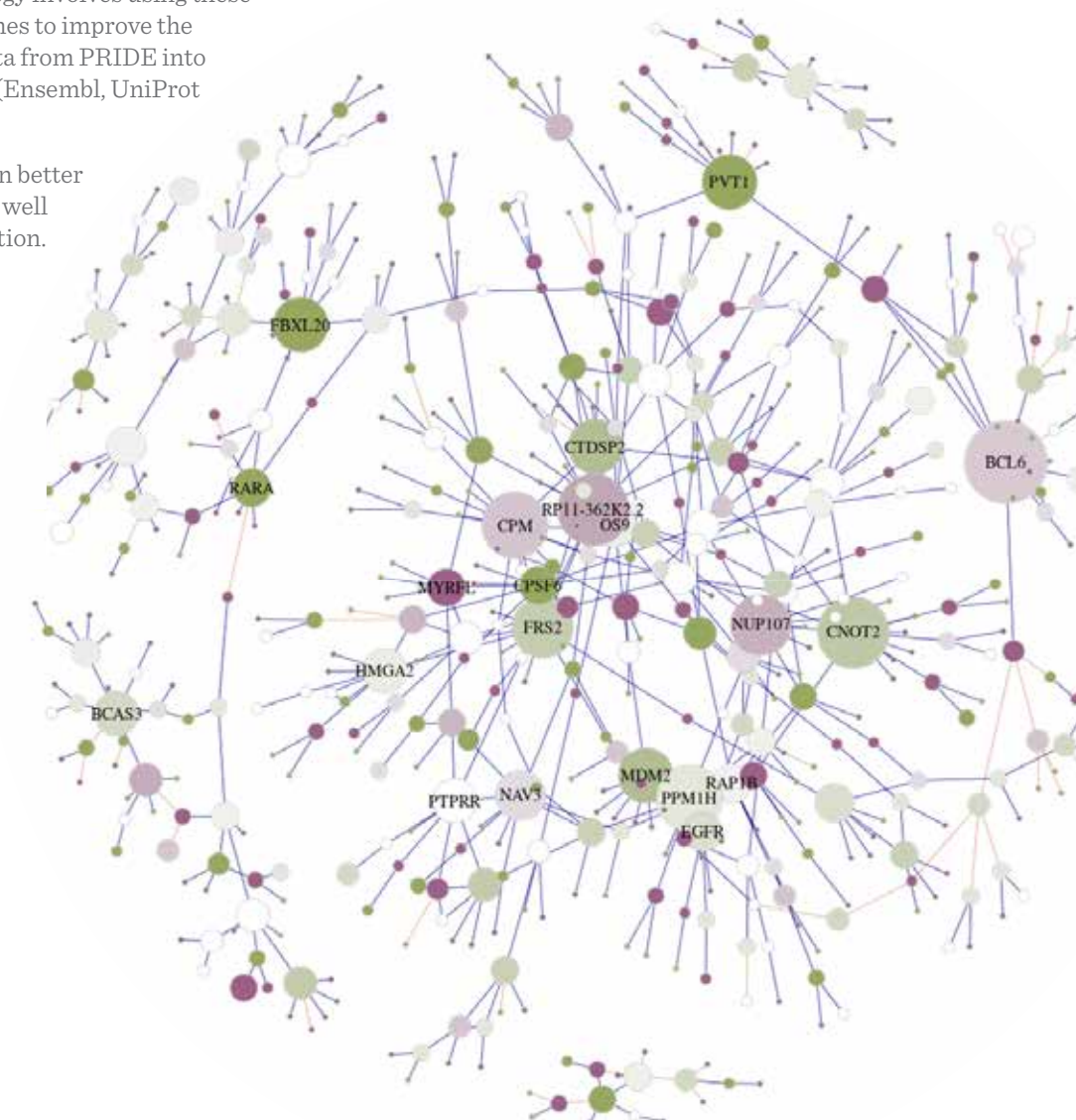
Alongside our collaborators at Dundee University and other groups at EMBL, we will continue building the BioImage Archive.



Alvis Brazma

Head of Molecular Atlas.
Senior Team Leader, Senior Scientist.

PhD in Computer Science, Moscow State University, 1987.
MSc in Mathematics, University of Latvia, Riga.
At EMBL-EBI since 1997.



Structural rearrangements associated with RNA fusions

Molecular Atlas

Data resources

ArrayExpress

The ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides data for reuse to the research community.

www.ebi.ac.uk/arrayexpress/

Expression Atlas

The Expression Atlas is an open science resource that allows users to find information about gene and protein expression across species and biological conditions. Expression Atlas aims to help answer questions such as “where is a certain gene expressed?” or “how does its expression change in a disease?”.

www.ebi.ac.uk/gxa

PRIDE

The PRoteomics IDentifications (PRIDE) database is a centralised, standards-compliant, public data repository for proteomics data. It includes protein and peptide identifications, post-translational modifications and supporting spectral evidence.

www.ebi.ac.uk/pride

Team achievements



Robert Petryszak

Gene Expression team

- ⦿ Improved the functionality of Expression Atlas and its Baseline Atlas, which comprises large genomics and proteomics studies
- ⦿ Produced guidelines for single-cell RNA-seq data curation
- ⦿ Developed a pipeline for single-cell RNA-seq data
- ⦿ Led the submissions handling effort for ArrayExpress
- ⦿ Brokered sequencing data for the ENA



Juan Antonio Vizcaino

Proteomics team

- ⦿ Record number of submitted datasets to PRIDE database (approximately 200/month)
- ⦿ Largest volume of data downloads from PRIDE in a single year (295 TB)
- ⦿ PRIDE named an ELIXIR Core Data Resource
- ⦿ Expansion of the ProteomeXchange Consortium to include a 5th member (iProX, China)
- ⦿ First phase of the development of open and reproducible proteomics data analysis pipelines to be deployed in cloud environments
- ⦿ Integration of the first proteomics datasets coming from PRIDE in the Expression Atlas



Ugis Sarkans

Functional Genomics Development team

- ⦿ Improved scalability of the BioStudies infrastructure, as the number of datasets exceeded one million
- ⦿ Implementation of a better information-finding mechanism and an API to efficiently explore study data packages from papers imported from Europe PMC

Gene Expression

The Gene Expression team handles the acquisition, curation, quality control, statistical analysis and visualisation of functional genomics data. It focuses on microarray, high-throughput sequencing-based gene expression and related proteomics data. The team runs several core EMBL-EBI resources, including the Expression Atlas and ArrayExpress.

The team also ensures RNA-sequencing quality control and analysis, the results of which are used by numerous resources at EMBL-EBI and externally.

We are part of Open Targets, the Cancer Genome Atlas Pan-Cancer analysis (PCAWG) projects and recently we joined the Human Cell Atlas project. Analysis and visualisation on plant data is also a major component of our work through our involvement in the Gramene project.

We collaborate closely with the Brazma, Marioni and Stegle research groups at EMBL-EBI, as well as the Teichmann and Hemberg groups at the Wellcome

Sanger Institute, developing new methods and algorithms, integrating new types of data across multiple platforms, and investigating relationships between transcriptomics and proteomics data in the context of cancer genomics.

Team members contribute substantially to online and face-to-face training in transcriptomics, in particular relating to our team's resources, but also for related topics such as next-generation sequencing analysis.

Expression Atlas contains gene expression data for many large landmark studies.



GTEx ENCODE



FANTOM5
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Basic research



PCAWG
PanCancer Analysis of WHOLE GENOMES
EurocanPlatform



CCLE
Genentech

Cancer research



wellcome trust
sanger
institute

Zebrafish development



HDBR
HUMAN DEVELOPMENTAL BIOLOGY RESOURCE

Prenatal human brain



BLUEPRINT
epigenome



HipSci

Key cell line models



THE HUMAN
PROTEIN ATLAS

Proteomics



DMDD
Deciphering the Mechanisms of Developmental Disorders



UC DAVIS
KOMP Repository
KNOCKOUT MOUSE PROJECT

Mouse models

Major achievements

In 2017 the growth of high-quality transcriptomic data continued, reaching approximately 3200 experiments and nearly 120 000 assays. These assays included nearly 600 RNA-seq experiments, nearly 8500 differential comparisons across 32 organisms, and almost 750 plant experiments. At the end of 2017 the Baseline Expression Atlas contained 134 RNA-seq studies, including data from many high-impact studies.

We improved the Expression Atlas interface, applying enhancements in its home page, making datasets more easily accessible and highlighting large-scale projects. We also improved the presentation of experiment pages (e.g. new filters on metadata), and prototyped new functionalities (e.g. transcript quantification). We have developed a meta-analysis pipeline for baseline RNA-seq and are currently applying it to human datasets. These results are being submitted to the Open Targets platform regularly.

We have joined the Human Cell Atlas collaboration. At the same time, we developed production-quality analysis pipelines for Smart-Seq2 and droplet-based single cell RNA-seq, curation guidelines for single-cell submissions to ArrayExpress, and are preparing for the release of a new cross-species, added-value resource, the Single Cell Expression Atlas.

Our RNA-seq pipeline has now processed all eligible (over 310 000) public RNA-seq runs in 300 species, and its results have been made available via REST API, Perl and BioPython libraries that allow for ontology-powered queries. The results of the analysis are used in production processes by Expression Atlas, Ensembl, WormBase ParaSite and by a variety of external users.

Future plans

In 2018 our development efforts will focus on the release of the first version of the Single Cell Expression Atlas, as well as building further its data production pipeline to include differential expression, pseudotime inference and cell type alignment. We will develop further our web functionality to accommodate visualisation and queries based on these new types of analyses (e.g. cell type queries, marker gene search and comparison).

We will also focus on identifying high-quality studies for inclusion and facilitating their flow into the Single Cell Expression Atlas by improving our single-cell curation pipeline to handle larger experiments in a semi-automatic way.



Robert Petryszak

Team Leader –
Gene Expression

*MPhil in Computer Speech and
Language Processing, University of
Cambridge, UK.
At EMBL-EBI since 2003.
Team Leader since 2015.*

We will continue to load suitable experiments into the existing Expression Atlas and to add plant crops species through our collaboration within the Gramene consortium. Alongside the Thornton Group, we will start work on meta-analysis of differential gene expression data in multiple RNA-seq experiments studying human disease. In the future, we aim to apply the methods resulting from this research to our meta-analysis pipeline, to produce gene expression signatures for disease phenotypes.

Selected publications

Papatheodorou I, et al. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1158

Tello-Ruiz MK, et al. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1111

Petryszak R, et al. (2017). The RNASeq-er API – a gateway to systematically updated analysis of public RNA-seq data. *Bioinformatics.* doi: 10.1093/bioinformatics/btx143.

Koscielny G, et al. (2017). Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* doi: 10.1093/nar/gkw1055

Proteomics

The Proteomics team is responsible for the maintenance and further development of the PRIDE database of mass spectrometry-based proteomics data and its sister resource, the peptide-centric PRIDE Cluster. The team is also leading the global ProteomeXchange (PX) Consortium of proteomics resources, aiming to standardise public proteomics data submission and dissemination worldwide.

In addition, the team develops open source, stand-alone tools such as the PRIDE Inspector (for data visualisation and basic analysis), the PX submission tool (for performing dataset submissions), and different software libraries for handling proteomics data in open standard formats. In this context, we heavily contribute to the activities of the Proteomics Standards Initiative (PSI). Finally, we develop open proteomics data analysis pipelines for different proteomics approaches with the aim of improving scientific reproducibility.

The team's primary goals are to help researchers to access and make the most of public proteomics data, and to encourage the adoption of data standards in the field. We also aim to integrate proteomics data from PRIDE into other bioinformatics resources such as Ensembl, Expression Atlas and UniProt.

Major achievements

In 2017 the PRIDE database received a record number of 2443 datasets (approximately 200 datasets/month), further strengthening its role as the world-leading proteomics data repository. Reuse and reanalysis of public proteomics data continues to grow, for diverse applications.

To demonstrate this, in 2017, approximately 295 TB of data were downloaded from PRIDE, the largest volume in a single year so far. In July 2017, PRIDE became an ELIXIR Core Data Resource and, consequently, it has a prominent role in the context of the new ELIXIR Proteomics Community.

PRIDE is a founding member of the global ProteomeXchange Consortium of proteomics resources. During 2017, the Consortium incorporated its fifth (and second Asian) member: the iProX repository, based at the Beijing Phoenix Centre, China. At present, the ProteomeXchange members are: PRIDE, PeptideAtlas/PASSEL and MassIVE (USA), jPOST (Japan) and iProX.

In 2017 we started a new activity: the development of open and reproducible analysis proteomics pipelines as a starting point for the popular data dependent acquisition (DDA) approaches. We are working on deploying them in a cloud infrastructure (using the EMBL-EBI Embassy Cloud, as a proof of concept), so that in the future they can be reused by anyone in the scientific community.

Last, but not least, in 2017 we integrated data coming from several quantitative proteomics datasets to the Expression Atlas. We also finalised an initial version of the infrastructure needed to seamlessly integrate proteomics data in Ensembl and in the UCSC genome browser, using TrackHubs.

Future plans

In 2018 we will continue to work on the development of open analysis pipelines for mass spectrometry proteomics data and their deployment in a cloud infrastructure. In addition to DDA workflows, we will start working on Data Independent Acquisition (DIA) approaches (e.g. SWATH-MS), in collaboration with the Stoller Centre at the University of Manchester.

We also plan to refine and stabilise the infrastructure required to represent and integrate proteomics data into genome browsers such as Ensembl. Furthermore, we will work on the development of more stable and robust pipelines to disseminate PRIDE data to UniProt.

Selected publications

Deutsch EW, et al. (2017). The ProteomeXchange Consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* doi: 10.1093/nar/gkw936

Martens L, Vizcaíno JA. (2017). A golden age for working with public proteomics data. *Trends Biochem Sci.* doi: 10.1016/j.tibs.2017.01.001

Perez-Riverol Y, Vizcaíno JA. (2017). Synthetic Human Proteomes for accelerating protein research. *Nat Methods.* doi: 10.1038/nmeth.4191



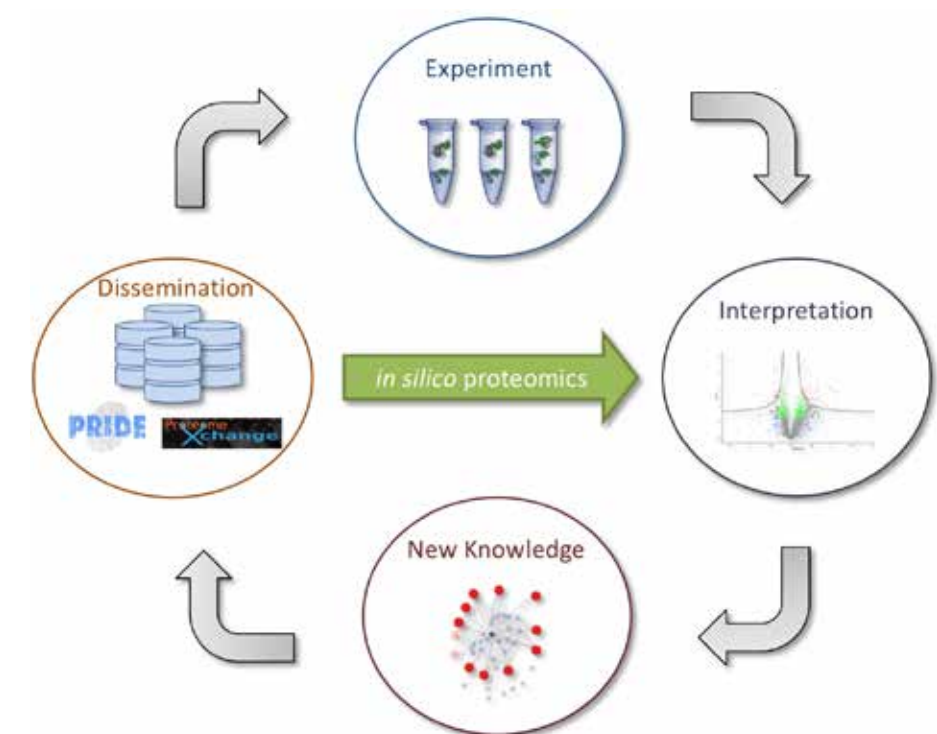
Juan Antonio Vizcaino

Team Leader – Proteomics

PhD in Molecular Biology, University of Salamanca, Spain, 2005. Postdoc position at University of Seville, Spain, 2005-2006. At EMBL-EBI since 2006.

Perez-Riverol Y, et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol.* doi: 10.1038/nbt.3790.

Vizcaíno JA, et al. (2017). The mzIdentML data standard v1.2, supporting advances in proteome informatics. *Mol Cell Proteomics.* doi: 10.1074/mcp.M117.068429



The rapidly growing volume of publicly available proteomics datasets in the PRIDE database (as part of ProteomeXchange) opens up the opportunity for *in silico* proteomics, which is using bioinformatics to test hypotheses directly through the available data, instead of going via the generation of new experimental data (PMID: 26449181).

Functional Genomics Development

Our team develops and maintains software for ArrayExpress, an ELIXIR Core Data Resource, and the BioStudies database, a resource for biological datasets that do not have a dedicated home within the EMBL-EBI services.

Together with the Expression Atlas team, we build and maintain data management tools, user interfaces, programmatic interfaces, and annotation and data submission systems for functional genomics resources. We also collaborate on a number of European ‘multi-omics’ and medical informatics projects in a data-management capacity.

Major achievements

The BioStudies database holds descriptions of biological studies, and links to data from these studies in other databases at EMBL-EBI and beyond.

In 2017 the number of individual study records in the BioStudies database exceeded one million. Most of these datasets were imported from Europe PMC – we receive study data packages for all papers with supplementary materials and/or links to life sciences databases.

To enable efficient exploration of this information, we developed a flexible data faceting mechanism that can be quickly adapted to various BioStudies projects. BioStudies now has an easy to use RESTful API.

We started using BioStudies for data capture and sharing in EU-ToxRisk – a large European collaborative project generating widely heterogeneous data. The data content hosted in BioStudies for another multi-partner project, HeCaToS, increased threefold in 2017. We worked on improving the scalability of the system, to prepare it for multi-terabyte datasets, in particular for imaging data.

We contributed to the THOR project that aimed to establish seamless integration between articles, data, and researchers across the research lifecycle. We implemented a data claiming mechanism that enables users to associate datasets that belong to them to their ORCID records.

In 2017 ArrayExpress was named as one of ELIXIR’s Core Data Resources. We continued running the ArrayExpress infrastructure, brokering sequence data to the European Nucleotide Archive. We also started preparations for moving ArrayExpress operations related to data distribution to the BioStudies platform.

Future plans

The BioStudies data submission tool will be developed further, allowing the definition of submission templates that will steer the data submission process according to the needs of a specific project like HeCaToS or EU-ToxRisk.

The process of migrating the existing ArrayExpress data and future functional genomics data submissions to BioStudies has started. We will refine the presentation of ArrayExpress data in BioStudies, document the migration path for the current ArrayExpress API users, and ensure that all the essential data management pipelines keep working.

We will continue working with the existing toxicogenomics projects and will add new data sources to BioStudies. For example, we are collaborating with the SourceData initiative that aims to publish figure source data alongside the figures, helping authors to archive research data and readers to analyse published results.

BioStudies hosts a number of large imaging studies, such as data from the Tara Oceans expeditions. We will provide better ways for users to explore and download these datasets.

Facets for exploring data imported from Europe PMC



Ugis Sarkans
Team Leader – Functional Genomics Development.
PhD in Computer Science, University of Latvia, 1998. Postdoctoral research at the University of Wales, Aberystwyth, 2000. At EMBL-EBI since 2000.

Selected publications

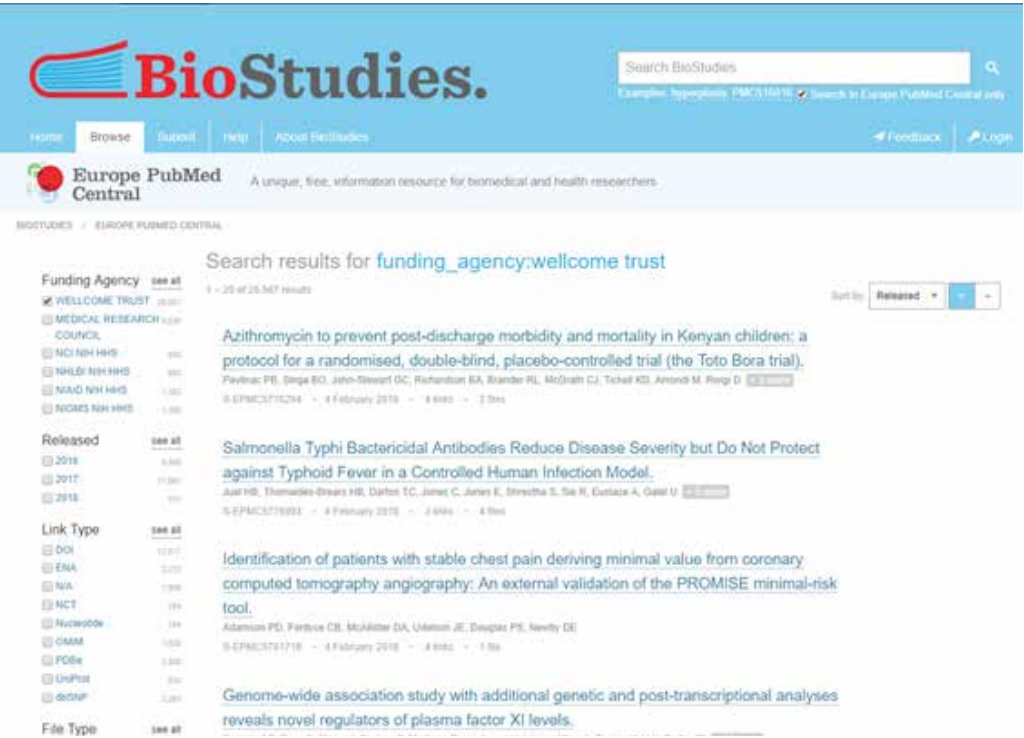
Sarkans U, et al. (2018). The BioStudies database - one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* doi:10.1093/nar/gkx965

Kuepfer L, et al. (2018). A model-based assay design to reproduce in vivo patterns of acute drug-induced toxicity. *Arch Toxicol.* doi:10.1007/s00204-017-2041-7

Williams E, et al. (2017). The Image Data Resource: A Bioimage Data Integration and Publication Platform. *Nat Methods.* doi:10.1038/nmeth.4326

Perez-Riverol Y, et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol.* doi:10.1038/nbt.3790

Brandizi M, et al. (2017). Orchestrating differential data access for translational research: a pilot implementation. *BMC Med Inform Decis Mak.* doi:10.1186/s12911-017-0424-6.



Proteins and Protein Families

EMBL-EBI provides foundational resources for researchers who work with protein sequences and protein families, including the UniProt, InterPro and Pfam data resources, and the HMMER homology search tool, among others. The team also provides unique resources for studying non-coding RNA with Rfam and RNACentral.

During 2017 there were some important changes in the organisation of the cluster. With Claire O'Donovan leaving to lead the EMBL-EBI Metabolomics team, we were delighted to recruit Sandra Orchard to join us to lead the Protein Function Content team. Sandra brings a wealth of experience and she actually began her career at EMBL-EBI as a UniProt curator. The Protein Families team, led by Rob Finn, has recast itself as the Sequence Families team to reflect its continuing focus on both protein and non-coding RNA.

Major achievements

UniProt continues to scale with the massive growth in sequence data. We provide selected Reference Proteomes that offer representative examples across the broad tree of life. In 2017 UniProt doubled the number of Reference Proteomes selected to over 10 000.

It is said that a picture paints a thousand words. The Protein Function Development team has continued to advance UniProt's graphical visualisations, which now also include protein interactions and subcellular location. These help our users rapidly gain an understanding of the components of often complex biological systems.

The Protein Function Development team have created new, accurate and comprehensive computational systems for functional annotation of UniProt proteins based on domain architectures (DAAC – Domain Architecture Alignment and Classification) and association rules (ARBA – Association Rule Based Annotator).

The Protein Function Development and Content teams have worked on developments aimed to facilitate biomedical and clinical research. UniProt has established an exchange pipeline with ClinVar for sharing variants of clinical interest. Work is ongoing to design a platform for interpretation of variants based on the role and functional mechanisms of proteins in disease.

The Sequence Families team has also seen a growth in the area of metagenomics, with EBI Metagenomics receiving more data than ever. The group has experimented with producing assemblies from the increasingly deep sequencing sets provided with submitted samples. These assemblies generate more useful contiguous sequences and help to reduce the number of peptide sequence fragments for analysis.

Future plans

A major challenge for the coming years will be to organise the flow of information from metagenomic assemblies into other areas of EMBL-EBI. There have been several papers reporting the identification of thousands of novel genomes using these approaches.

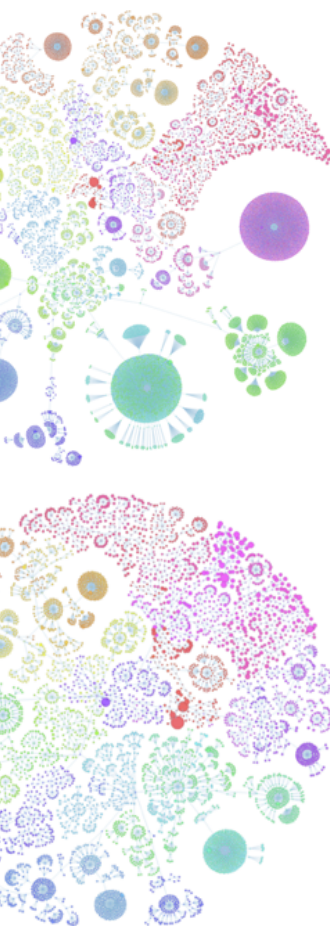
A second important challenge will be to enable our growing community of clinical users to access UniProt information to interpret the consequences of mutations that lead to human disease.



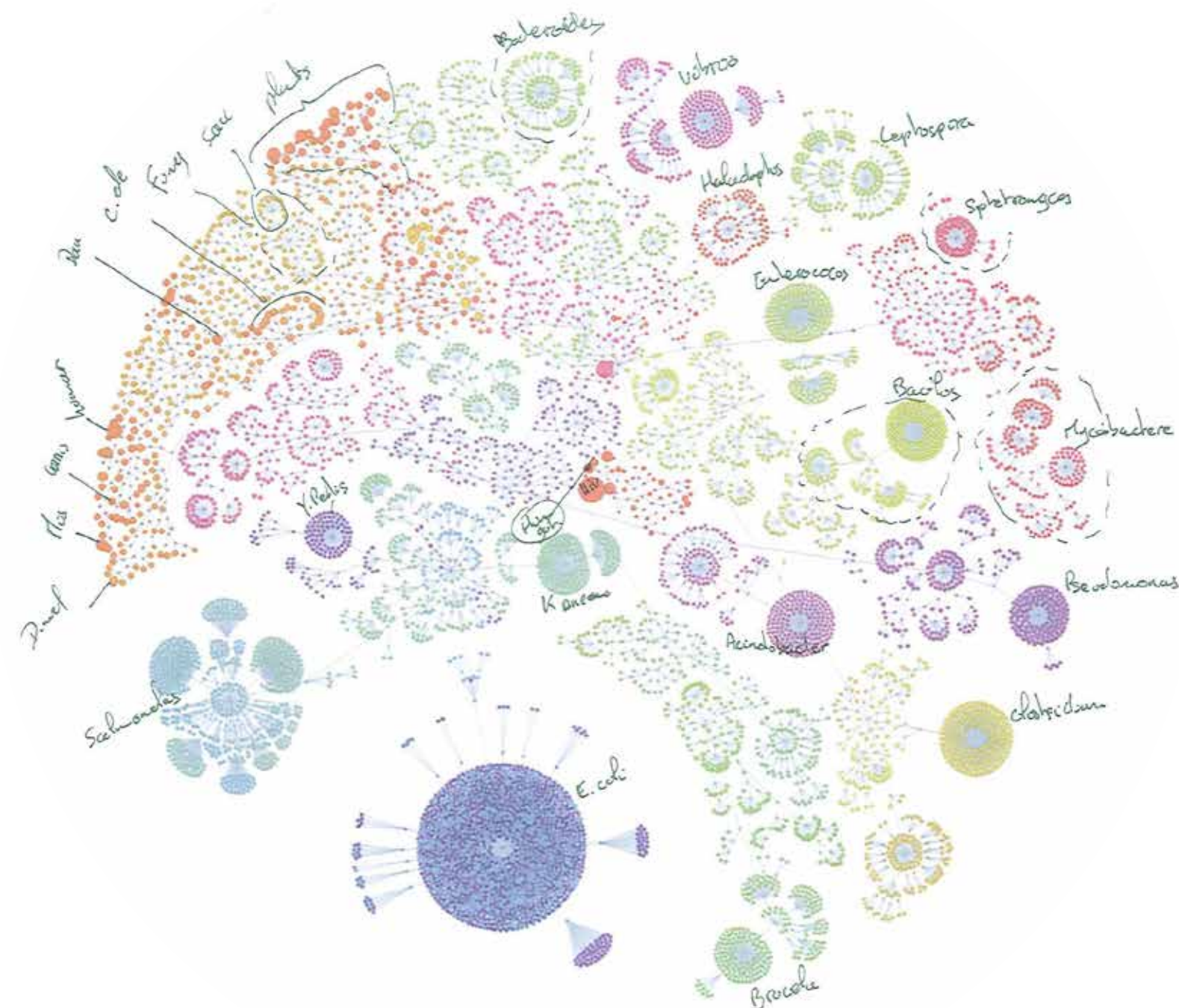
Alex Bateman

Head of Protein Sequence Resources

PhD 1997, University of Cambridge. Postdoctoral work at the Sanger Centre. Group Leader at Wellcome Sanger Institute 2001-2012. Head of Protein Sequence Resources. At EMBL-EBI since 2012.



Distribution of UniProt proteomes before and after removing redundancy



(Right) Distribution of UniProt proteomes across the Tree of Life

Proteins and Protein Families

Data resources

UniProt

UniProt provides a single, centralised, authoritative resource for protein sequences and functional annotation. The UniProt Consortium supports biological research by maintaining a freely accessible, high-quality database that serves as a stable, comprehensive, fully classified, richly, accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces.

www.uniprot.org

Enzyme Portal

The Enzyme Portal integrates publicly available information about enzymes, such as small-molecule chemistry, biochemical pathways and drug compounds.

www.ebi.ac.uk/enzymeportal

GO

The GO annotation programme aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB).

www.ebi.ac.uk/GOA

EBI Metagenomics

EBI Metagenomics is a free resource for the analysis, archiving and browsing of all types of metagenomic data, which aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample.

www.ebi.ac.uk/metagenomics

IntEnz

IntEnz (Integrated relational Enzyme database) is a freely available resource focused on enzyme nomenclature. IntEnz is created in collaboration with the Swiss Institute of Bioinformatics (SIB).

www.ebi.ac.uk/intenz/

InterPro

InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. InterPro uses predictive models, known as signatures, provided by several member databases that make up the InterPro consortium.

www.ebi.ac.uk/interpro

Pfam

Pfam is a database of protein sequence families. Each Pfam family is represented by a statistical model (a profile-hidden Markov model), trained using a curated alignment of representative sequences. These models can be searched against all protein sequences to find occurrences of Pfam families, thereby aiding the identification of evolutionarily related sequences.

<http://pfam.xfam.org>

Rfam

Rfam is a curated database of non-coding RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models.

<http://rfam.xfam.org/>

RNAcentral

RNAcentral is a database of non-coding RNA sequences that serves as a single entry point for searching and accessing the data from an international consortium of established RNA resources. RNAcentral provides a unified view of non-coding RNA sequence data and aims to represent all non-coding RNA types from all organisms.

<http://rnacentral.org>

Team achievements



Sandra Orchard

Protein Function Content team

- Complete human sequence set reviewed to ensure full mapping from UniProtKB entries to Ensembl identifiers. Targeted manual curation added both protein function and disease variant data to key entries
- Eukaryotic protein nomenclature agreed between the INSDC, UniProt Consortium and Vertebrate Gene Nomenclature Committee
- Enhancement of entries by manually created automatic annotation rules extended
- Taxonomic range of complete proteome sets significantly extended
- Completion of a manually-curated set of human proteins for the validation of the computational approaches by the Critical Assessment of Function Annotation experiment (CAFA)
- Demonstrated the scalability of expert, literature-based curation, showing that it can keep up with the growing body of biomedical literature
- Contributed to increased understanding of kinase function through curation of the *Caenorhabditis elegans* kinome
- Major contribution to improvements to description of cell signalling events by the Gene Ontology resource



Maria-Jesus Martin

Protein Function Development team

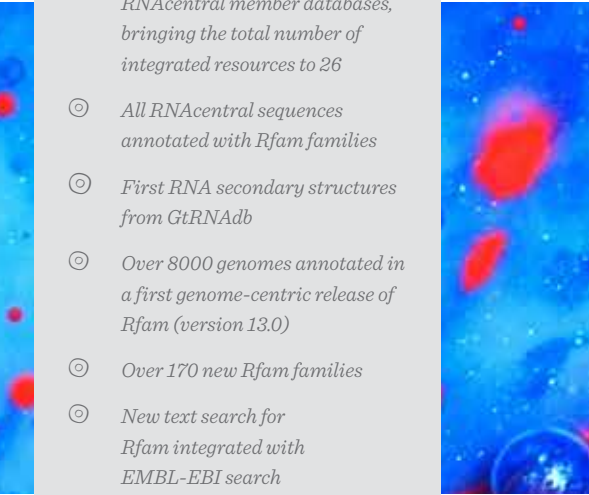
- Strong focus on developments aimed at facilitating biomedical and clinical research. Consequently, we:
- Established an exchange pipeline with ClinVar for sharing variants of clinical interest
- Designed a platform for interpretation of variants based on the role and functional mechanisms of proteins in disease
- Extended the Protein REST API for accessing key integrated protein and genome information (genomic coordinates for each protein, isoform-specific annotations, large-scale variation and proteomics, antigens and tailored proteomes, and UniParc datasets)
- New visualisations for protein interactions and subcellular location
- Provided 15K Reference proteomes with enriched functional annotations
- Collaborated in the evaluation of recently released Critical Assessment of protein Function Annotation (CAFA3) challenge
- Released a new UX-guided QuickGO browser, which uses state-of-the-art web technologies
- Developed new, accurate and comprehensive computational systems for functional annotation of proteins based on domain architectures (Domain Architecture Alignment and Classification- DAAC) and association rules (Association Rule Based Annotator- ARBA)



Rob Finn

Sequence Families team

- Six releases of InterPro, eleven data releases to UniProt and 2868 new InterPro entries added
- New InterPro entry type - homologous superfamily
- Pfam release 31.0 in March 2017 including a total of 16 712 families and 604 clans
- Periodic updates to the HMMER search target databases, including the addition of the MEROPS database. The HMMER API also modified to allow a greater range of services
- EBI Metagenomics reached over 100 000 datasets. Analysis pipeline version 4.0 upgraded the entire taxonomic profiling section and performs prokaryotic and eukaryotic taxonomic classification based on LSU and SSU rRNA genes
- Provision of a new API for analysis of metagenomic data
- Metagenomic assembly added as a new analysis service
- Non-redundant peptide database of almost 50 million sequences produced from an initial 400 metagenomic assembly datasets
- The addition of four new RNAcentral member databases, bringing the total number of integrated resources to 26
- All RNAcentral sequences annotated with Rfam families
- First RNA secondary structures from GtRNAdb
- Over 8000 genomes annotated in a first genome-centric release of Rfam (version 13.0)
- Over 170 new Rfam families
- New text search for Rfam integrated with EMBL-EBI search



Protein Function Content

The Protein Function Content team manages the biocuration of the Protein Sequences databases, interpreting and integrating information relevant to biology. The primary goals of biocuration are accurate and comprehensive representation of biological knowledge, as well as facilitating easy access to the data for working scientists and providing a basis for computational analysis.

The curation methods we apply to UniProtKB/Swiss-Prot include manual extraction and structuring of experimental information from the literature, manual verification of results from computational analyses, quality assessment, integration of large-scale datasets and continuous updating as new information becomes available.

UniProt has two complementary approaches to automatic annotation of protein sequences with a high degree of accuracy. UniRule is a collection of manually-curated annotation rules, which define annotations that can be propagated based on specific conditions. The Statistical Automatic Annotation System (SAAS) is an automatic, decision-tree-based, rule-generating system. The central components of these approaches are rules based on the manually-curated data in UniProtKB/Swiss-Prot from the experimental literature and InterPro classification.

The UniProt GO annotation (GOA) program aims to add high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). We supplement UniProt manual and electronic GO annotations with manual annotations supplied by external collaborating GO Consortium groups. This ensures that users have a comprehensive GO annotation dataset. UniProt is a member of the GO Consortium.

Major achievements

The UniProt Consortium continues to supply protein sequence and function data to the entire scientific community with a strong focus on the needs of biomedicine and biotechnology. An application for renewal of NIH funding was submitted at the end of 2017.

The Consortium continues to support links between genomic and proteomic data, and the curation team worked to review our human sequence set to enable a more complete mapping from UniProtKB entries to Ensembl identifiers. This will enable improved integration of genome-related resources with those that are protein-centric, allowing users to track nonsynonymous variants from genomes through to their effect on protein function. Targeted manual curation has added both protein function and variant data to many human entries and improved annotation of links between variants and disease, when known.

Eukaryotic protein nomenclature guidelines have now been agreed between key players in the sequence database field, including the INSDC, UniProt Consortium and Vertebrate Gene Nomenclature Committee. These guidelines will be used to enhance our manual annotation and will be propagated to unreviewed entries through the UniRule system.

Future plans

Our team aims to work with UniProt Consortium partners to improve the standardisation of variant interpretations between UniProt, ClinVar and ClinGen, mapping UniProt variant interpretation annotations to ACMG-AMP categories.

We will also be extending the curation of UniRules for the annotation of eukaryotic proteins, with a focus on those of biomedical importance, and archaeal, bacterial and viral proteins, particularly those of pathogens.

Last but not least, we will evaluate emerging standards for protein function annotation, for example the Gene Ontology Common Annotation Model and improved annotation of protein complexes.



Sandra Orchard

Team Leader – Protein Function Content

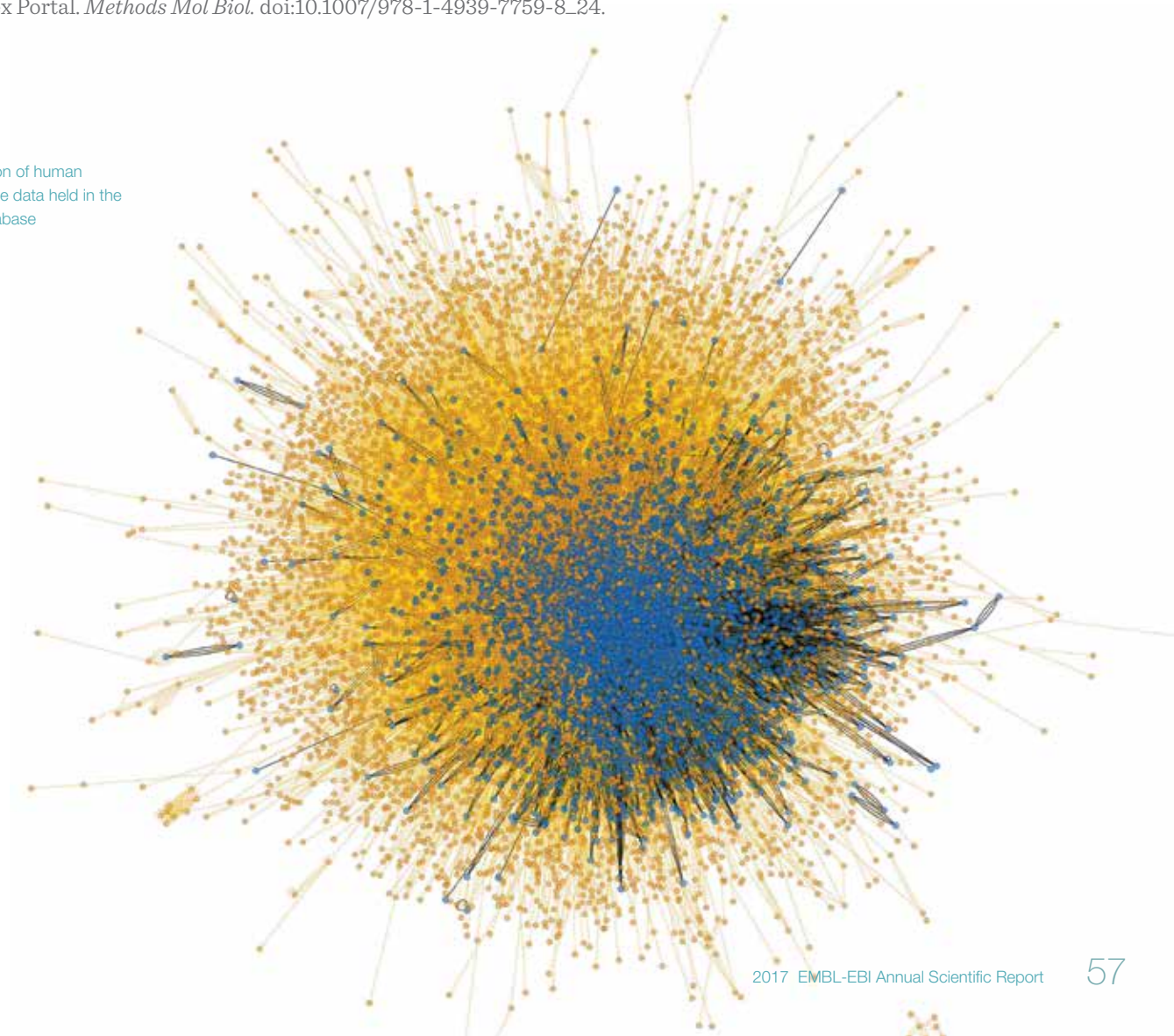
BSc (Hons) Biochemistry, University of Liverpool 1982. Team Leader Inflammatory Diseases Dept, Roche Products Ltd UK, until 2001. Team Leader of the Molecular Interaction Team as of April 2015, and of Protein Function Content since August 2017. At EMBL-EBI since 2002.

Selected publications

Sivade DM et al. (2018). Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*. doi:10.1186/s12859-018-2118-1.

Meldal BHM, et al. (2018). Searching and Extracting Data from the EMBL-EBI Complex Portal. *Methods Mol Biol*. doi:10.1007/978-1-4939-7759-8_24.

Visualisation of human interactome data held in the IntAct database



Protein Function Development

The work of the team spans several major resources under the umbrella of UniProt, the comprehensive resource of protein sequences and functional annotation: the UniProt Knowledgebase, the UniProt Archive and the UniProt Reference Clusters. The team develops software, tools and services for protein information and protein function prediction in the UniProt, Gene Ontology (GO) annotation and enzyme data resources at EMBL-EBI.

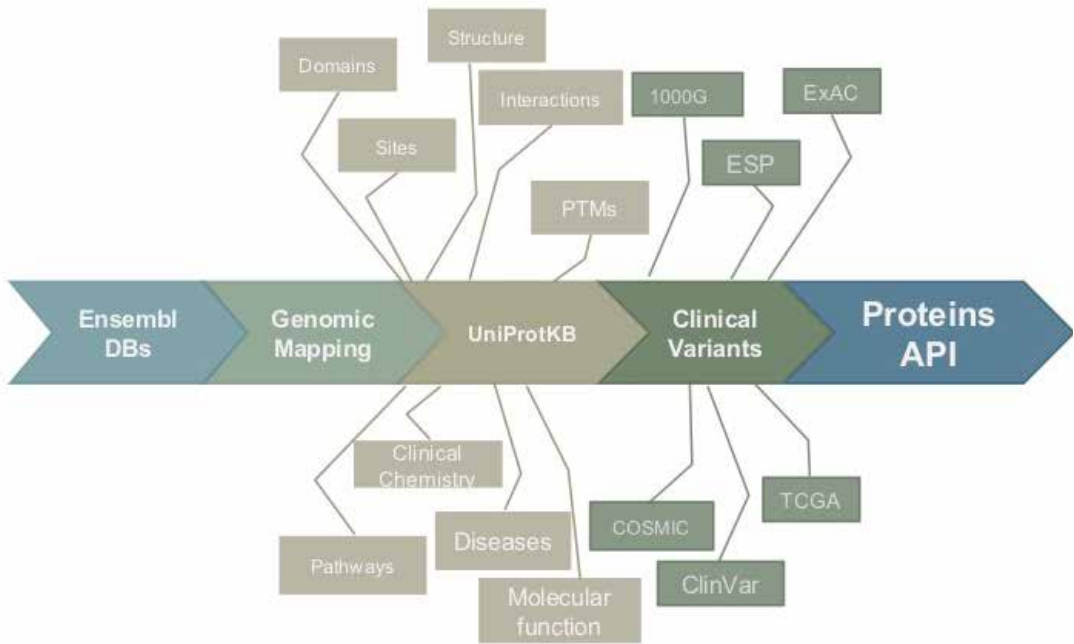
UniProt aims to enable advances in scientific research and innovation by integrating and enhancing protein data and knowledge into a freely accessible resource.

Major achievements

Our team has a strong focus on developments aimed to facilitate biomedical and clinical research. UniProt is key to deciphering the functional genome. To facilitate this, in 2017 we established an exchange pipeline with the ClinVar data resource for sharing variants of clinical interest. We also designed a platform for interpretation of variants based on the role and functional mechanisms of proteins in disease.

We extended the Protein REST API for accessing key integrated protein and genome information, including genomic coordinates for each protein, isoform-specific annotations, large scale variation and proteomics, antigens and tailored proteome, and UniParc datasets.

Integration of genomes and proteins in UniProt



New visualisations for protein interactions and subcellular locations are now available alongside 15 000 reference proteomes with enriched functional annotations.

The team also collaborated in the evaluation of the recently released Critical Assessment of protein Function Annotation (CAFA3) challenge.

The newly released UX-guided QuickGO browser uses state-of-the-art web technologies.

The team developed new, accurate and comprehensive computational systems for functional annotation of proteins based on domain architectures (Domain Architecture Alignment and Classification – DAAC) and association rules (Association Rule Based Annotator – ARBA).

Future plans

In 2018, we plan to release the ProtVista protein browser with reusable web components, which will enable users to select specific features (e.g. variation data) and integrate them within their workflows. The browser will also be extended to incorporate and visualise functional residues in 3D structures.

We will continue to engage with user communities working in functional prediction and explore methods and data-exchange mechanisms to improve accuracy



Maria-Jesus Martin

Team Leader – Protein Function Development

BSc in Veterinary Medicine, University Autònoma de Madrid. PhD in Molecular Biology (Bioinformatics), 2003. Team Leader since 2009. At EMBL-EBI since 1996.

and coverage of protein annotations. We will also continue development of computational methods for function prediction and share these with the users through standard formats and tools.

The team will explore deep learning methodologies to expand our annotations beyond the functional predictions (e.g. associations to drugs). It is important to maintain our focus on usability and engage with our users to ensure that we maintain a global genome/proteome- and gene product-centric view of the sequence space as well as serve biomedical researchers focusing on variation data and disease annotations. We will develop a platform for interpretation of variation data at the protein functional level.

Selected publications

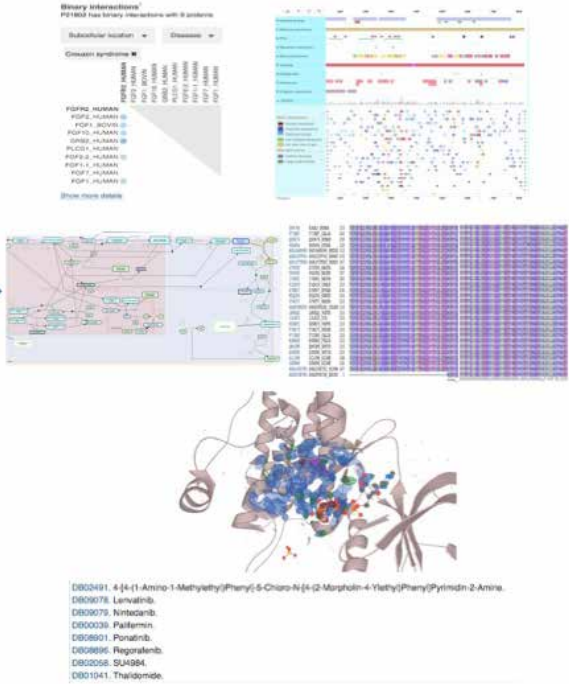
Watkins X et al. (2017). UniProt Consortium. ProtVista: visualization of protein sequence annotations. *Bioinformatics*. doi:10.1093/bioinformatics/btx120

Nightingale A et al. (2017). The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res*. doi:10.1093/nar/gkx237

Saidi R et al. (2017). Rule Mining Techniques to Predict Prokaryotic Metabolic Pathways. *Methods Mol Biol*. doi:10.1007/978-1-4939-7027-8_12

Mirdita M et al. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res*. doi:10.1093/nar/gkw1081

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. doi:10.1093/nar/gkw1099



Sequence Families

The Sequence Families team is responsible for the InterPro, Pfam, RNACentral and Rfam data resources, and also coordinates the EBI Metagenomics project. In addition, we are responsible for the range of protein-homology searches provided by the HMMER web services.

Major achievements

InterPro, Pfam and HMMER

InterPro entries are labelled with a ‘type’ (family, domain, repeat or site), reflecting the biological entity that the constituent signatures represent. In September 2017, we added a new entry type, homologous superfamily, to complement the existing types. A homologous superfamily is defined as a group of proteins that share a common evolutionary origin, indicated by their structural similarities.

In addition to this change, we have continued to integrate other member database signatures, adding 2868 new InterPro entries in the last year. InterPro has continued to be regularly released, issuing six public releases in 2017, along with eleven internal data releases to UniProt.

We released Pfam 31.0 in March 2017. This was the third release based on UniProtKB reference proteomes and included a total of 16 712 families and 604 clans. In Pfam 31.0, over 36% of Pfam entries are placed within a clan. Curation of this hierarchy is an ongoing effort.

Since release 31.0, a total of 830 new Pfam entries have been added. Furthermore, 18 new clans were created, 783 families were added to clans (totalling 6779 entries, 39% in total), and 52 entries have been updated from being families of unknown function (DUFs) to families with more functionally-relevant names and annotation.

The HMMER website has continued to provide millions of searches per month (search volume has increased by around 6.5% over the last year). The HMMER sequence database has been updated monthly, with releases in sync with the UniProt database. During the last year we have also added the complete set of Ensembl peptides (previously just those from Ensembl Genomes were included).

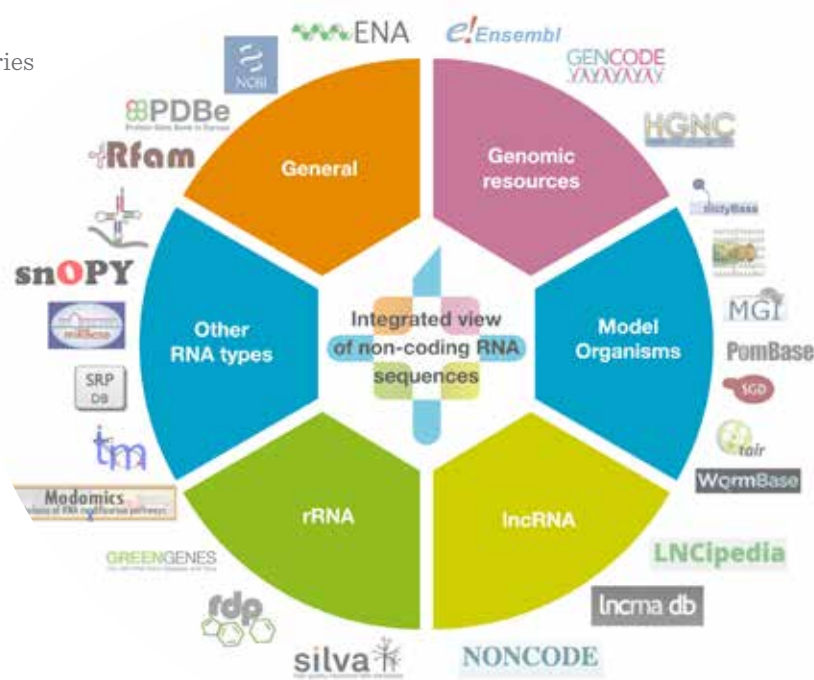
Metagenomics

EBI Metagenomics has grown considerably over the last year and now contains over 1200 publicly available projects, comprising ~75 000 samples and ~100 000

runs (compared with 60 000 last year). As well as the hundreds of small projects, EBI Metagenomics contains the analysis of many showcase studies, such as Tara Oceans, Ocean Sampling Day, American & British Gut, MetaHIT and METASOIL.

The metagenomics analysis pipeline underwent a substantial update in August 2017 to version 4.0, with the entire taxonomic profiling section replaced. As a result of this upgrade, the pipeline is now able to offer both prokaryotic and eukaryotic taxonomic classification based on LSU and SSU rRNA genes using the MAPSeq taxonomic analysis software and the SILVA database.

To provide a richer search and retrieval interface, in December 2017 we released a RESTful API, providing programmatic access to the data. This has several top-levels (termed resources), such as studies, samples, runs, experiment-types, biomes and annotations. Appropriate relationships and links are provided between the resources, allowing complex queries to be



Schematic of the RNACentral consortium, highlighting the different data types contributed by each resource.

constructed (for example: “retrieve all oceanographic samples from metagenomic studies taken at temperature less than 10°C”). The provision of such complex queries allows metadata to be combined with annotation for powerful data analysis and visualisation.

In 2017, we undertook a feasibility study aimed at investigating assembly of metagenomic datasets given current infrastructural resources. Based on the results, we believe it is feasible to offer assembly of user-submitted metagenomic datasets, subject to request. We have chosen a panel of three assemblers for use with the pipeline: metaSPAdes, MEGAHIT and Minia. As part of the assembly tool evaluation process, we assembled a number of publicly available metagenomic datasets from ENA. To date, we have assembled 2298 different shotgun metagenomics datasets from 78 different projects.

As a compendium to the assemblies and their associated analysis results, we also developed a workflow to produce a non-redundant set of peptides. From an initial 400 assembly datasets, a non-redundant peptide database of almost 50 million sequences has been produced. Over 15 million of these are predicted to be full length, yet only ~1 million have exact counterparts in the UniProtKB database. To allow the querying of this sequence database by users with a target sequence, we have deployed a HMMER web search engine and server. This interface can be accessed via a tab on the front page of the EBI metagenomics website.

RNA resources

RNACentral continued growing in 2017, with an addition of four new member databases. All sequences were comprehensively annotated with Rfam families, and new quality control metrics were introduced based on Rfam annotations, including detection of partial sequences and potential contamination. To reach a broader audience and engage with our users, we delivered multiple presentations at international scientific conferences, including an exhibition booth with interactive website demonstration at the RNA Society meeting in Prague.

In the past year, we released a major new version of Rfam. Rfam 13.0 adopts a new genome-centric approach that annotates a non-redundant set of over 8000 reference genomes with RNA families. A new text search was developed to allow the users to explore the data more efficiently. Over 170 new RNA families were added to Rfam, bringing the total to 2686.

Future plans

In 2018 we will release a new InterPro website that encompasses much of the functionality from Pfam, thereby allowing it to supersede the Pfam website.



Rob Finn
Team Leader –
Sequence Families
PhD in Biochemistry, Imperial College London; Wellcome Sanger Institute, 2001-2010; Janelia Research Campus, 2010-2013. At EMBL-EBI since 2014.

We will make a release of Pfam as well as continue to add hundreds of new entries to Pfam from a variety of sources, in particular those from plant genomes. In InterPro, we will implement additional semi-automatic procedures to streamline integration of member database signatures into InterPro entries.

Subsequent to the provisioning of the EBI Metagenomics assembly service, we will investigate alternative ways to present this data via the website. We will also evaluate different approaches for recovering genomes of metagenomics assemblies. This has the potential to provide important references datasets for further analysis into different microbiomes. Finally, we will develop a new website for EBI Metagenomics using the API.

We will ensure that the HMMER website continues to scale with ever-growing sequence databases. We will expand the genome-related functionality in Rfam and RNACentral and will issue two releases of each database. We will also improve the RNACentral sequence search and develop a new online portal enabling expert users to build new Rfam families.

Selected publications

Mitchell AL, et al. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* doi: 10.1093/nar/gkx967

Finn RD, et al. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* doi: 10.1093/nar/gkw1107

RNACentral Consortium (2017). RNACentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* doi: 10.1093/nar/gkw1008

Finn RD, et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* doi: 10.1093/nar/gkv1344

Kalvari I, et al. (2017). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1038

Molecular and Cellular Structure

Understanding the structure of molecules is key to understanding their function. The team aims to bring structure to biology by making this complex field more accessible to non-specialists. The cluster manages three major archives in structural biology: the Protein Data Bank (PDB), the Electron Microscopy Data Bank (EMDB) and the Electron Microscopy Public Image Archive (EMPIAR).

In 2017, the Molecular and Cellular Structure (MCS) cluster finished its restructuring, which began in 2016, with the appointment of Ardan Patwardhan as Team Leader. Ardan leads the Cellular Structure and 3D Bioimaging team (CS3DB), which manages the EMDB (with international partners) and EMPIAR archives.

This team is growing rapidly and will be well placed to support the archiving needs of the 3D cellular bioimaging community. EMDB celebrated its 15th anniversary in 2017 and, together with the establishment of EMPIAR (in 2014), showed that EMBL-EBI wisely anticipated the explosive growth of the electron cryo-microscopy (cryo-EM) field, a fact further recognised by the 2017 Chemistry Nobel Prizes for three pioneers of the field (Jacques Dubochet, Joachim Frank and Richard Henderson). Both Joachim Frank and Richard Henderson have been involved with the work on EMDB and EMPIAR at EMBL-EBI, serving on advisory committees and task forces, and supporting crucial grant applications.

EMDB archives 3D volume maps derived from cryo-EM experiments, whereas EMPIAR holds the raw image data from cryo-EM (and other) experiments. Established as recently as 2014 at the request of the cryo-EM community, EMPIAR has since developed into an important resource by addressing new needs (e.g. archiving data from other imaging modalities, and storing data for community challenges). In addition, EMPIAR has become an exemplar archive for the wider field of bioimaging and will be a cornerstone of EMBL-EBI's BioImage Archive (BIA) initiative.

The Protein Data Bank in Europe (PDBe), led by Sameer Velankar, is the other team in the MCS cluster. This team is part of the Worldwide Protein Data Bank partnership (wwPDB), and is a leading contributor towards the management of the Protein

Data Bank (PDB) that archives atomic models (and the underpinning experimental data) of biomacromolecular structures determined by X-ray crystallography, NMR spectroscopy, and cryo-EM. Additionally, the PDBe team works with partners in EMBL-EBI, the UK and elsewhere to provide value-added annotations of the molecular structure data in the PDB.

Major achievements

A total of 11 129 PDB entries were released by wwPDB in 2017 (an increase from 10 852 in 2016), with new PDB depositions numbering 13 049 in 2017 (up from 11 614). Amongst the new structure depositions, 4044 were processed at EMBL-EBI (4051 in 2016), accounting for 31% of the worldwide total.

In 2017, EMDB released 1106 maps (up from 1071 in 2016). The number of new EMDB depositions rose by 29%, from 1074 in 2016 to 1390 in 2017. Of these new EMDB depositions, 523 (38% of the worldwide total) were annotated at EMBL-EBI (up from 358 in 2016).

EMPIAR released 42 entries in 2017 (increase from 30 in 2016) and processed 71 new depositions compared to 32 in 2016.

Future plans

Key components of the MCS cluster strategy are: to be actively involved and lead current and new data-archiving activities in the field of molecular and cellular structural biology; to engage actively with the scientific communities that use and produce such data as well as with strategically important collaborators; to continuously improve the functional annotation, integration and dissemination of such data; and to actively engage in training, outreach and public engagement. The reports of the PDBe and CS3DB teams on the following pages show how their activities and plans align with our cluster's overall strategy.

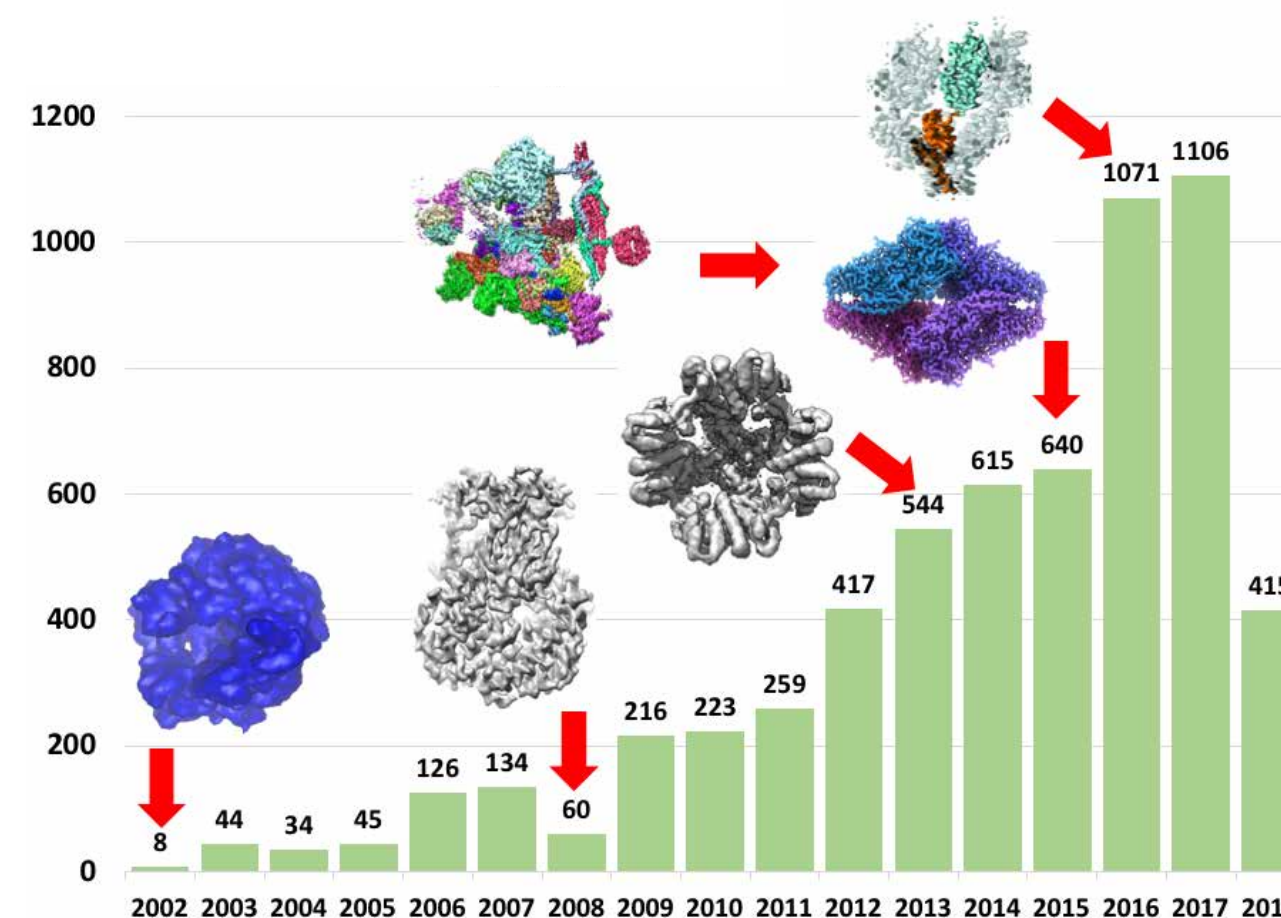


Gerard Kleywegt

Head of Molecular and Cellular Structure

PhD University of Utrecht, 1991. Postdoctoral researcher, then independent investigator, University of Uppsala, 1992-2009. Co-ordinator, then Programme Director of the Swedish Structural Biology Network, 1996-2009. Research Fellow of the Royal Swedish Academy of Sciences, 2002-2006. Professor of Structural Molecular Biology, University of Uppsala, 2009-2012. At EMBL-EBI since 2009.

The number of EMDB entries released per calendar year since its inception in 2002



Molecular and Cellular Structure

Data resources

Electron Microscopy Data Bank

EMDB is a public repository for electron microscopy electric potential maps of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography and electron crystallography.

www.emdb-empiar.org

Electron Microscopy Public Image Archive

EMPIAR is a public resource for raw, 2D images from molecular and cellular 3D bioimaging experiments using transmission or scanning electron microscopy and electron or soft X-ray tomography. The purpose of EMPIAR is to facilitate methods development, validation, training and community challenges, which will lead to better 3D structures.

www.empiar.org

Protein Data Bank in Europe

PDBe is the European resource for the collection, organisation and dissemination of 3D structural data on biological macromolecules and their complexes. Together with international partners, PDBe manages the global public repository of molecular structure data, the Protein Data Bank (PDB). PDBe's goal is to develop infrastructure and tools to enable translation of macromolecular structure data into knowledge.

www.pdbe.org

Team achievements



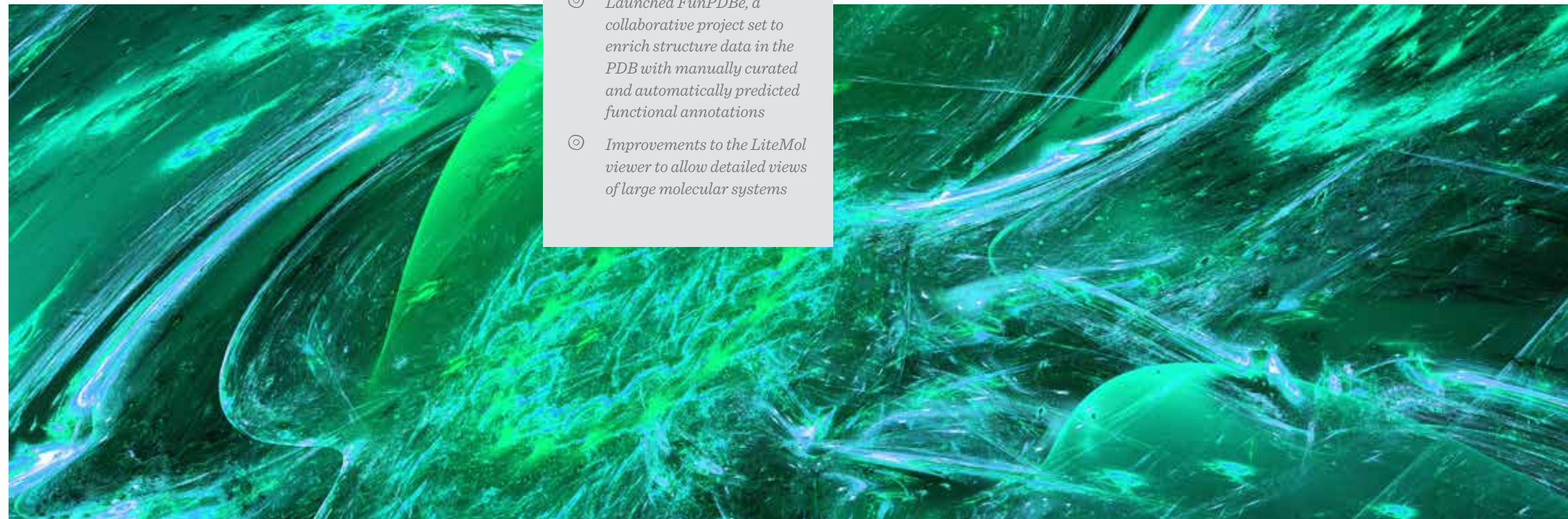
Sameer Velankar Protein Data Bank Europe team

- ⦿ Development and management of the OneDep system for deposition, validation and curation of new structure submissions, including a possibility for authors to be linked to their ORCID identifiers
- ⦿ Development of a OneDep module supporting interoperability with SASBDB (partner data resource at EMBL Hamburg)
- ⦿ Improved mappings between PDB and UniProt entries, expanding the PDB structure coverage of protein space from 43 000 canonical UniProt accessions to over 1.5 million entries
- ⦿ Launched FunPDBe, a collaborative project set to enrich structure data in the PDB with manually curated and automatically predicted functional annotations
- ⦿ Improvements to the LiteMol viewer to allow detailed views of large molecular systems



Ardan Patwardhan Cellular Structure and 3D Bioimaging team

- ⦿ Curated 523 EMDB entries
- ⦿ Released 42 new EMPIAR entries
- ⦿ Released EMDB-Segmentation File Format (EMDB-SFF) to the public
- ⦿ Held expert workshop on "Public Archiving of Cellular Electron Microscopy and Soft X-ray Tomography Data"



Protein Data Bank in Europe

PDBe aims to serve the biomedical community by curating and providing easy access to high-quality, complete and enriched macromolecular structure data, designing tools and services for efficient data discovery and visualisation, and engaging with different user communities via development training material.

PDBe continues to be actively engaged in the wwPDB collaboration – it is responsible for processing all PDB entries originating from European and African institutions, which amounted to 4044 depositions in 2017 (31% of the 13 049 worldwide). In the same period, 523 EMDB entries were processed at PDBe. Despite this high workload, PDBe continues to process over 90% of all received depositions within 48 hours.

Major achievements

Together with the other wwPDB partners, PDBe continued the development and management of the OneDep system for deposition, validation and curation of new structure submissions, including an option for authors to be linked to their ORCID identifiers. In 2018, supplying ORCID identifiers will become mandatory for corresponding authors of PDB entries. Through a BBSRC-funded project, PDBe has developed a OneDep module supporting interoperability with SASBDB (partner resource at EMBL Hamburg). This module serves as a prototype for supporting a federation of archives in structural biology. The wwPDB partners also released version 5 of the mmCIF dictionary

underpinning the PDB archive and made significant progress towards supporting full versioning of individual PDB entries. The wwPDB partners also published four joint peer-reviewed papers.

In the BBSRC-funded SIFTS project, we improved the mappings between PDB and UniProt entries. These now include mappings to isoforms and to UniRef90 clusters, expanding the PDB structure coverage of protein space from ~43 000 canonical UniProt accessions to over 1.5 million entries. The SIFTS resource has also added mappings to genomic positions, preliminary annotations of Pfam domains identified with PHMMER and additional PubMed cross-references.

In October 2017, PDBe started FunPDBe, a new BBSRC-funded collaborative project that will enrich structure data in the PDB with manually curated and automatically predicted functional annotations, which are currently held across a large number of smaller niche resources. The collaborating teams have already agreed on and developed minimal data-exchange standards, and PDBe has implemented a prototype deposition system.

PDBe improved its weekly release process in 2017 using AirFlow technology. It is anticipated that in 2018 this improvement will result in a number of efficiency gains with respect to maintenance of the release process. In addition, PDBe released three API updates providing access to the new SIFTS mappings and cross-references discussed above, as well as information relating to validation of NMR structures and X-ray experimental data.

Improvements to the PDBe website include the reorganisation of the pages describing experimental details, which now support hybrid structure-determination techniques. A number of new web components have been developed, such as a summary view of raw experimental datasets used in determining PDB structures and held at EMPIAR, SBGrid Data Bank and the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRM), as well as a detailed view of experimental data held at SASBDB.

PDBe put into production a coordinate server and a density server, which stream compressed PDB coordinate data and X-ray and cryo-EM maps on demand. They underpin the LiteMol suite and in particular the LiteMol viewer, integrated into PDBe web pages. They also offer access to these datasets for other resources.

The LiteMol viewer has undergone a number of improvements. Relying on the coordinate and density servers, the viewer now displays large molecular systems at reduced resolution and allows the user to quickly zoom into detailed views at the best possible resolution. Displaying the 3D maps for structures solved by X-ray crystallography or cryo-EM allows users to immediately identify regions of structures where experimental evidence may be unconvincing.

The PDBe team continued its work on training and outreach and participated in more than 40 outreach events and workshops, and sponsored three poster prizes at major international conferences. The team also conducted four webinars, which are available on the PDBe YouTube channel. As part of the outreach programme, we continued a collaborative project with art departments in four senior schools in Cambridge, UK, culminating in an art exhibition in central Cambridge. This public engagement project aims to introduce school students to the wealth of information available in the PDB. The number of followers of PDBe's various social media accounts increased by over 20%.



Sameer Velankar

Team Leader – Protein Data Bank in Europe

PhD, Indian Institute of Science, 1997. Postdoctoral researcher, Oxford University, 1997–2000. At EMBL-EBI since 2000.

Future plans

The plans for 2018 encompass continued active participation in the wwPDB partnership through biocuration of PDB and EMDB entries, further enhancements of the OneDep system and maintenance of the PDB archive, and active outreach and public engagement. The completion of the first year of the FunPDBe project will include the release of the deposition system and the first release of functional-site-prediction data.

Planned PDBe website improvements include:

- ⦿ Adding functional annotation on co-factor-like molecules
- ⦿ Inclusion of RNA annotations from Rfam
- ⦿ Redesign of the sequence-feature viewer, which will be shared with the UniProt and InterPro teams
- ⦿ Integration of THOR web components for claiming and displaying ORCID data
- ⦿ Updates to the PDBe search system to provide advanced search functionality and include relevant functional data, and to improve the presentation of the search results
- ⦿ Redesign of pages describing small molecules in the PDB
- ⦿ Integration of functional annotation data from the FunPDBe project

Selected publications

Mir S, et al. (2018). PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.* doi:10.1093/nar/gkx1070

Sehna D, et al (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat Methods.* doi: 10.1038/nmeth.4499

Gore S, et al. (2017) Validation of Structures in the Protein Data Bank. *Structure.* doi: 10.1016/j.str.2017.10.009

(Left): PDBe updated its 3D viewer, enabling users to load molecular structures instantly on any Internet-connected device.



Cellular Structure and 3D Bioimaging

The Cellular Structure and 3D Bioimaging (CS3DB) team aims to provide problem-centric views of biology by developing tools and resources that integrate multi-scalar structural and bioimaging data and improved validation of structural data to facilitate its reuse. The team is responsible for the EMDB and EMPIAR data resources and also develops stand-alone tools and web-based resources for searching, validating, visualising and programmatically accessing structural and bioimaging data.

The team is anticipated to grow rapidly over the following years as EMBL-EBI increasingly focuses on digital biology (in which imaging plays a major role), and aspires to become a global hub for archiving imaging data.

Major achievements

EMDB reached 5542 entries of which 1106 were released in 2017 alone. EMDB received 1390 depositions, of which 523 (38%) were annotated at EMBL-EBI. An impressive 49% of the entries released in 2017 were at reported resolutions of better than 4 Å: these include an increasing number of membrane-channel complexes (e.g. the TRPM4 cation channel, EMD-8871), the Human Mitochondrial Respiratory Supercomplex (e.g. EMD-6775), and a structure of Tau filaments that constitute the neurofibrillary lesions abundant in Alzheimer's disease (EMD-3741). Among the more exotic entries was a tomographic reconstruction of the Brazilian Giant Samba virus (EMD-8599), which has a capsid diameter of ~500 nm.

In 2018 EMPIAR released 42 new datasets, compared to 30 in 2016. An increasing fraction of these entries are from cellular EM experiments. In December 2017, we held an expert workshop on “Public Archiving of Cellular Electron Microscopy and Soft X-ray Tomography Data” which helped us gain a better understanding of the data and metadata requirements for these imaging modalities and we plan to improve the capture of cellular EM data in 2018. EMPIAR continues to be a vital source of cryo-EM datasets relating to cutting-edge developments in fields such as the use of phase-plates (e.g. haemoglobin, EMPIAR-10084) and the reconstitution of membrane proteins in lipid nanodiscs (which is partly responsible for the increasing number of such structures in EMDB, e.g., EMD-8702 and EMPIAR-10093).

As EMPIAR continues to grow, it is increasingly important to consider the back-end storage requirements in order to scale up seamlessly into the petabyte scale and beyond. In January 2017, EMBL-EBI, EMBL Heidelberg and the University of Dundee organised an expert workshop on bioimage archiving and one of the recommendations was to set up a scalable, generic public bioimage archive at EMBL-EBI. CS3DB and several other teams have received internal funding to pilot the development of this BioImage Archive (BIA) and to migrate EMPIAR to use it as the back-end.

An important consideration with regards to cellular structure data is that there is no obvious means of linking it to the wealth of other publicly available bioinformatics resources. Following recommendations from the 2015 expert workshop on “3D segmentations and transformations – building bridges

between cellular and molecular structural biology” held at EMBL-EBI, we have developed the EMDB Segmentation File Format (EMDB-SFF) in consultation with community experts and released it publicly in 2017. We have also developed a Segmentation Annotation Tool (SAT), which facilitates the biological annotation of segmentations and a web-based Volume Browser that enables the integrated visualisation of cellular and molecular structure data.

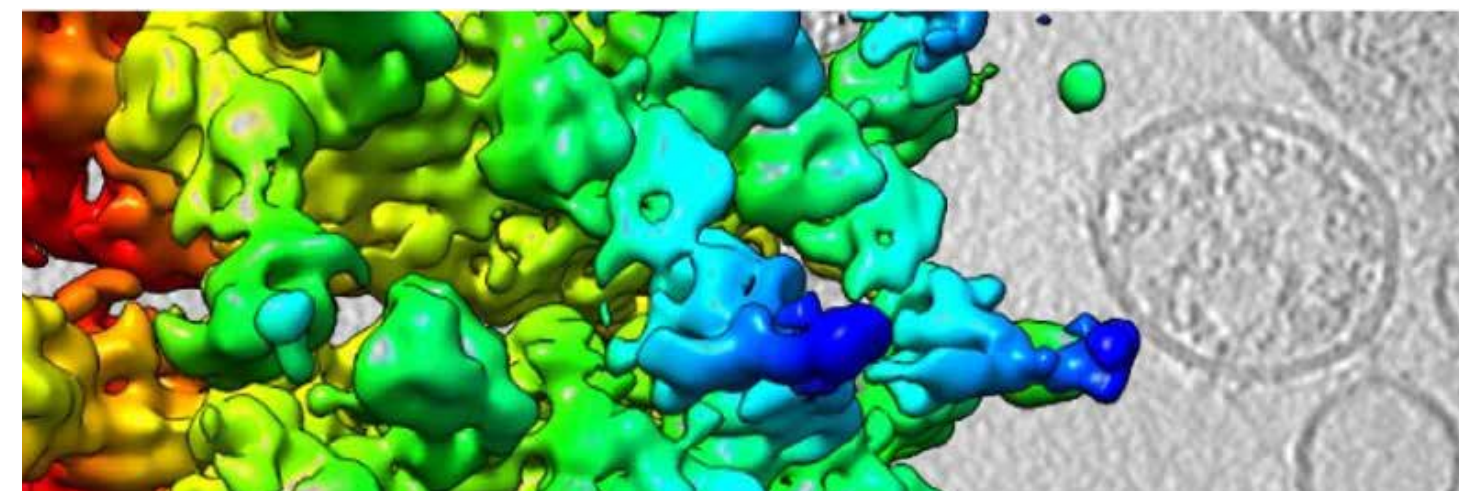
Our team plays a key role in facilitating community-wide initiatives, for example, working with journals to encourage deposition. We organise expert thematic workshops to encourage best practices and knowledge exchange, and participate in the EMBL-EBI Training Programme.

Future plans

In 2018 the CS3DB team is expected to more than double in size. This will bolster support to handle the increasing number, diversity and complexity of EMDB and EMPIAR depositions, firm-up production processes and expand outreach and training activities.

Two grants (one each from the Wellcome Trust and BBSRC) will enable us to develop substantial archive-related EM validation resources. Another grant, funded jointly by MRC and BBSRC, will allow us to develop further support for cellular imaging data in EMPIAR and to develop data-submission pipelines from microscopy centres such as the Electron Bio-Imaging Centre (eBIC), based at the Diamond Light Source. We will work

Structure of the Ebola virus nucleocapsid overlaid on an electron tomography image (EMD-3873; emdb-empiar.org/emd-3873)



Ardan Patwardhan

Team leader – Cellular Structure and 3D Bioimaging

Lecturer, Imperial College London, 1999 to 2010. Masters in Engineering Physics and PhD from the Royal Institute of Technology, Stockholm, Sweden. At EMBL-EBI since 2011.

with other EMBL-EBI teams to build the EMBL-EBI BioImaging Archive and start using it as the back-end for EMPIAR.

The biological annotation of cellular structure data is of paramount importance for data integration but it is an Achilles heel in the exploitation of data currently archived. For instance, there is no straightforward way to query EMDB for all membrane-channel complexes. We will exploit a combined approach of stepped-up manual annotation using ‘tagathons’, crowd-sourced annotation and machine learning to address this issue.

Selected publications

Joseph AP, et al. (2017). Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J Struct Biol.* doi: 10.1016/j.jsb.2017.05.007

Patwardhan, A. (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallographica Section D.* doi: 10.1107/S2059798317004181

Patwardhan, A, et al. (2017). Building bridges between cellular and molecular structural biology. *Elife.* doi: :10.7554/eLife.25835

Structure of Tau filaments that constitute the neurofibrillary lesions abundant in Alzheimer's disease (EMD-3741)

Molecular Systems

Our Molecular Systems teams provide network-oriented public databases as reference implementations for community standards, in particular the IntAct molecular interaction database, the Reactome pathway database, and the BioModels repository of computational models of biological systems. In addition, we provide infrastructure for data discovery (OmicsDI) and stable referencing (Identifiers.org) of data objects.

Major achievements

In 2017 we published the Omics Discovery Index (OmicsDI), a new resource providing dataset discovery across a heterogeneous, distributed group of transcriptomics, genomics, proteomics and metabolomics data resources spanning more than 100 000 datasets from 15 repositories in four continents, including both open and controlled access data resources. OmicsDI provides harmonised metadata across its partner repositories, allows users to “claim” datasets they have contributed to, and to propagate these associations to their ORCID profile. It also provides extensive links to related datasets, based on metadata similarity and other criteria.

The IntAct database of molecular interactions has released a major new dataset, providing more than 15 000 instances where mutations have been experimentally shown to affect a protein-protein interaction. This dataset has been curated in collaboration with other partners in the context of the International Molecular Exchange consortium (IMEx) over several years, and has undergone extensive harmonisation in terms of curation and representation standards. It is expected to be a valuable resource both in the analysis of functional effects of mutations, and in the development and testing of prediction software in the domain.

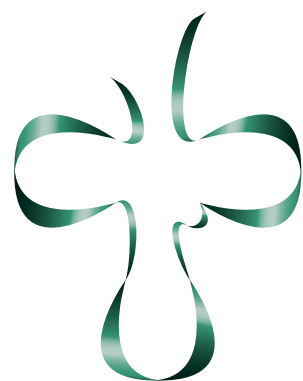
The Complex Portal has been completely redeveloped, focusing on integration with other resources; widgets from the Expression Atlas, Reactome, and PDBe have been integrated to provide additional information on annotated complexes.

With the deployment of the new website design, we have completed a major redevelopment of the Reactome user interface, which now provides a multi-scale, stable, high performance platform for the visualisation and analysis of biomolecular pathway data. We added professionally designed, interactive pathway overview diagrams for rapid navigation of the Reactome pathway hierarchy. Visual elements of these diagrams, such as representations of organelles or tissue types, are also separately available as a library which currently features more than 800 reusable icons for biomolecular illustrations.

We have also added the capability to download Reactome pathway diagrams in .pptx format, with all entities represented as PowerPoint objects, rather than raster images. This allows scientists to reuse and rearrange Reactome diagrams to include their own emphasis or discoveries, for use in publications and presentations. All Reactome major components are available as web services or JavaScript-based widgets suitable for integration into third-party applications. For efficient computational access to Reactome, we have also released a regularly updated copy of Reactome data as a graph database in Neo4J format.



rRNA icon from the Reactome icon library



tRNA icon from the Reactome icon library

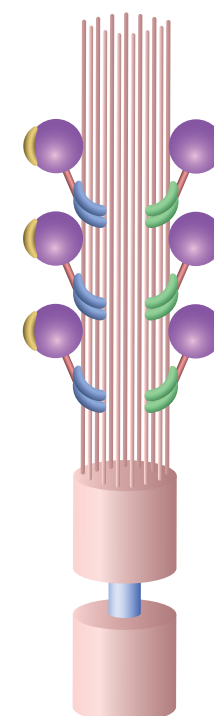
In collaboration with the Uhlen group at the Karolinska Institute, the BioModels resource of systems biology models has more than quadrupled its content of literature-based models through the addition of more than 6500 Genome Scale Metabolic Models (PDGSMM) derived from patients with 21 different types of cancer. In addition, we have released the new BioModels web platform, providing improved performance, as well as the capacity to host models from a broad range of representations, allowing us to broaden the scope of BioModels.

The Identifiers.org resolving system provides stable identifiers for resources in the life science domain. In collaboration with the California Digital Library, Identifiers.org now supports resolution of unique, compact identifiers, like `pdb:2gc4`, facilitating efficient data referencing and citation.

We reached hundreds of scientists in outreach and training events in 2017, including courses and workshops locally, internationally and virtually through EMBL-EBI Train online, YouTube and webinars.

Future plans

The Omics Discovery Index (OmicsDI) is expected to grow through the integration of additional data sources, and a broadening of scope through additional data types, as initiated with the addition of systems biology models from BioModels.



Cilium icon from the Reactome icon library



Henning Hermjakob
Head of Molecular Systems.
Team Leader – Molecular Networks

*MSc Bioinformatics University of
Bielefeld, Germany, 1995.*

*Research Assistant at the German
National Centre for Biotechnology
(GBF), 1996.*

At EMBL-EBI since 1997.

The Complex Portal will reach a major milestone with the completion of the yeast complexome annotation project.

After completion of the redevelopment of the Reactome web-based user interface, we will improve the Reactome search and pathway analysis capabilities, and enhance the integration with external molecular interaction and expression data resources.

Based on the new BioModels web platform, we will improve the support for representation standards in addition to the Systems Biology Markup Language (SBML).

In the context of the Identifiers.org and OmicsDI projects, we will continue to pilot agile strategies for data discovery and integration, including cloud-based deployment.

Selected publications

Perez-Riverol Y et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol*. doi: 10.1038/nbt.3790.

Sidiropoulos K et al. (2017). Reactome enhanced pathway visualization. *Bioinformatics*. doi: 10.1093/bioinformatics/btx441

Lloret-Villas A et al. (2017). The Impact of Mathematical Modelling in Understanding the Mechanisms Underlying Neurodegeneration: Evolving Dimensions and Future Directions. *CPT Pharmacometrics Syst Pharmacol*. doi: 10.1002/psp4.12155

Wimalaratne S et al. (2017). Uniform Resolution of Compact Identifiers for Biomedical Data. *Biorxiv (Preprint)*. doi: 10.1101/101279

Combe CW et al. (2017). ComplexViewer: visualization of curated macromolecular complexes. *Bioinformatics*. doi: 10.1093/bioinformatics/btx497

Molecular Systems

Data resources

Reactome

Reactome is an open-source, open-access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff, and cross-referenced to many bioinformatics databases.

www.reactome.org

BioModels

BioModels Database is a repository of peer-reviewed, computational models, primarily from the field of systems biology, but also of general biological interest. BioModels allows biologists to store, search, and retrieve published mathematical models.

www.ebi.ac.uk/biomodels

Complex Portal

The Complex Portal is a manually curated, encyclopaedic resource of macromolecular complexes from a number of key model organisms.

www.ebi.ac.uk/intact/complex

IntAct

IntAct provides a freely available, open-source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions.

www.ebi.ac.uk/intact

OmicsDI

OmicsDI provides a Knowledge Discovery framework across heterogeneous data (genomics, proteomics, transcriptomics and metabolomics).

www.omicsdi.org

Identifiers.org

Identifiers.org is an established resolving system that enables the referencing of data for the scientific community, with a focus on the life sciences domain.

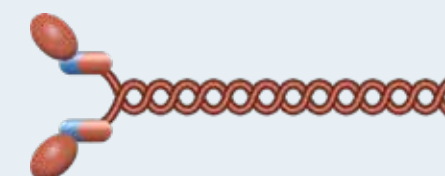
www.identifiers.org



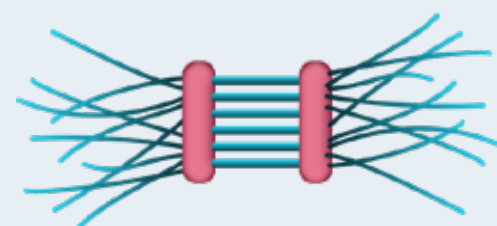
Actin filament



Keratin tetramer



Non-Muscle Myosin II



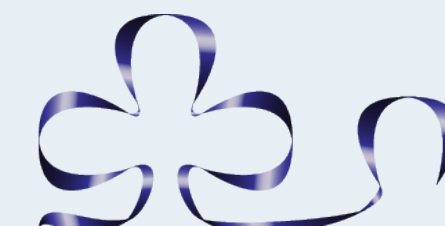
Adherens Junction



VLDL Particle



Mitochondrion



snRNA

Team achievements



Henning Hermjakob

Molecular Networks team

- © The Omics Discovery Index (OmicsDI) now comprises more than 100 000 omics datasets from 15 databases in four continents. It allows users to “claim” their datasets, and to associate them with their ORCID ID
- © IntAct released a new dataset of 15 000 instances where mutations have been experimentally shown to affect protein-protein interactions
- © Reactome completed the redevelopment of its user interface, providing a new website, high quality pathway overview diagrams for efficient navigation through the Reactome pathway hierarchy, and a library of more than 800 icons for biomolecular illustration
- © BioModels released its new website and has more than quadrupled its content of literature-derived models through the addition of more than 6500 Patient Derived Genome Scale Metabolic Models (PDGSMs)
- © In collaboration with the California Digital Library, Identifiers.org now supports resolution of unique, compact identifiers, facilitating data citation

Snapshot of the Reactome icon library, showing re-usable elements for biomolecular visualisation, here from the section “cell elements”.

Chemistry Services

EMBL-EBI's chemistry resources help researchers design and study small molecules and their effects on biological systems. On their own and via integration with other resources, they enable scientists in industry and academia to explore life-science data in new ways.

In January 2017, we welcomed Claire O'Donovan as the Head of Metabolomics. Previously, she was the Protein Function Content Team Leader between 2009 and 2016, providing essential resources to the biological community through the biocuration of UniProt, the Gene Ontology Annotation project (GOA) and the Enzyme portal.

Major achievements

ChEMBL grew to more than 1.7 million compounds and 14.7 million bioactivities, serving approximately 3.5 million page requests from about 16 000 unique hosts per month. At the end of 2017 the number of novel chemical entities annotated in the SureChEMBL patent resource stood at approximately 19 million. In addition to access via a web interface, users can download a regular stream of patent-derived data for integration with in-house resources.

The ChEMBL team is involved in a number of collaborations, resulting in very fruitful outcomes in 2017. For example, as part of our work with Open Targets, we developed a workflow to provide up-to-date information on marketed drugs and compounds in clinical development, thereby linking target with disease.

The ChEBI database also saw significant growth, reaching over 53 500 fully-curated chemical entity entries and serving around 80 000 unique URLs per month. ChEBI is well integrated with MetaboLights, BioModels, Reactome and the Rhea enzymic reaction database and is a critical component of several internal and external resources, including the BRENDA enzyme database and the Gene Ontology.

The metabolomics field continues to evolve, as the data being submitted increases and the user community continues to evolve in its diversity. Therefore, besides delivery of the core MetaboLights database, we have enhanced the online creation and editing of study metadata, developed a new user interface for the submission process and expanded data annotation and analysis. MetaboLights now contains more than 500 datasets and serves as a reference database for metabolomics studies and individual metabolites and associated information such as spectra and chemical structure together with host species/organism information.

EMBL-EBI also coordinated the launch of the PhenoMeNal project in 2017, an international endeavor that aims to use data generated by metabolomics applications to improve our understanding of the causes and mechanisms underlying health, ageing and disease.

In 2017, the MetabolomeXchange consortium also grew to include the Japanese database "Metabolonote", and its platform had more than 1100 datasets publicly available at the end of 2017.

Future plans

In 2018 we aim to continue to broaden the utility and content of ChEMBL, SureChEMBL and related resources. We will further explore the incorporation of bioactivity data from patents into the ChEMBL database and extend our methods to identify and curate relevant information on marketed drugs and compounds in clinical trials

The software infrastructure of ChEBI will be updated to provide a more sustainable and supportable implementation. We will design and deliver this new infrastructure whilst continuing to maintain the existing capability to the user community.

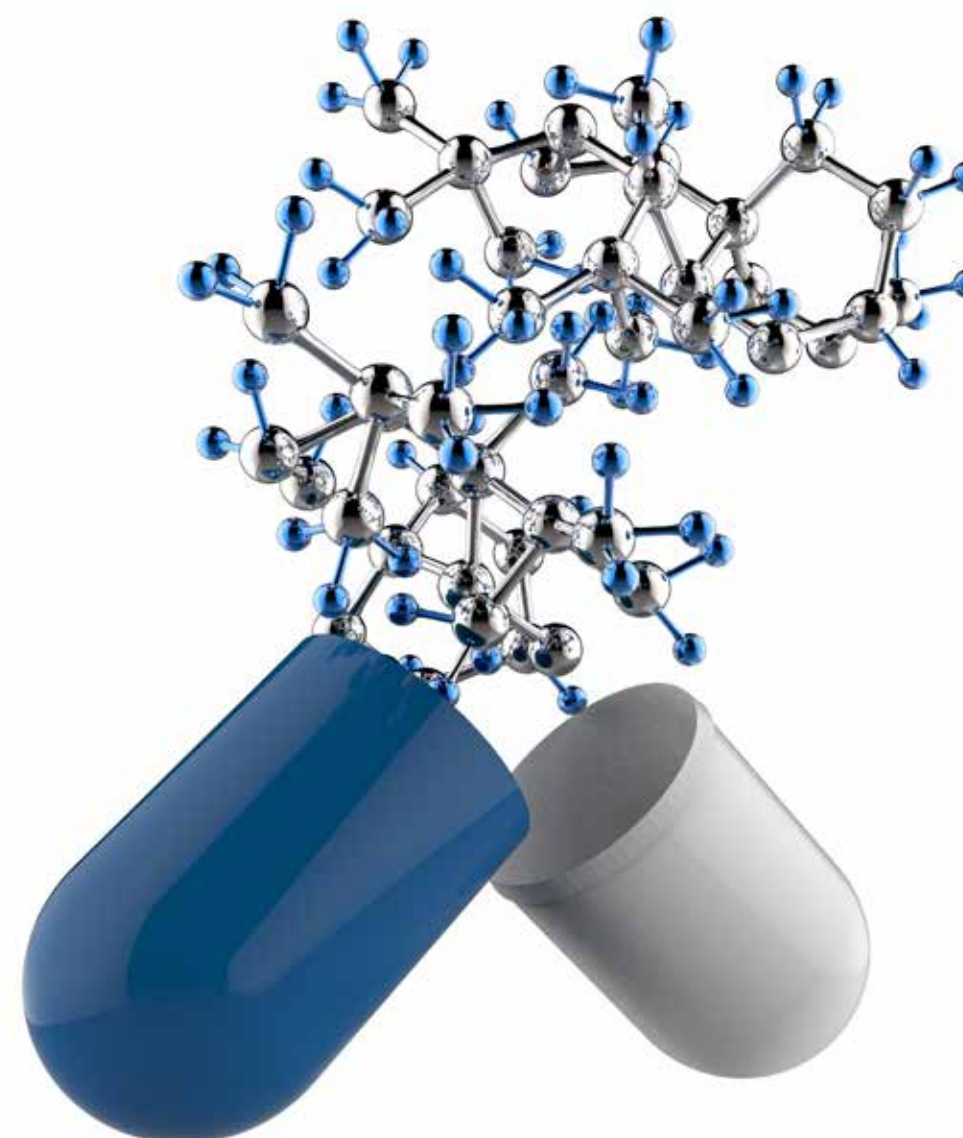
For MetaboLights, we intend to release the new website and the new submission process in 2018. On the data side, we will extend our activities to enable us to reflect the changing nature of metabolomics and new communities such as genomics and clinical specialists.



Andrew Leach

Head of Chemistry Services.
Team Leader – Chemical
Biology

GSK Research and Development, 1994-2016. Trustee of the Cambridge Crystallographic Data Centre, 2006-2015. Editor, Journal of Computer-Aided Molecular Design, 1997-2012. DPhil in Chemistry, Oxford University. At EMBL-EBI since August 2016.



Chemistry Services

Data resources

ChEMBL

ChEMBL, a database of bioactive compounds, provides curated quantitative bioactivity data that links compounds to molecular targets, phenotypic effects, exposure and toxicity endpoints. ChEMBL focuses on interactions relevant to pharmaceutical and agro-chemical development. Most of the data is curated from the literature, supplemented by a growing number of deposited data sets.

www.ebi.ac.uk/chembl

SureChEMBL

SureChEMBL extracts chemical structure data from the full text and images of patents on a daily basis using an automated pipeline. It provides a valuable source for scientific research, as the data in patents is complementary to that in the scientific literature.

www.surechembl.org

UniChem

UniChem is an IUPAC International Chemical Identifier (InChI)-based resolver that enables rapid lookup of chemical structure objects across both EMBL-EBI and external resources.

www.ebi.ac.uk/unichem

ChEBI

Chemical Entities of Biological Interest (ChEBI) is a freely available, manually annotated dictionary and ontology of molecular entities focused on small chemical compounds. It provides a wide range of related chemical information including chemical structures, such as formulae, names and synonyms, links to other databases and a controlled vocabulary that describes the chemical space.

www.ebi.ac.uk/chebi

MetaboLights

MetaboLights is a database for metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra, as well as their biological roles, locations and concentrations, plus experimental data from metabolomics experiments.

www.ebi.ac.uk/metabolights

PhenoMeNal

PhenoMeNal (Phenome and Metabolome aNalysis) is a comprehensive and standardised e-infrastructure that supports the data processing and analysis pipelines for molecular phenotype data generated by metabolomics applications. EMBL-EBI coordinates the project, funded by the EU's Horizon 2020.

<http://phenomenal-h2020.eu/home/>

Team achievements



Andrew Leach

Chemical Biology team

- ChEMBL database reached 14.7 million bioactivity values on over 1.7 million distinct compounds
- SureChEMBL grew to over 19.6 million unique compounds extracted from 18 million patents
- The development and detailed evaluation of a much more powerful ChEMBL database schema
- Multiple collaborative projects in the safety/toxicology area
- Pilot study to identify the value of extracting bioactivity data from patents
- Workflow able to provide information on marketed drugs and compounds in clinical development
- Over 154 million unique chemical structures referenced in UniChem
- Development of a new ChEMBL user interface
- Automation and rationalisation of the complex processes used to create each ChEMBL release
- Continued growth in the coverage of the ChEBI database with 3000 fully curated entries being added during the year, bringing the total to over 53 000.
- ChEBI used UniChem to provide links to several resources, including ChEMBL, SureChEMBL, PubChem, CID, Brenda Lingands and Bksm-React.



Claire O'Donovan

Metabolomics team

- Started a major technology refresh for MetaboLights to simplify data submission, general use and responsiveness
- Review of the MetaboLights website to highlight the full functionality of the data and services including the new analysis platform LABS
- Deployment of integrated, secure, permanent, on-demand, sustainable e-infrastructure for the processing, analysis, and information-mining of PhenoMeNal data
- EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists and the EMBL Metabolomics Workflows course
- Collaboration with EU-funded METASPACE project resulting in an automated data transfer of metabolomics-related image data between METASPACE and MetaboLights

Chemical Biology

The Chemical Biology team develops and manages ChEMBL, a widely-used resource for the drug discovery, agrochemical and consumer healthcare communities. It also manages SureChEMBL, the patent resource containing chemical structures extracted from patents on a daily basis; and UniChem, a resource to link chemical structures across databases, both internal and external to EMBL-EBI.

Our research interests centre on the use of informatics and modelling techniques to tackle problems relevant to translational drug discovery, including aspects of molecular recognition, drug safety, target selection and protein structure.

Major achievements

ChEMBL

Currently there are over 1.7 million distinct compound structures in ChEMBL with 14.7 million activity values. ChEMBL is not limited to particular data types and includes pharmacological data ranging from straightforward protein-ligand binding data to more complex cell-based, tissue and *in vivo* functional assays and disease models.

One continued area of focus in 2017 has been the therapeutic target and indication annotation of marketed drugs, withdrawn drugs and compounds in clinical development. This information enables the establishment of high-confidence target-compound-disease links, valued by collaborators including Open Targets and the Illuminating the Druggable Genome project (IDG). Drug discovery data is increasingly complex and we have undertaken significant work to extend the ChEMBL database schema to capture information from, for example, phenotypic screens, *in vivo* toxicity assays or pharmacokinetic endpoints in a more structured format.

We worked on a number of key technology projects during 2017. A project to redesign and implement a more efficient, streamlined and automated process has made significant progress. A second major project is a completely new web interface with a modern “look and feel” that will also provide more accessible ways for users to visualise and analyse the results of searches. This project was guided by a “user experience” approach to define common workflows.

SureChEMBL

SureChEMBL now contains more than 19.6 million unique compounds that have been extracted from 18 million patents. These datasets grow at a rate of around 80 000 novel chemicals per month from roughly 50 000 new patents.

We have been working to make SureChEMBL more robust and to limit some of the dependencies. For example, the system is now deployed on EMBL-EBI's Embassy cloud compute infrastructure rather than a commercial provider. In a pilot study as part of the Illuminating the Druggable Genome project, we explored the scope of patent data as a source of bioactivity data and information on “underexplored” targets. A semi-automated workflow was developed to identify patents of potential interest, for subsequent manual inspection. Appropriate patents were then fed into our established ChEMBL curation and annotation pipeline for inclusion in the database.

We continued to provide weekly updates of the indexed compounds in UniChem and quarterly downloads of the mapping between compounds and patents.

Currently, UniChem contains links to over 153 million chemical structures drawn from 34 separate resources.

ChEBI

ChEBI continued its steady growth during 2017 with over 3000 fully curated entries being added during the year, bringing the total to over 53 000. In 2017, we made use of the capabilities of UniChem to provide links from ChEBI to several new resources, including ChEMBL, SureChEMBL, PubChem CID, Brenda Ligands and Bkms-React.

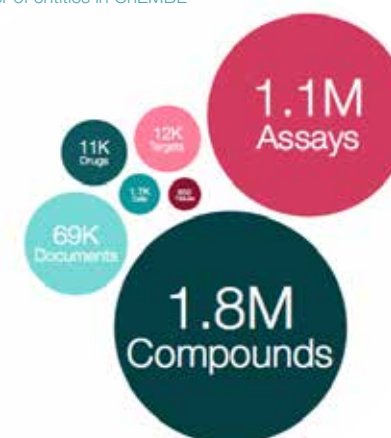
We provided new links from ChEBI to the DrugCentral database. Accessions from ChEBI to the GlyTouCan, FooDB, and FAO/WHO Food Standards databases are now available to curators, and we will aim to include links to these resources during the coming year.

Collaborations

The ChEMBL team continues to be involved in a number of external collaborations. Three new projects started in 2017 (IMI-TransQST, IMI-eTRANSAFE and Open Targets target tractability), the first two of which are in the area of safety and toxicology. We contribute in a number of ways to these and related projects, including the curation of relevant data, the development of new ways to visualise complex data types, the creation of predictive models and the delivery of underlying infrastructure components.

We have also continued to contribute to the Open Targets collaboration and to the NIH-funded Illuminating the Druggable Genome project, with a focus on clinical pipeline and patent-derived data respectively. With Open Targets, we are exploring methods for target tractability assessment and delivery of a practical tractability decision-making workflow for use in drug discovery. As one of the CORBEL research infrastructures, we have engaged with a number of scientists wishing to pursue their research projects using ChEMBL data.

Number of entities in ChEMBL



Andrew Leach

Head of Chemistry Services.
Team Leader – Chemical Biology

GSK Research and Development, 1994-2016. Trustee of the Cambridge Crystallographic Data Centre, 2006-2015. Editor, Journal of Computer-Aided Molecular Design, 1997-2012. DPhil in Chemistry, Oxford University. At EMBL-EBI since August 2016.

Future plans

In 2018 we aim to continue to broaden the utility and content of ChEMBL, SureChEMBL and related resources. We will further explore the incorporation of bioactivity data from patents into the ChEMBL database.

We will also expand our use of relevant ontologies as these gain acceptance by the wider user community. We aim to complete the development of our new web interface and release it to the community. We will deploy the new ChEMBL database schema and complete the improvements to the release process.

The software infrastructure of ChEBI will be updated to provide a more sustainable and supportable implementation. We will design and deliver this new infrastructure whilst continuing to maintain the existing capability to the user community.

Last, but not least, we will continue our active involvement in external collaborations, seeking new ways to capitalise on safety and toxicity data and to further exploit our resources in active research projects.

Selected publications

Gaulton A et al. (2017). The ChEMBL Database in 2017. *Nucleic Acids Research*. doi: 10.1093/nar/gkw1074

Nowotka M et al. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opinion on Drug Discovery*. doi: 10.1080/17460441.2017.1339032

Perez-Riverol Y et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol*. doi: 10.1038/nbt.3790

Vita R et al. (2017). Better living through ontologies at the Immune Epitope Database. *Database*. doi: 10.1093/database/bax014

Hastings J et al. (2017). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. doi: 10.1093/nar/gkv1031



Protein families covered in ChEMBL according to the number of compound activity measurements for each.

Metabolomics

The Metabolomics team established and manages the MetaboLights database, which covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.

Major achievements

MetaboLights

In 2017, we introduced a new online submission model, which greatly reduces dependencies on external applications. This enables us to ensure sustainability and make the submission process easier and more intuitive.

Our secure web services were extended to support more focused searching and more granular editing of metadata. The new quick-edit, updated web services and the new design is currently in internal testing and will be released to the public in the near future. We also performed a thorough review of our website and redesigned it with a fresh new look that highlights even more relevant metadata to our users.

We finalised the first automatic study submission pipeline for Mass Spectrometry imaging studies in collaboration with the EMBL Heidelberg core metabolomics facility in an EU Horizon 2020 grant-funded activity. This is an exciting development in the metabolomics field, providing new insights into metabolomic processes at the cellular and tissue levels. We expect there to be great growth in this area of metabolomics. We extended the MetaboLights Labs workspace to facilitate this integration, in addition to include features to automatically upload data from external systems, like Galaxy workflow environments.

We now link more than 25 000 compounds in MetaboLights to ChEBI, enriching compounds with submitted spectra from Mass Spectrometry (MS) and Nuclear Magnetic Resonance Spectroscopy (NMR) experiments. In December 2017, MetaboLights contained just over 520 studies covering about 340 unique organisms, 117 000 sample records, 431 000 raw files, 805 000 metabolite features (spectra) reported. Over 45 000 of the metabolite features are now linked to ChEBI identifiers and it will be a focus for the coming year to enhance our identification processes in collaboration with the ELIXIR Metabolomics community.

Community standards

The MetabolomeXchange consortium continues to grow with the addition of the Japanese database “Metablonote” and the MetabolomeXchange platform had more than 1100 datasets publicly available at the end of 2017. We are also part of the Proteomics (HUPO) consortium developing mzTab standards for capturing and reporting metabolomics identified metabolites. We are also involved in the development of new qcML data format for quality control jointly led by the Proteomics Standards Initiative (PSI) and Metabolomics Standards Initiative (MSI) communities.

PhenoMeNal

A large volume of medical molecular phenotyping and genotyping data will be generated by metabolomics applications now entering research and the clinic. The PhenoMeNal project aims to develop and deploy an integrated, secure, sustainable e-infrastructure for the processing, analysis, and information-mining of such data.

In 2017, PhenoMeNal established an e-infrastructure for data processing of medical metabolic phenotype data. In this period, the consortium work was focused on the development of the Virtual Research Environment (VRE) portal with automatic user-controlled deployment. Deployments of the VRE are fully supported on EMBL-EBI Embassy Cloud (OpenStack), Amazon (AWS) and Google (GCP).

Outreach and training

In 2017 MetaboLights and PhenoMeNal were represented through talks, posters and training sessions at 23 scientific conferences, training sessions, larger meetings and knowledge-exchange events in the UK and other European countries, as well as China, Japan and Australia. We successfully hosted the oversubscribed 2017 EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists. Following feedback from this course, we also successfully developed and delivered a complementary Metabolomics Workflows course for advanced users in October 2017.

Future plans

We intend to release the new website and the new submission process in 2018. On the data side, we will extend our activities to enable us to reflect the changing nature of metabolomics and the new communities such as genomics and clinical specialists.

We will further enrich MetaboLights with curated knowledge, including reference spectra, pathways, protocols and references to a wider range of resources. We will continue to develop new online data analysis capabilities in LABS to strengthen the position of MetaboLights as an important research platform.

PhenoMeNal will provide the third release in February 2018, and it is planned to add about 30 more tools to enhance the analysis processes. August 2018 will see the fourth stable release, including full Microsoft (Azure) support, in addition to workflow enhancements and an extended service catalogue.



Claire O'Donovan

Team Leader – Metabolomics

BSc (Hons) in Biochemistry, University College Cork, 1992. Diploma in Computer Science University College Cork, 1993. UniProt Content Team Leader 2009-2016; Head of Metabolomics since January 2017. At EMBL since 1993. At EMBL-EBI since 1994.

Selected publications

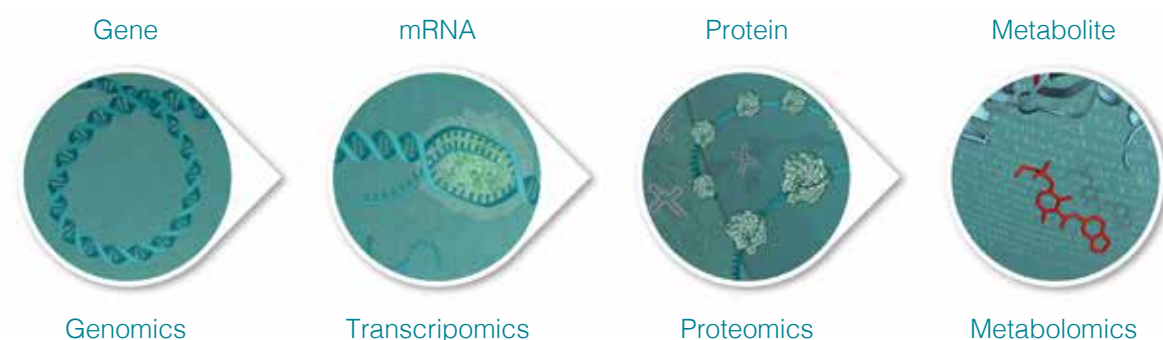
Larralde M et al. (2017). mzML2ISA & nmrML2ISA: generating enriched ISA-Tab metadata files from metabolomics XML data. *Bioinformatics*. doi: doi.org/10.1093/bioinformatics/btx169

Van Rijswijk M et al. (2017). The future of metabolomics in ELIXIR. *F1000Research*. doi: 10.12688/f1000research.12342.1

Haug K et al. (2017). Global open data management in metabolomics. *Current opinion in chemical biology*. doi: 10.1016/j.cbpa.2016.12.024

Salek RM et al. (2017). Automated assembly of species metabolomes through data submission into a public repository. *GigaScience*. doi: 10.1093/gigascience/gix062

Schober D et al. (2017). nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. *Analytical Chemistry*. doi: 10.1021/acs.analchem.7b02795



Where Metabolomics fits in the world of 'omics

Literature Services

The EMBL-EBI Literature Services team runs Europe PMC, the database for life-science literature and platform for text-based innovation. Linking research articles with the underlying supporting data and using articles to provide biological context across all data resources are two critical factors that enable data-driven discovery.

Europe PMC 2017 highlights



Search snippets



Suggested authors



Data behind the article



Annotations API

To achieve these goals, we work collaboratively with other service providers, publishers and databases, and engage with the scientific community and our users. In this regard, the databases developed at EMBL-EBI, and more widely through ELIXIR, are of primary importance.

The core of our work is providing fast, reliable and powerful access to scientific literature. We also place article narratives in the wider context of related data and credit systems, such as article citations. We engage with individual scientists, text miners, developers and database managers to understand how we can build layers of value upon the basic article content. Our team provides the infrastructure that enables individuals to enrich scientific literature, either manually or using computational methods, and to publish the results, maximising the usefulness of the core content. This allows the widest possible reuse of publicly-funded experimental data.

Europe PMC grows at a rate of over 1 million abstracts and around 300 000 full-text articles per year. As of December 2017 Europe PMC offered more than 33 million abstracts and around 4.6 million full-text research articles. It also includes agricultural science records from Agricola, biological patents and clinical guidelines.

Europe PMC adds value through the development of article-citation networks, links articles to underlying datasets and makes text-mined terms of biological interest discoverable. It provides programmatic access via REST and SOAP web services, and allows users to bulk-download open-access articles (over 1.8 million) via FTP. Users can also search over 60 000 biomedical research grants that have been awarded to nearly 29 000 PIs supported by Europe PMC's 28 funders. These funders include the World Health Organization, the European Research Council and many national funding agencies and charities, led by the Wellcome Trust.

Major achievements

In 2017 use of Europe PMC continued to increase with more than 17 million unique IP addresses visiting the website during the year. Programmatic access via RESTful web services has served up to 1.8 TB per month in XML and JSON formats.

To facilitate information discovery, search results in Europe PMC now include snippets. Snippets reveal the most relevant article excerpts containing the searched terms and can be very helpful to identify relevant articles, and get context for the search. Users can also locate the search terms in the article by following the snippet link under each excerpt. The link points to the sentence in the publication where the snippet was retrieved.

Our integration with ORCID iDs – unique identifiers for researchers – continued to be central to developments of Europe PMC. An author search in Europe PMC now brings up a Suggested Authors box linking to matching researchers that have an ORCID. It displays up to the two most prolific researchers and links to their Author Profile page. With over 650 000 authors in Europe PMC actively publishing nearly 5 million scientific articles using their ORCID, we expect this feature to be of increasing interest to our users.

Data integration is a unique feature of Europe PMC. To provide easy access to all primary data associated with a study, Europe PMC has integrated with the BioStudies database. A BioStudies record is generated for every full text article in Europe PMC that either has supplemental data files or mentions data identified by text-mining accession numbers for over 20 major data resources, including ENA, PDB, and UniProt. Over a million Europe PMC articles now have corresponding BioStudies records.

One of the goals of open access is to stimulate innovation: to discover and use the content in new ways. We have recently expanded the Europe PMC programmatic tools suit with the Annotations API, which provides access to targeted information text-mined from millions of biomedical abstracts and open-access, full-text articles. It allows users to retrieve, for example, all articles that discuss involvement of a specific gene or protein in their disease of interest. There are over 488 million annotations on 19 million articles in Europe PMC – more than the number of known variations in the human genome.

Future plans

Using a variety of user-based research approaches, we will continue to improve user experience on the website, and, in particular, the article pages and search experience over the course of 2018.

Data resource

Europe PMC

Europe PMC provides open access to full-text scientific literature resources and supports innovation by engaging users, enabling contributors, and integrating related research data. Europe PMC hosts a full mirror of PubMed, but also includes several million more abstracts of global patents, Agricola (agricultural research) and other sources.

www.europepmc.org



Johanna McEntyre

Team Leader – Literature Services

PhD in Plant Biology, Manchester Metropolitan University, 1990. Editor, Trends in Biochemical Sciences, Elsevier, 1997. Staff Scientist, NCBI, National Library of Medicine, NIH, US, 2009. At EMBL-EBI since 2009.

Selected publications

Levchenko M, et al. (2018). Europe PMC in 2017. *Nucleic Acids Res.* doi: 10.1093/nar/gkx1005

Sarkans U, et al. (2018). The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* doi: 10.1093/nar/gkx965

Kafkas S, et al. (2017). Literature evidence in open targets - a target validation platform. *J Biomed Semantics.* doi: 10.1186/s13326-017-0131-3

McMurry JA, et al. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol.* doi: 10.1371/journal.pbio.2001414

Anderson W, et al. (2017). Towards coordinated international support of core data resources for the life sciences. *bioRxiv.* doi:10.1101/110825

Team achievements

Johanna McEntyre

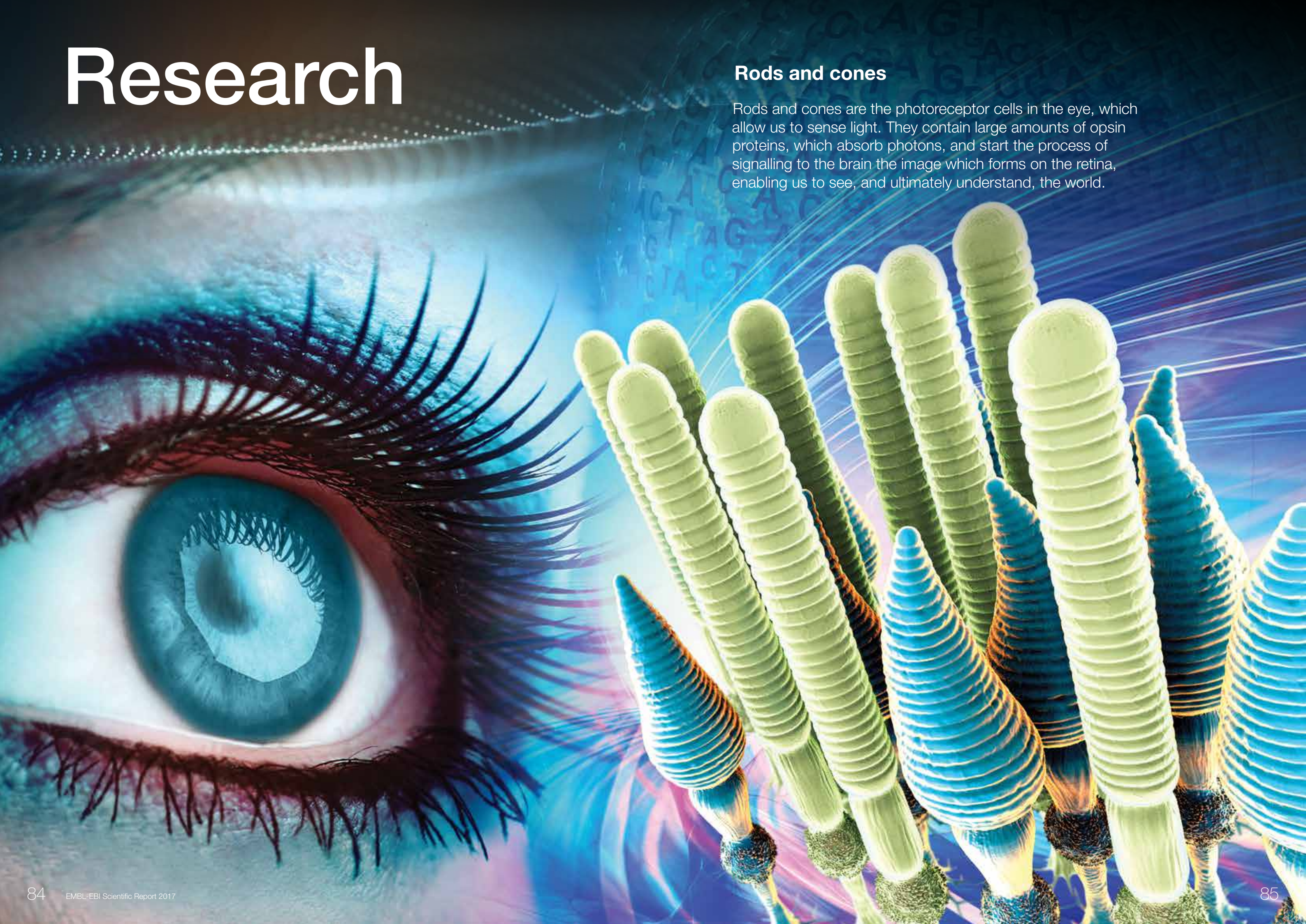
Literature Services team

- ② Launched search snippets to help identify relevant articles and get search context
- ② Developed "Suggested Authors" feature that suggests researchers with an ORCID matching an author search
- ② Integrated with BioStudies database to make supplemental data citable and make it more available for re-use and discovery
- ② Developed Annotations API

Research

Rods and cones

Rods and cones are the photoreceptor cells in the eye, which allow us to sense light. They contain large amounts of opsin proteins, which absorb photons, and start the process of signalling to the brain the image which forms on the retina, enabling us to see, and ultimately understand, the world.



Research achievements in 2017

In 2017 we welcomed new group leaders: Evangelia Petsalaki and Zamin Iqbal. Petsalaki's work focuses on human cell signalling in healthy and disease conditions, while Iqbal's computational genomics research group explores genetic variation in microbes, and develops methods for investigating surveillance and diagnostics of antimicrobial resistance.

The Marioni group showed that in mice immune cells in older tissues lack coordination.



Cancer genetics

An international research consortium led by Moritz Gerstung has shown proof of concept that personalised therapy will be possible in the future for people with cancer (Gerstung M et al, 2017). The study provides details of how a knowledge bank could be used to find the best treatment option for people with acute myeloid leukaemia (AML). The researchers also explored how a patient's genetic details can be incorporated into predicting the outcome and treatment choice for that patient.

In 2017, for the first time, scientists completed a detailed study of many of the proteins in bowel cancer cells. Research led by the Wellcome Sanger Institute and the Saez-Rodriguez group investigated the role proteins play in predicting how common mutations affect proteins in cancer cells, and whether such proteins are important in predicting the cancer's response to treatment (Roumeliotis T et al, 2017). The results gave scientists a better picture of the cellular processes behind bowel cancer, and could enable researchers to predict which drugs would be effective in treating individual patients.

Ageing insights

Working alongside the University of Cambridge, the Wellcome Sanger Institute and the Cancer Research UK–Cambridge Institute (CRUK-CI), researchers in the Marioni group have shed light on a long-standing debate about why the immune system weakens with age (Martinez-Jimenez CP et al, 2017). Their findings show that immune cells in older tissues lack coordination and exhibit much more variability in gene expression compared with their younger counterparts.

Social genetic effects

Researchers in the Stegle group have shown that the health of individual mice is influenced by the genetic makeup of their partners. Unexpectedly, the genetics of social partners were found to affect wound healing and body weight as well as behaviour (Baud A et al, 2017). The methods used to detect 'social genetic effects' help future research into the mechanisms whereby one individual influences another.

Understanding disease

The Stegle group, in collaboration with the Wellcome Sanger Institute, King's College London, the University of Dundee and the University of Cambridge, launched one of the largest public collections of high-quality human induced pluripotent stem cell lines (iPSCs) from healthy individuals. Scientists can use these cells to study human development and disease. The new resource,

comprising of the systematic generation, genotyping and phenotyping of hundreds of stem cell lines by the Human Induced Pluripotent Stem Cells Initiative, also outlined the major sources of genetic and phenotypic variation in iPSCs and established their suitability as models of complex human traits and cancer (Kilpinen H et al, 2017).

Using knock-out mice, researchers from the Medical Research Council (MRC) Harwell Institute and the Mouse Informatics group at EMBL-EBI found 52 previously unidentified genes that are critical for hearing (Bowl MR et al, 2017). These newly discovered genes could provide novel insights into the causes of hearing loss in humans.

The Iqbal group and collaborators at the University of Oxford developed a method for extracting *Mycobacterium tuberculosis* DNA directly from sputum (with no culture) and then sequencing it directly. This reduced the sequencing time from around two weeks, most of which was spent waiting for bacteria to grow in culture, to less than 24 hours, demonstrating the potential for a portable, handheld point of care test for tuberculosis (Votintseva AA et al, 2017).

A collaboration between the Beltrao group and the Wellcome Sanger Institute discovered new insights into the life cycle of malaria-causing *Plasmodium* parasites, as they are transmitted from mammal to mosquito (Invergo B, et al, 2017). Using mass spectrometry and computational biology, the collaborators explored phosphorylation signaling during the first minute of *Plasmodium berghei* gametocyte activation in the vector. This revealed an unexpectedly broad response, with proteins related to distinct cell cycle events undergoing simultaneous phosphoregulation. Determining the underlying cell signaling events and how the parasites detect the change in environment from host to vector is a primary step toward a fundamental understanding of transmission of malaria.

Deep learning

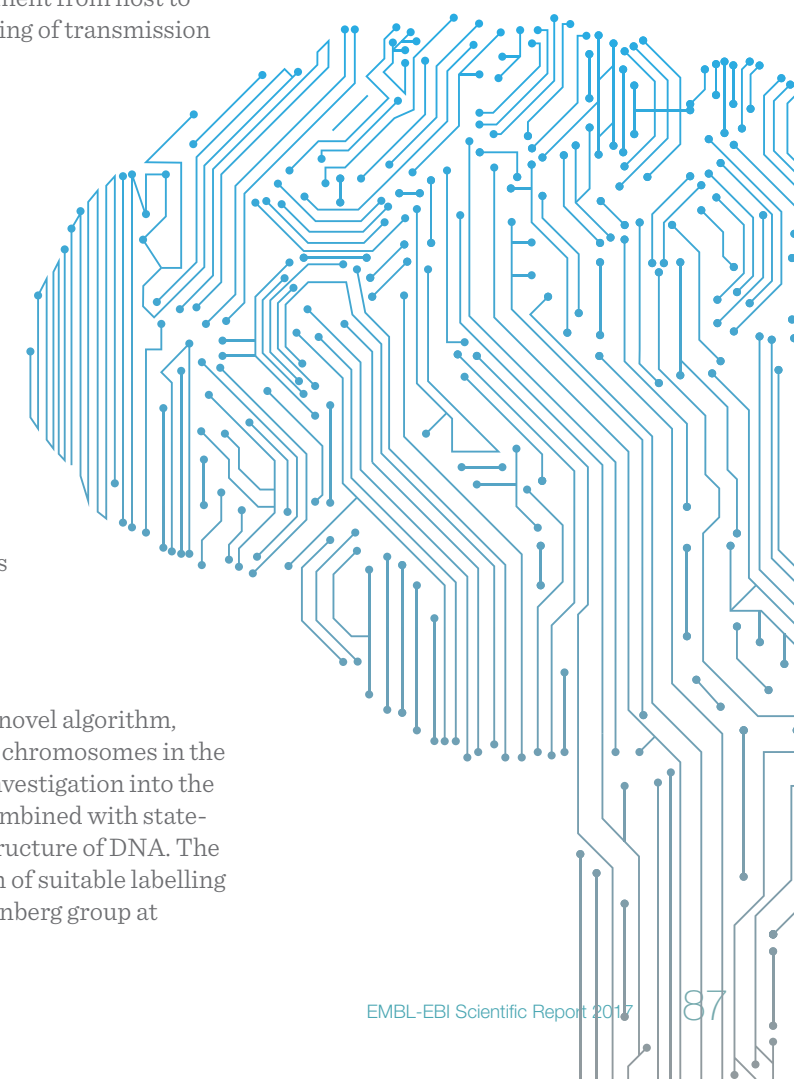
To better understand how DNA sequences relate to biological changes, the genomics community is turning to artificial neural networks – a class of machine learning methods first introduced in the 1980s and inspired by the wiring of the brain. More recently, these models have been rebranded as “deep neural networks”, which form the field of deep learning.

The Stegle group developed a new deep-learning method, DeepCpG, which helps scientists better understand the epigenome – the biochemical activity around the genome (Angermueller C et al, 2017). This method, which predicts missing methylation states, enables genome-wide analyses of DNA methylation assayed at single-cell resolution.

Super-resolution imaging

The Birney group successfully completed and published a novel algorithm, ChromoTrace, for determining the 3D structure of human chromosomes in the nucleus (Barton C et al, 2018). This work is a theoretical investigation into the feasibility of using advanced computational techniques combined with state-of-the-art laboratory labelling protocols to study the 3D structure of DNA. The information provided by this research is guiding the design of suitable labelling strategies for real super resolution experiments in the Ellenberg group at EMBL Heidelberg.

The Beltrao group explored the life cycle of *Plasmodium* parasites as they are transmitted from mammal to mosquito.



EMBL International PhD Programme at EMBL-EBI

Students in the EMBL International PhD Programme at EMBL-EBI receive advanced, interdisciplinary training in molecular biology and bioinformatics, which results in a joint degree from EMBL and 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 2017, eight of our 24 PhD students obtained their degree, with the successful theses focusing on topics such as genetic analysis of molecular traits in skeletal muscle (Leyland Taylor), deep neural networks and statistical models for studying single-cell DNA methylation (Christof Angermuller) and the evolution, modifications and interactions of proteins and RNAs (Ananth Prakash Surappa-Narayanappa).

EMBL Postdoctoral Programme at EMBL-EBI

Our international, interdisciplinary environment and world-class facilities set our Postdoctoral Fellowships apart. Our postdocs work closely with experts in our data service and infrastructure teams, and benefit from the rich scientific exchange and collaborative culture of the Wellcome Genome Campus.

In addition to general EMBL postdoctoral students, we have three postdoctoral schemes. The first one is ESPOD, the EMBL-EBI–Sanger Postdoctoral Programme, a perfect combination of wet lab and dry lab science for researchers who take both an experimental and computational approach to their work. The second scheme is EBPOD, the EMBL-EBI–Biomedical Research Centre Postdoctoral Programme, aimed at researchers who apply computational approaches to translational clinical research. Last, but not least, EIPOD is the EMBL Interdisciplinary Programme, aimed at candidates whose research crosses multiple scientific boundaries.

In 2017, EMBL-EBI had 40 postdoctoral researchers, with eight researchers on the EIPOD scheme, three researchers on our ESPOD scheme and three on the EBPOD scheme.



EMBL-EBI Predoctoral Fellows in 2017.
Top row, left to right: Matthew Jeffries, Marta Strumillo, Melike Donertas, Alistair Dunham, Jose Guilherme de Almeida, Borgthor Petursson, Conor Walker, Hannah Currant, Sergio Miguel Santos, David Bradley. Bottom row, left to right: Claudia Hernandez, Umberto Perron, Ricard Argelaguet, Anna Cuomo, Harald Vohringer, Nadezda Volkova, Elsa Kentepozidou, Aleix Lafita Masip.

Genomics research



Birney group

Algorithms and outbred genetic variation

- ⊙ Developing novel algorithms, such as ChromoTrace, for determining the 3D structure of chromosomes in the nucleus from super-resolution microscopy data
- ⊙ Whole genome sequencing and initial analysis of the first vertebrate near-isogenic panel in medaka fish, the Kiyosu panel
- ⊙ Collaborations with clinical researchers at Addenbrookes Hospital for using electronic healthcare resources (EPIC) to predict important factors related to patient treatment plans
- ⊙ Initial analysis of 65 000 OCT scans from the UK Biobank and testing of large scale genome wide association testing using biobank genetic data



Brazma group

Functional genomics research

- ⊙ Developed methods for single-cell RNA-seq data analysis
- ⊙ Lead a working group on data integration for the Pan-Cancer Analysis of Whole Genomes initiative (PCAWG)
- ⊙ Analysed the data for expression Quantitative Traits (eQTLs) for the PCAWG project



Enright group

Functional genomics and analysis of small RNA function

- ⊙ Utilised differences in gene expression as the readout of potential post-transcriptional regulation in Schistosoma parasites
- ⊙ Developed mirnovo, a machine learning based algorithm able to identify miRNAs directly from small RNA-Seq data
- ⊙ Demonstrated how long terminal repeats contribute to post-mitotic spermatogenic transcriptome diversity



Flicek group

Evolution of transcriptional regulation

- ⊙ In collaboration with the Odom group at the University of Cambridge, described the role of regulatory complexity in the maintenance of gene expression across mammalian evolution
- ⊙ Quantified the role and contributions of cis and trans variation to transcription factor binding intensity and gene expression using an F1 model in mice
- ⊙ Using the genomes of a rat model of metabolic syndrome, demonstrated how an integrative analysis of genome variation and conserved regulation can be used to identify genomic regions responsible for the observed phenotypes

Research summaries



Gerstung group

Computational cancer biology

- ⊙ First comprehensive analysis showing that cancer-causing mutations often arise years, and in some cases decades, before diagnosis in a patient's cells
- ⊙ Development of a predictive model to determine the risk of developing acute myeloid leukaemia 5-10 years in advance
- ⊙ Analysis of genetic and genotoxic causes of mutation spectra seen in human cancers



Goldman group

Evolutionary tools for genomic analysis

- ⊙ Completed and published our description of a newly discovered small-scale genome mutation process, in which short genome regions are replaced by material copied from the other strand of replicating double-stranded DNA
- ⊙ Development of the IRaPPA method for ranking structural models of protein complexes
- ⊙ With support from the BBSRC, continued our work towards re-purposing DNA as a practical medium for archiving digital information
- ⊙ Continued our work on how the choice of genes with different evolutionary rates impacts on the accuracy and reliability of inferences of phylogenetic history
- ⊙ Investigated the comparative accuracy of reconstruction of ancestral genome sequences based on different multiple sequence alignment algorithms
- ⊙ Developed phylogenetic algorithms designed to provide unbiased estimates of evolutionary divergence from data sets in which only SNPs are recorded, and non-variable sites have been omitted



Iqbal group

Computational microbial genomics

- ⊙ Laying the foundations of scalable "search index" technologies for DNA archives (e.g. "has anyone seen this mutation before in this species?")
- ⊙ Led the sequence analysis of 100 000 M. tuberculosis genomes for the CRyPTIC global consortium, which is building a catalog of drug-resistance mutations to enable partial replacement of phenotyping for clinical diagnostics
- ⊙ Developed two types of "graph genome" approaches. The first approach models eukaryotic recombination, and we are primarily focusing on surface antigens of the malaria parasite P. falciparum. The second approach models the bacterial pan-genome and uses new long read (nanopore) technology

Protein, structure and chemical biology research

Research summaries



Bateman group

Analysis of protein and RNA sequence

- ⦿ Defined the complete set of known RNA interactions for the yeast genome and analysed the network properties compared to protein interactions
- ⦿ Identified a novel protein family found in the Eros proteins that plays a role in mouse macrophage reactive oxygen bursts
- ⦿ Identified the evolutionary relationships of domains found in the structure of the RNA chaperone ProQ



Thornton group

Proteins: Structure, function and evolution

- ⦿ Study of enzyme evolution revealed how enzyme families evolve to include members with different functions, often retaining their catalytic mechanism
- ⦿ Analysis of de novo variants that cause developmental disorders in children
- ⦿ Development of computational methods to prioritise drugs for testing in model organisms, based on their likely effects on ageing



Beltrao group

Evolution of cellular networks

- ⦿ Investigated the changes in protein phosphorylation occurring during the first seconds of Plasmodium gametogenesis
- ⦿ Studied how gene copy-number alterations are sometimes attenuated from mRNA levels to the protein abundance levels
- ⦿ Compiled and studied a collection of approximately 900 strains of E. coli in order to attempt to predict their growth phenotypes from their genome sequences



Marioni group

Computational and evolutionary genomics

- ⦿ Developed computational approaches for normalisation, batch correction and downstream interpretation of single-cell RNA-sequencing data
- ⦿ Used scRNA-seq to characterise cell type heterogeneity in an entire mouse embryo
- ⦿ Used scRNA-seq to show that the immune response is more variable in older than in younger mice
- ⦿ Member of the steering committee of the Data Coordination Platform for the Human Cell Atlas and co-chair of the Human Cell Atlas Analysis Working Group



Petsalaki group

Whole-cell signaling

- ⦿ Development of approaches towards the creation of a reference kinase-kinase network to be used as a basis to study context specific signalling networks across tissues
- ⦿ Development of a method that uses random walk with restart to propagate signals in noisy networks.



Stegle group

Statistical genomics and systems genetics

- ⦿ Developed improved statistical methods to identify genotype-environment interactions
- ⦿ Devised integrative analytical strategies to investigate genetic variation in the social environment on biomedical traits
- ⦿ Published the first manuscript on human genomic variation in induced pluripotent stem cells
- ⦿ Established an epigenetic clock model based in DNA methylation in the mouse



Birney group

Algorithms and outbred genetic variation

DNA sequence remains at the heart of molecular biology and bioinformatics. The Birney research group focuses on developing sequence algorithms and using genetic variation to explore elements of basic biology within and between species.

Major achievements

In 2017 our group successfully completed and published ChromoTrace, a novel algorithm for determining the 3D structure of human chromosomes in the nucleus using super resolution microscopy imaging. This work is a theoretical investigation into the feasibility of using advanced computational techniques combined with state-of-the-art laboratory labelling protocols to study the 3D structure of DNA. The results are encouraging and information provided by our simulation framework and algorithm performance is guiding the design of suitable labelling strategies for real super resolution experiments in the Ellenberg group at EMBL Heidelberg.

The other major strand of the group research focuses on using genetic and phenotypic variation from specialised key resources and large-scale data collections to understand basic biology.

For example, we have completed the whole genome sequencing and initial sequence analysis of the Kiyosu medaka fish inbred panel. Amazingly, over 75% of the lines from the panel are over 80% homozygous, providing a truly unique model organism resource. Working with collaborators in Heidelberg, Professor Wittbrodt (Heidelberg University) and Dr Loosli (KIT), we have developed pilot phenotyping assays and analytical techniques for high throughput heart beat screening, behavioural assays, morphometric measurements and CT scanning of individual fish.

The last decade has seen an astonishing increase in the use of genome-wide association studies to understand human diseases. The availability of large-scale datasets, such as the UK Biobank, delivers tremendous opportunities for researchers. Association analysis can be applied to nearly any measurable phenotype in a cellular or organismal system where an accessible, outbred population is available. The Birney group works on UK Biobank data with a primary focus on using high dimensional phenotypes (e.g. heart, eye and brain scans) combined with state-of-the-art statistical genetics techniques to explore aspects of basic biology in humans (LiMMBo, PhenotypeSimulator).

Our two major projects involving UK Biobank data include a collaboration with Moorfields eye hospital, for which we are performing genome wide association testing on over 65 000 individuals who have both genetic data and Optical Coherence Tomography (OCT) scans. Secondly, in collaboration with Dr Declan O'Regan at Imperial College London, we are involved in genome wide association testing using reduced representations of heart MRI imaging data within the biobank.

The Birney group also uses its experience in analysing phenotype information and deep learning to develop methods for using big data in healthcare.

Last but not least, we continue to be involved in the opportunities arising from the application of new sequencing technologies, in particular with the new developments from Oxford Nanopore (ONT). The group is involved in active development of methods for improving base calling quality, sequence alignment and the detection of base modifications from direct RNA sequencing ONT data. Genetic variation influences important biological traits and the accurate prediction of deleterious variants is essential for understanding disease. The group has developed a novel variant scoring method for missense mutations, which incorporates ancestral allele frequencies and protein structure information from UniProt, information that is generally lacking from other approaches.

Future plans

In 2018 the Birney group will continue to work on the development of sequence algorithms as well as the use of genetic and phenotypic variation datasets to explore basic biology.

Algorithm development will continue in a number of areas. ChromoTrace will be applied to data from real super resolution microscopy experiments and will undoubtedly require some algorithmic updates to account for issues not captured during simulation. Additionally, the ChromoTrace method will be improved by allowing deeper and more flexible searching of the combinatorial space when seeding initial search paths.



Ewan Birney

Research Group Leader and Senior Scientist

PhD 2000, Wellcome Sanger Institute. At EMBL since 2000. Joint Associate Director from 2012 to 2015. EMBL-EBI Director since June 2015.

Work on the Kiyosu panel will continue in a number of key areas. A major focus will be investigations into gene to environment interaction terms. Initial studies will use a change in seasonal conditions as the environmental shift and RNA expression levels across a number of organs as the molecular measurements.

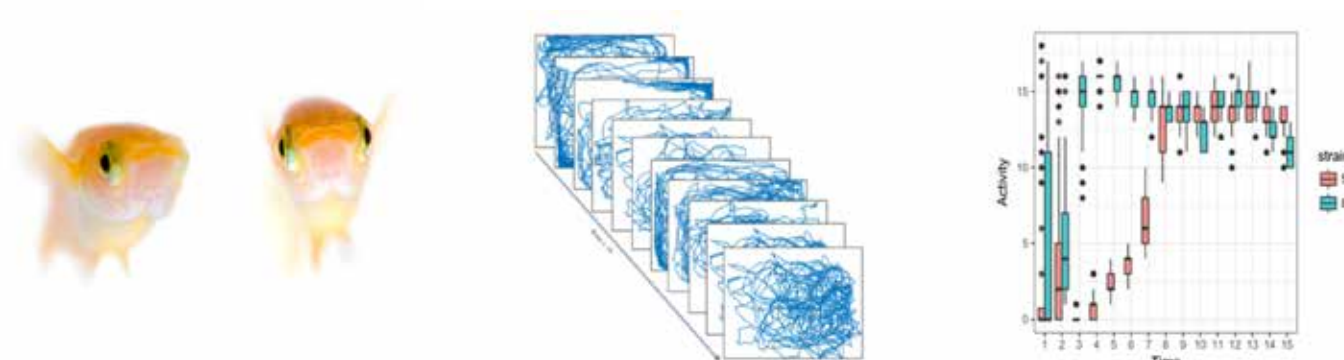
Our work with human genetic data will continue primarily using the UK Biobank resource with a particular focus on genome-wide association testing using reduced representations of both heart MRI and eye OCT imaging data. This work is likely to uncover new genetic loci involved in the development and aging process of the heart and eye.

The group will work on phenotypic data analysis (electronic healthcare records) for individuals greater than 75 years old admitted non-electively to the Department of Medicine for the Elderly at Addenbrookes Hospital, Cambridge, UK. Data from the EPIC database will be extracted to allow high dimensional algorithms and patient trajectory models to be developed using deep learning. This work should discover major driver variables for predictions into certain outcome classes providing meaningful insight into when patients are likely to display particular issues.

Selected publications

Barton, et al. (2018). ChromoTrace: Reconstruction of 3D Chromosome Configurations by Super-Resolution Microscopy. *PLoS Computational Biology* (accepted)

Meyer VH, et al. (2018). LiMMBo: a simple, scalable approach for linear mixed models in high-dimensional genetic association studies. *bioRxiv*. doi: 10.1101/255497



(Opposite): Middle panel shows the stacked swimming patterns for each medaka fish across the entire assay duration. Right hand panel shows the activity of two different strains after introduction into the videoing tank. Strain 5 takes longer to reach a baseline activity level than strain 8.

Brazma group

Functional genomics

The Brazma research group complements the Expression Atlas services cluster, analysing and integrating new types of data across multiple platforms. The group is interested in cancer genomics and elucidating relationships between transcriptomics and proteomics.

In 2017 we continued our work on the comparison of transcript and protein expression levels, and on isoform-level gene expression. We also started to develop methods for single-cell RNA-seq data analysis in collaboration with the Stegle and Marioni groups at EMBL-EBI and the Teichmann group at the Wellcome Sanger Institute.

As participants in the Pan-Cancer Analysis of Whole Genomes initiative, we lead a working group on data integration jointly with our collaborators from the University of California Santa Cruz and the University of Zurich. The data analysis has now been completed and the manuscript posted in Bioarchive, and is under review in *Nature*. One of the highlights of this research, which was done in our group jointly with Zemin Zhang's group from Pecking University, was the discovery and analysis of gene fusions. We also analysed the data for



Alvis Brazma
Head of Molecular Atlas.
Senior Team Leader, Senior Scientist.

PhD in Computer Science, Moscow State University, 1987.
MSc in Mathematics, University of Latvia, Riga.
At EMBL-EBI since 1997.

expression Quantitative Traits (eQTLs) in collaboration with the Stegle group at EMBL-EBI and the Korbel group at EMBL Heidelberg, and prepared the results of these analyses for publication.

Selected publications

Roumeliotis TI, et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. *Cell Reports*. doi: 10.1016/j.celrep.2017.08.010

Perez-Riverol Y, et al. (2017). Synthetic Human Proteomes for accelerating protein research. *Nature Methods*. doi: 10.1038/nmeth.4191

Roumeliotis TI, et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. *Cell Reports*. doi: 10.1016/j.celrep.2017.08.010

Enright group

Functional genomics and analysis of small RNA function

Complete genome sequencing projects are generating enormous amounts of data, and while progress has been rapid, a significant proportion of genes in any given genome are either un-annotated or possess a poorly characterised function. The group aims to predict and describe the functions of genes, proteins and regulatory RNAs, as well as their interactions in living organisms.

Regulatory RNAs have recently entered the limelight as the roles of a number of novel classes of non-coding RNAs have been uncovered. Our work involves the development of algorithms, protocols and datasets for functional genomics. We focus on determining the functions of regulatory RNAs, including microRNAs (miRNAs), piwiRNAs and long non-coding RNAs. We collaborate extensively with experimental laboratories on commissioning experiments and analysing experimental data.

Major achievements

The detection of known miRNAs and the prediction of novel miRNAs are important for our understanding of the evolution of these molecules and their roles in functional specialisation.

MiRNAs are approximately 19–22 nucleotides in length and their most characterised function so far is regulating the bioavailability of messenger RNA for the production of protein. This effect is called post-transcriptional regulation.

Schistosomes are parasitic helminths that cause schistosomiasis, a disease affecting approximately 200 million people, primarily in underprivileged regions of the world. *Schistosoma mansoni* is the most experimentally tractable Schistosome species due to its ease of propagation in the laboratory and the high quality of its genome assembly and annotation. Although there is growing interest in microRNAs in parasitic worms, we know very little about the role these molecules play in the context of developmental processes.

Previously, it has been shown that different stages of the *Schistosoma* parasites express different types of miRNAs. In a recent paper, we focused on utilising differences in gene expression as the readout of potential post-transcriptional regulation. Using bioinformatics tools we found that members of one miRNA family called miR-277/4989 might be responsible for the



Anton Enright
Research Group Leader

PhD in Computational Biology, University of Cambridge, 2003.
Postdoctoral research at Memorial Sloan-Kettering Cancer Center, New York.
At EMBL-EBI since 2008.

change in gene expression observed between juvenile and adult worms (Protasio, 2017). Furthermore, the effect of this miRNA seems to be more prominent in the sexually mature females rather than in immature females.

Generally speaking, although a large number of miRNAs have already been annotated, many other miRNAs that are expressed in very particular cell types remain elusive. Sequencing allows us to quickly and accurately identify the expression of known miRNAs from small RNA-Seq data. Previous approaches to the prediction of novel miRNAs usually involve the analysis of structural features of miRNA precursor hairpin sequences obtained from genome sequence.

We surmised that it may be possible to identify miRNAs by using these biogenesis features observed directly from sequenced reads, solely or in addition to structural analysis from genome data. To this end, we have developed mirnovo, a machine learning based algorithm, which is able to identify known and novel miRNAs in animals and plants directly from small RNA-Seq data, with or without a reference genome (Vitsios, 2017). This method performs comparably to existing tools, however is simpler to use with reduced run time. Its performance and accuracy have been tested on multiple datasets, including species with poorly assembled genomes.

Selected publications

Protasio AV, et al (2017). MiR-277/4989 regulate transcriptional landscape during juvenile to adult transition in the parasitic helminth *Schistosoma mansoni*. *PLOS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0005559

Davis MP, et al (2017). Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. *EMBO Reports*. doi: 10.15252/embr.201744059

Vitsios DM, et al (2017). Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research*. doi: 10.1093/nar/gkx836

As part of the Pan-Cancer Analysis of Whole Genomes project, the Brazma group and Zemin Zhang's group from Pecking University, discovered and analysed gene fusions.



Flicek group

Evolution of transcriptional regulation

The group's research projects leverage comparative regulatory genomics approaches to understand the evolution of transcriptional regulation with a long-term goal of understanding the mechanisms and maintenance of tissue-specific genome regulation.

Major achievements

The leading result from our research into the evolution of transcriptional regulation was a study in collaboration with Duncan Odom's group at the University of Cambridge, jointly led by former postdoctoral fellow, and current Research Associate at INSERM in Paris, Camille Berthelot. This work was the next stage in a line of research addressing regulatory evolution across a wide swath of the mammalian clade, including a previous study that analysed the evolution of genomic promoter and enhancer elements across 20 mammalian species.

We analysed gene expression data from more than 20 mammals to understand how gene expression levels and stability across evolutionary time relate to the active enhancer and promoter elements in the same individuals. Our results demonstrated that the primary driver of both expression level and expression stability is the regulatory complexity (i.e. the number) of regulatory elements rather than whether they are conserved. We found that conserved regulatory elements are more effective at influencing gene expression, but it is also clear that recently evolved and lineage-specific regulatory elements are functional as well.

David Martin-Galvez, former postdoctoral fellow and now faculty at Universidad Complutense Madrid, led an investigation into the feasibility of combining genome variation with conserved transcriptional regulation as a way to identify genomic regions associated with metabolic syndrome. This work, which featured collaborators at the University of Iowa, USA, identified plausible target genes associated with circadian rhythm and with regulation of the TGF-beta signalling pathway. It also provided a proof of principle of using regions of the genome that are functionally conserved in a tissue-specific manner to prioritise sequence variants that may be implicated in disease.

Finally, Emily Wong, former EMBO Advanced Fellow and now a Discovery Early Career Researcher Award (DECRA) fellow at the University of Queensland in Brisbane, Australia, co-led a study designed to tease apart the exact role of genetic variation in transcription factor binding differences between species and integrate these results for both chromatin and gene expression. The experiments used F1 crosses of closely related mice, which enabled us to determine the role of cis and trans variation in the heritability of transcription factor binding intensity. We were also able to definitively show that binding intensity is under genetic control and estimate the genomic scope of influence of a variant on a transcription factor binding site to be about 10Kb.

Future plans

Our line of research addressing regulatory evolution across a wide swath of the mammalian clade is expanding to incorporate multiple tissues in multiple species. This experimental design will enable us to consider both the tissue axis and the species axis in our analysis of regulatory evolution. Preliminary results suggest similar dynamics across all of the somatic tissues that we have thus far profiled, but far less sharing of regulatory regions across tissues than we had predicted.

In addition, we are currently exploring the functional roles of evolutionarily very young CTCF binding sites by identifying those that have arisen recently within the *Mus* genus. This has involved an investigation of CTCF binding profiles across the *Mus caroli* and *Mus pahari* genomes, which we have sequenced and are in the process of fully analysing. We are also investigating CTCF binding in the same F1 model that was described above. These and other projects are building on our long-term collaboration with the Odom group at the University of Cambridge.

Selected publications

Berthelot C, et al. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*. doi: 10.1101/125435

Jasinska AJ, et al. (2017). Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. *Nature Genetics*. doi: 10.1038/ng.3959

Martín-Gálvez D, et al. (2017). Genome variation and conserved regulation identify genomic regions responsible for strain specific phenotypes in rat. *BMC Genomics*. doi: 10.1186/s12864-017-4351-9

Wong ES, et al. (2017). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nature Communications*. doi: 10.1038/s41467-017-01037-x

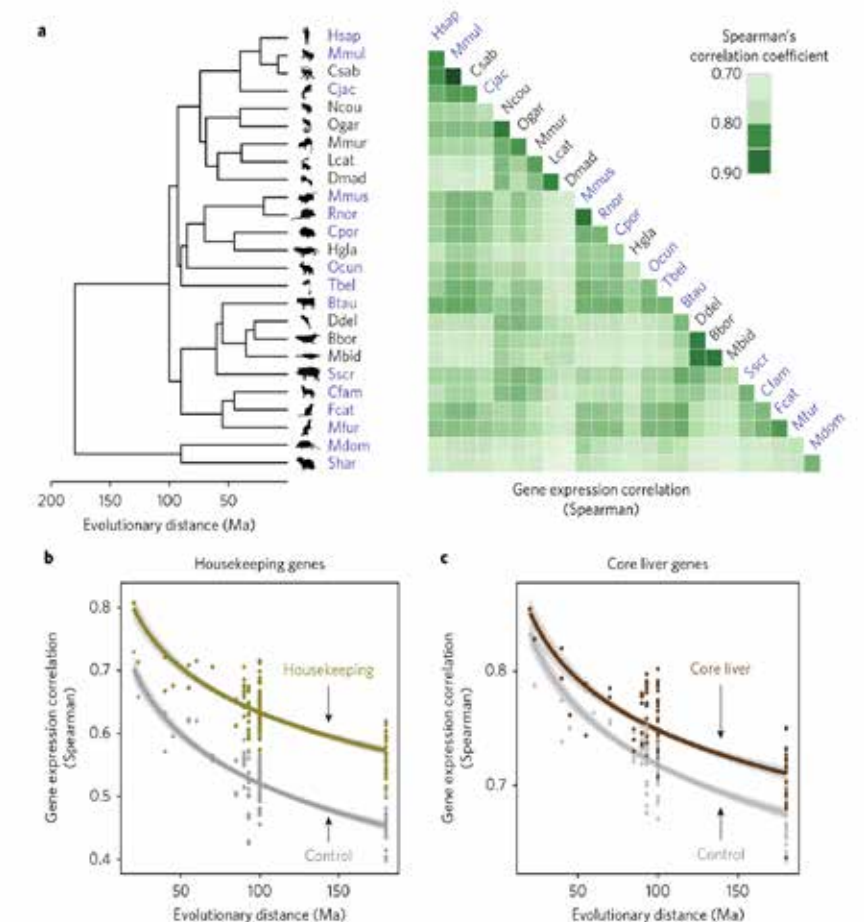


Paul Flicek

Research Team Leader and Senior Scientist

DSc Washington University, 2004.
Honorary Faculty Member, Wellcome Sanger Institute since 2008. Team Leader at EMBL-EBI since 2007, Senior Scientist since 2011.

At EMBL since 2005.



Liver gene expression levels are highly conserved across 25 mammalian species.

Gerstung group

Computational cancer biology

Cancer is a genetic disease caused by mutations to the genome. International efforts such as the International Cancer Genome Consortium (ICGC) have charted the genomic lesions leading to cancer at unprecedented detail and in tens of thousands of patients. A revelation of these projects was an even greater genomic complexity of cancer genomes than previously anticipated: despite having the same disease, each patient harbours a unique constellation of mutations.

Our group uses statistical approaches to enhance the quantitative understanding of cancer. This is critical to extract meaningful signals from big molecular data sets, such as genomics and transcriptomics, as well as imaging and large longitudinal records for thousands to millions of patients.

Specific research questions address the molecular mechanism of mutations, the evolutionary dynamics driving cancer, translational applications to predict the future trajectory from pre-malignant to malignant disease, and prognostic and clinical decision support algorithms.

To address these questions, we develop and use statistical algorithms to discriminate signal from noise in large data sets using high-dimensional statistical learning theory, but also employ machine and deep learning methods. In addition to answering quantitative research questions and developing algorithms we also build clinical decision support tools.

Major achievements

As part of the international Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium the group has provided the first extensive analysis of when mutations arise during cancer development. This has been a long-standing and unanswered question because the transformation of cells to cancer in a given patient is usually unobservable.

It is clear that the mutations found in tumour cells must have been acquired between fertilisation and diagnosis and it emerges that the accumulation of certain mutations is a consequence of normal tissue development as cells are constantly facing mutagenic exposures and add, on average, about one mutation per cell division. It is unclear, however, when this normal process converts into a cancerous one.

The analysis conducted on 2658 cancers revealed that many lesions seen in cancer genomes date back many years and in some cases decades prior to diagnosis. This confirms that the development of cancer is, at least initially, a very slow process spanning many years. The findings may also offer a window of opportunity for specifically picking out those cells that have already acquired mutations making cells more prone to become cancerous.

In line with these findings, members of the group have developed predictive models to calculate the risk of developing cancer based on sequencing of non-malignant tissue samples. In a collaborative study with George Vassiliou (Wellcome Sanger Institute), John Dick (University of Toronto) and Liran Slush (Weizman Institute) we have sequenced blood samples from patients with acute myeloid leukaemia (AML) and healthy controls. All samples were collected in a healthy state 5-10 years prior to leukaemia diagnosis. Many individuals who developed AML harboured mutations characteristic of leukaemia up to ten

years prior to disease. Compared to healthy individuals, such mutations were more frequently found, affected a greater proportion of blood cells and more often co-occurred in the same individual, thereby predicting the risk of AML with about 80% accuracy.

In a third stream of work, researchers from the group have analysed the patterns of mutations generated by deficiencies of DNA mismatch repair (MMR) in human cancers and *C. elegans* as a model system. These analyses demonstrated that the mutations found in MMR-deficient cancers have contributions from errors occurring during DNA replication, but also spontaneous mutation of DNA independent of replication. In primary cancers, these two sources are difficult to delineate, but analyses in worms, in which different components of MMR and DNA replication were genetically manipulated, helped to separate these sources. Together these analyses led to a better understanding of the mutation spectra found in human cancers and characterise the dual functions of DNA mismatch repair.

Future plans

The group will expand its work characterising the evolutionary paths leading to cancer and will create catalogues of genes involved at different stages of cancer development. To achieve this goal we will analyse the genomes of more than 10 000 cancer patients to obtain a comprehensive picture of cancer development. We will also conduct more comprehensive molecular and single-cell analyses of pre-cancerous syndromes to understand the exact transformation between benign and malignant somatic evolution.

We will continue to develop statistical risk models to predict disease progression based on large volumes of genomic, molecular and clinical data. These models will be used to predict prognostic events posterior to cancer diagnosis, to establish risk-adjusted treatment decisions. The same type of models will also be employed to calculate cancer risk based on pre-malignant tissue samples as outlined above with the aim of preventing cancer.

A third pillar of analysis will focus on tools for delineating mutational processes using both systematic mutagenesis screens in model systems and large data sets from cancer samples. At the same time we will develop more comprehensive algorithms to characterise the multifaceted nature of mutational processes, which often generate a variety of mutation types ranging from single nucleotide variants to alterations at the scale of Megabases.

A novel emerging stream of research is driven by the dramatic progress of computer vision due to deep learning, which allows to quantify histopathological

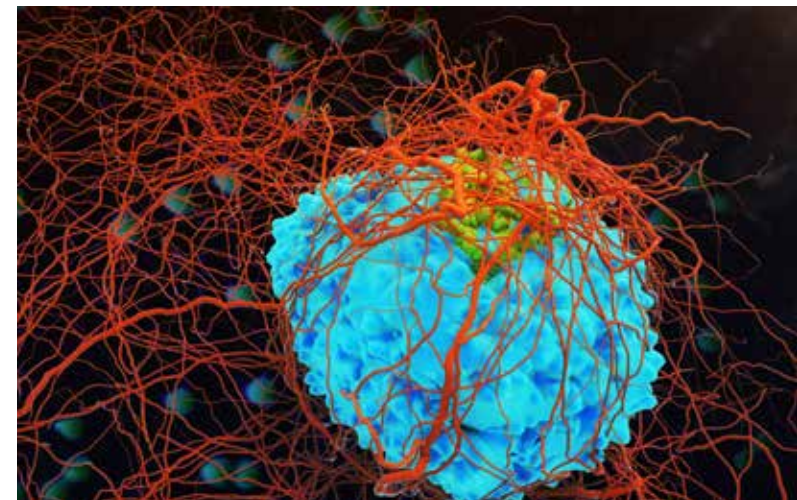


Moritz Gerstung

Research Group Leader

PhD in Computational Biology, ETH Zurich, 2012. Postdoctoral research at the Wellcome Sanger Institute. At EMBL since 2015.

abnormalities. This will help quantitatively understand the associations of tissue architecture and genetic and molecular changes and also utilise imaging alongside genetic and molecular data in diagnostic and prognostic algorithms.



As part of the international PCAWG consortium, the Gerstung group has provided the first extensive analysis of when mutations arise during cancer development.

Selected publications

González S et al. (2017). Immuno-oncology from the perspective of somatic evolution. *Seminars in Cancer Biology*. doi: 10.1016/j.semcancer.2017.12.001

Gerstung M et al. (2017). Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics*. doi: 10.1038/ng.3756

Martincorena I et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. doi: 10.1016/j.cell.2017.09.042

Meier B et al. (2018). Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *bioRxiv*. doi: 10.1101/149153

Gerstung M et al. (2017). The evolutionary history of 2,658 cancers. *bioRxiv*. doi: 10.1101/161562

Goldman group

Evolutionary tools for genomic analysis

The diversity of all life has been shaped by its evolutionary history. The group's research focuses on the processes of molecular sequence evolution, developing data analysis methods that allow us to exploit this information and glean powerful insights into genomic function, evolutionary processes and phylogenetic history.



To understand the evolutionary relationships between all organisms, it is necessary to analyse molecular sequences with consideration of their underlying structure. This is usually represented by an evolutionary tree indicating the branching relationships of organisms as they diverge from their common ancestors, and showing degrees of genetic difference between them.

We develop mathematical, statistical and computational techniques to reveal information from genome data, draw inferences about the processes that gave rise to these interrelationships and make predictions about the biology of the systems whose components are encoded in those genomes. We develop new evolutionary models and methods, sharing them via stand-alone software and web services, and apply new techniques to interesting biological questions. We participate in comparative genomic studies, both independently and in collaboration with others. Our evolutionary studies involve the analysis of next-generation sequencing (NGS) data, which enables enormous gains in our understanding of genomes but poses many new challenges.

Major achievements

A recurring feature observed in human genome sequencing data is a surprisingly high frequency of complex mutations, composed of apparent combinations of successive, nearby, base substitutions, insertion and deletions. We developed a generalised mutation model to describe local template-switching during DNA replication and used this to study the role of template switch events in the origin of these mutation clusters. Under this model, short genome regions (typically up to 25 bp, although sometimes longer) are replaced during replication by similar-length fragments copied from a nearby location on the complementary strand. The process appears to be similar to large-scale genome rearrangements that have been associated with genetic disease; previously overlooked, this local template-switching process explains many complex mutations more parsimoniously than invoking a process of successive base substitutions, insertions and deletions within a single cluster.

Applied to the human genome, our model successfully detects thousands of template-switch events during the evolution of human and chimp from their common ancestor, and hundreds of events between two independently sequenced human genomes. A number of human disease mutations have also been attributed to events that can be explained by our model.

We showed that human resequencing project reference data contain many erroneous variant annotations caused by these mutation clusters, as existing next-generation sequencing (NGS) mapping algorithms fail to place reads with multiple differences from the reference, leading to errors in subsequent variant calling.

In 2017 we completed an investigation into the impact of popular multiple sequence alignment (MSA) programs on reconstruction of ancestral sequences. Many researchers are interested in synthesising proteins of parental or extinct species to study their biochemical properties and compare them with those

of their extant relatives. Accurate reconstruction of ancestral states is vital to these studies; however, we discovered that different aligners introduce various biases, including a widespread tendency to overestimate the lengths of ancestral sequences. While all aligners give similar results under easier conditions, when faced with more challenging alignment problems (e.g., greater evolutionary divergence or higher insertion and deletion rates) there was considerable variation in the overall accuracy of inferred ancestral sequences and in the results achieved for more-recent or more-ancient ancestors. We were able to give guidelines for other researchers to choose the MSA method most likely to give them good ancestral sequence reconstructions.

Our phylogenetics work included an exploration of how the rate of evolution of a genomic region affects its usefulness for uncovering evolutionary relationships. Regions evolving slowly contain too few informative mutations; those that evolve quickly can have so many mutations that many informative ones are 'masked' by subsequent mutations, leading to more noise. Our work helps us understand what evolutionary rates are optimal for reliable phylogeny estimation.

The group also worked on statistical methods suitable for the unbiased estimation of phylogenies from "truncated" MSAs: alignments that omit sequence positions at which no mutations are observed. If no allowance for truncation is made in the phylogenetic analysis of such data, estimated evolutionary rates can be massively inflated. We devised, implemented and tested a correction to standard maximum likelihood phylogenetic estimation, showing excellent performance.

We have also completed two projects in the application of machine learning methods to long-standing problems in structural bioinformatics. The first of these concerns the structural changes that proteins undergo when moving from one conformation to another. We adapted LASSO regression, a machine-learning method, to show that in most cases a transition pathway between the two states can be calculated using a simple mechanical model of protein structure in which the protein is represented as a system of balls and strings.

The second machine-learning application concerns the prediction of the 3D structure of protein complexes from the structures of the unbound protein molecules comprising them. We developed the Integrative Ranking of Protein-Protein Assemblies (IRaPPA) method of ranking structural models of docked complexes by combining many different biophysical models together. When applied to structures generated with four state-of-the-art docking methods, IRaPPA showed a large increase in performance. IRaPPA now been implemented in the SwarmDock, ZDOCK and PyDock protein-protein docking web servers.



Nick Goldman

Research Group Leader and Senior scientists

PhD University of Cambridge, 1992. Postdoctoral work at National Institute for Medical Research, London, 1991-1995, and University of Cambridge, 1995-2002. Wellcome Trust Senior Fellow, 1995-2006. EMBL Senior Scientist since 2009. EMBL-EBI Joint Head of Research since 2017. At EMBL-EBI since 2002.

Future plans

We will continue to improve and develop new methods for phylogenetic analysis, and new techniques to analyse incomplete datasets. One example underway is the development of the concept of "effective sequence number", which will provide a statistically principled method for weighting the contribution to data analyses that should be given to sequence data that share common ancestry.

Having established the existence of short-range template-switching events in human evolution, we will seek to further characterise the process. In collaboration with Aylwyn Scally from the University of Cambridge Department of Genetics, we will investigate the consequences of the template-switching process for understanding population variation, evolution and disease by exploring the growing number of high quality de novo-assembled genomes

We also aim to study how evolutionary forces can shape protein-protein interaction networks. This includes building a model of how evolutionary pressures, acting on the networks' cellular information processing properties by the strengthening and weakening of interaction links, can be revealed in the sequences of the proteins making up the network.

Selected publications

Hayes TW, Moal IH (2017). Modeling protein conformational transition pathways using collective motions and the LASSO method. *Journal of Chemical Theory and Computation*. doi: 10.1021/acs.jctc.6b01110

Klopfstein S, et al. (2017). More on the best evolutionary rate for phylogenetic analysis. *Systematic Biology*. doi: 10.1093/sysbio/syx051

Löytynoja A, Goldman N (2017). Short template switch events explain mutation clusters in the human genome. *Genome Research*. doi: 10.1101/gr.214973.116

Moal IH, et al. (2017). IRaPPA: information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics*. doi: 10.1093/bioinformatics/btx068

Iqbal group

Computational microbial genomics

The group works on fundamental computational methods for analysing genetic variation in sequence data from microbial species, and on translating this into the clinic. Half of this work is “hardcore” computer science, building new algorithms and reliable software – in particular for describing how different individual genomes differ within a species. The other half is more applied, working on translating these methods into diagnostics and surveillance for drug-resistant bacterial infections.

We work closely with Public Health England on their sequencing pipeline for tuberculosis (TB), and with many collaborators worldwide on a project to study 100 000 *M. tuberculosis* genomes and phenotype half of them for drug resistance.

During the group’s first year at EMBL-EBI, we recruited a bioinformatician to work on the 100 000 TB genomes project, a C++ developer to take over development of our graph genome software, and a PhD student to work on nanopore analysis of tuberculosis infections. We also recruited an ESPOD postdoctoral researcher, who is an expert on mobile genetic elements (to work jointly with Prof. Nick Thomson at the Wellcome Sanger Institute on the evolutionary history of plasmids and integrons), and an EIPOD postdoctoral researcher who is an expert on TB (to work with me and Dr Misha Savitski at EMBL Heidelberg on proteomic and transcriptomic analysis of drug resistance in *M. tuberculosis*).

Major achievements

The world is accumulating stores of sequenced DNA at an exponential rate. However, the vast majority of these datasets are inaccessible to (sequence) search, making them mostly useless for anything except direct replication of the originating scientific study. However, if we could search all previous genomes for a specific mutation, or species, or plasmid, or drug resistance gene, we could revolutionise epidemiology (enabling automated outbreak tracking), clinical microbiology and the study of mobile elements.

To solve the issue, PhD student Phelim Bradley reinvented a forgotten idea from 1990s web search, and combined it with our knowledge of bacterial population genetics. Using this, we have successfully indexed the entire global corpus of bacterial and viral DNA as of December 2016, and made it for the first time accessible to search.

We published a preprint in December 2017, giving 3 demonstrations. First, we searched the ENA (nearly 500 000 samples) for the new MCR-1 drug resistance gene in under one second, showing it is now possible to provide real time global monitoring of antibiotic resistance genes. Second, we searched for over 2000 plasmids and estimated their host range, showing patterns of sharing between species and phyla, including some plasmids that have successfully made trans-phylum jumps. Finally, we plotted the rise over time of drug resistance in the archive.

Our group also focuses on *Mycobacterium tuberculosis*, the biggest killer of all bacteria, infecting 10 million people in 2016 and causing over one million deaths. Infections can develop slowly and asymptotically, and drug resistance can manifest during an infection within the patient if they are treated with just one drug. As a result, standard treatment involves four drugs, making it harder for the bacteria to evolve multiple resistances simultaneously. Unfortunately, drug resistance has spread sufficiently that it is now necessary to take a clinical sample and test it to see which drugs it is resistant to. The traditional approach is slow, taking months due to the slow growth of the bacteria. An alternative approach is to sequence the DNA of the bacteria and spot mutations known to cause resistance.

As of the start of 2017, this could be done in around two weeks, most of which is spent waiting for the bacteria to grow in culture. Working with collaborators at the University of Oxford over a few years, we developed a method for extracting DNA directly from the sputum (with no culture) and then sequencing directly. We managed to achieve a turnaround time of less than 24 hours, and, with the handheld Oxford Nanopore sequencer, we reduced it to about 12.5 hours. The potential of a portable, handheld point of care test for TB is enormous, and our paper, published in the *Journal of Clinical Microbiology* in March, was a big step forward.

Along with collaborators in Oxford, Vietnam and India, Dr Iqbal was awarded a Longitude Prize Discovery Award to trial nanopore sequencing of TB in Ho Chi Minh City and Mumbai. We want to see what the challenges are of doing this in the field.

Public Health England went live in March sequencing all tuberculosis samples, and using Dr Iqbal’s Mykrobe predictor tool as part of their workflow. England is the first country in the world to do this.

Future plans

We are planning a trip to Mumbai in early 2018 with the head of the Public Health England TB labs to share our experience of sequencing for TB, and as reconnaissance and requirement-gathering for our planned TB global surveillance tool. This will be complemented by trialling nanopore TB diagnostics in the field in Madagascar, Vietnam and India.

We also plan to work with Public Health from England, Scotland, Wales and Northern Ireland to trial our TB sequence analysis pipeline as a replacement for the current Public Health England solution.

Last, but not least, the group aims to work with the ENA to trial a live index of the bacterial and viral data within the archive.

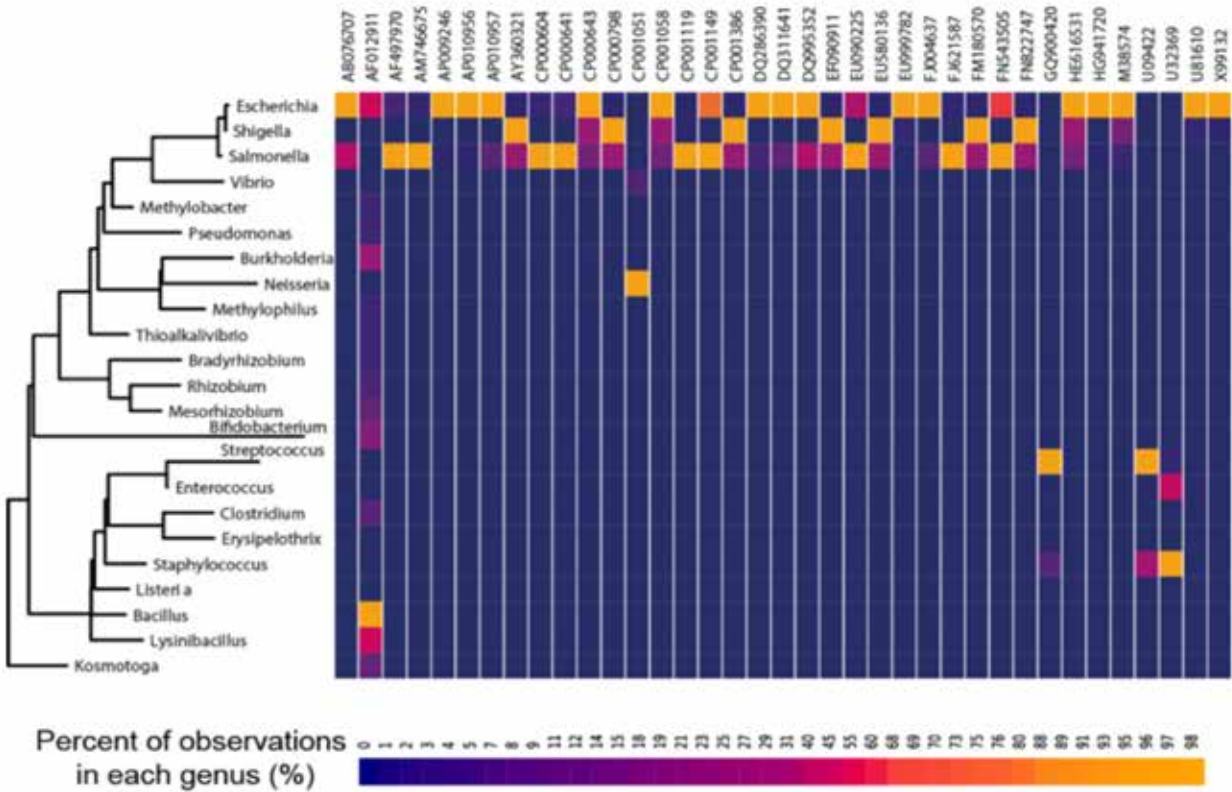


Zamin Iqbal
Research Group Leader
Phd in Mathematics, University of Oxford 2000. Scientific programmer for 1000 Genomes project EMBL-EBI, 2008. Postdoctoral researcher and then group leader, University of Oxford, 2009-2016. At EMBL-EBI since 2017.

Selected publications

- Votintseva AA, et al. (2017). Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *Journal of Clinical Microbiology*. doi: 10.1128/JCM.02483-16
- Bradley P, et al. (2017). Real-time search of all bacterial and viral genomic data. *bioRxiv*. doi: 10.1101/234955

Heatmap showing the frequency of plasmids within each genus, after performing query of our search index of the ENA for 2000 plasmids.



Thornton group

Proteins: Structure, function and evolution

The Thornton group aims to understand how biology works at the molecular level, with a particular focus on protein structure, evolution and ageing.

We explore how enzymes perform catalysis by developing novel software tools that allow us to characterise enzyme mechanisms and navigate the catalytic and substrate space. In parallel, we investigate the evolution of these enzymes to discover how they can evolve new mechanisms and specificities. This work is based on protein-structure classification data derived by colleagues at University College London (UCL). We aim to improve the prediction of function from sequence and structure and to enable the design of new proteins or small molecules with novel functions.

We also explore sequence variation between individuals, especially those variants related to rare diseases. We study the structure, function and impact of those variants which cause developmental diseases in children, in collaboration with the Wellcome Sanger Institute and the University of Cambridge.

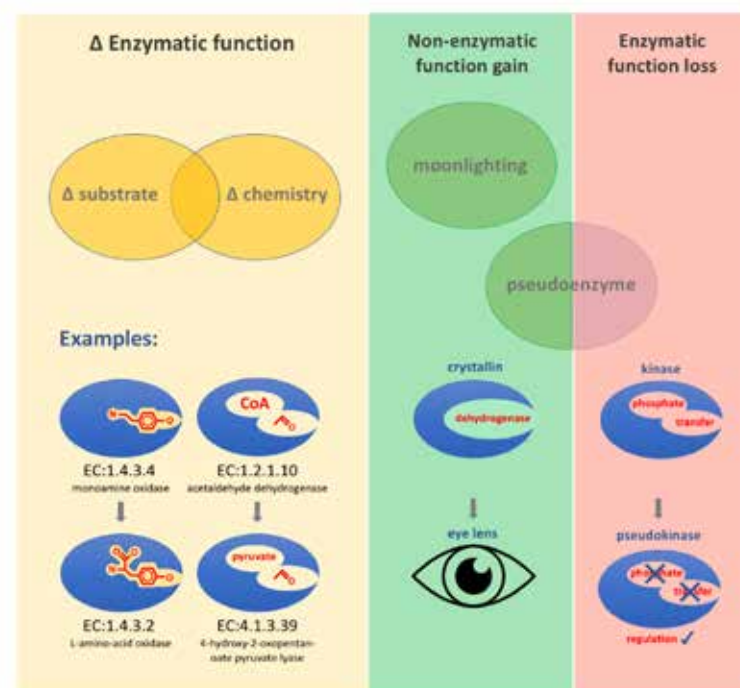
Our close collaboration with experimental biologists at UCL and the Babraham Institute allows us to EXPLORE the molecular basis of ageing in different organisms. We help analyse functional genomics data from model organisms and humans. Recent work has focused on two aspects: (1) the epigenetics of ageing and (2) the impact of small molecules on ageing as potential therapeutics and as probes to untangle the molecular hallmarks of ageing.

Major achievements

Evolution of new enzyme functions

We have used our computational tools for analysing enzyme structure and function to improve our understanding of the evolution of new functions in enzyme families. We have identified examples of creeping and leaping evolution.

Different types of functional changes in enzymes. The figure includes examples of (1) creeping evolution, involving incremental changes to the binding cavity leading to minor changes in functionality (example based on binding pocket mutations to give the substrate creep from EC:1.4.3.4 (monoamine oxidase) to EC:1.4.3.2 (L-amino-acid oxidase)) (2) leaping evolution involves a radical shift in function, dramatically altering substrate binding or chemistry (example based on binding pocket mutations to give the leap between EC:1.2.1.10 (acetaldehyde dehydrogenase) and EC:4.1.3.39 (4-hydroxy-2-oxopentanoate pyruvate lyase)). Inset molecular structures generated with MarvinSketch.



In the former, the substrate is changed slightly, which from a molecular perspective is a trivial change, but from a biological perspective can generate a completely new molecule of vital importance for metabolism of the organism.

At the other extreme, the enzyme may change or modulate its catalytic mechanism, sometimes using the natural limited promiscuity seen in many enzymes to facilitate the evolution of a new function (Tyzack et al, 2017). We are now examining individual enzyme families to identify a limited number of evolutionary paradigms to change function.

Structural analysis of variants causing developmental disorders

Our group explores whether information derived from 3D structures of protein molecules can help us understand the consequences of mutations on the molecules and the organism. Using data obtained by the Deciphering Developmental Disorders (DDD) Consortium, we have analysed some of the *de novo* variants found in children with developmental disorders (Evers et al, 2017).

We identified 19 *de novo* mutations in the DYRK1A gene, including five missense mutations. Protein structural analysis reveals that the missense mutations are either close to the ATP or peptide binding-sites within the kinase domain, or are important for protein stability, suggesting they lead to a loss of the protein's function mechanism. Furthermore, there is some correlation between the magnitude of the change and the severity of the resultant phenotype. Overall, we suggest that *de novo* dominant mutations in DYRK1A account for nearly 0.5% of severe developmental disorders due to substantially reduced kinase function.

Drug repurposing for ageing research

Many increasingly prevalent diseases share a common risk factor: age. However, little is known about pharmaceutical interventions against aging. An important challenge is to assess the potential to repurpose existing drugs for initial testing on model organisms.

To this end, we present a new approach to rank drug-like compounds with known mammalian targets according to their likelihood to modulate aging in the invertebrates *Caenorhabditis elegans* and *Drosophila* (Ziehm et al, 2017). Our approach combines information on genetic effects on aging, orthology relationships and sequence conservation, 3D protein structures, drug binding and bioavailability. Overall, we rank 743 different drug-like compounds for their likelihood to modulate aging. The top ranked compounds are promising as research tools and ultimately a step towards identifying drugs for a healthier human aging.



Janet Thornton

Director Emeritus,
Senior Scientist

PhD King's College & National Institute for Medical Research, London, 1973. Postdoctoral research, University of Oxford, NIMR & Birkbeck College, London. Lecturer, Birkbeck College 1983-1989. Professor of Biomolecular Structure, University College London since 1990. Bernal Professor at Birkbeck College, 1996-2002. Director of the Centre for Structural Biology at Birkbeck College and University College London, 1998-2001. Director of EMBL-EBI 2001-2015, EMBL Senior Scientist since 2001.

Future plans

In our quest to understand enzymes, we have recently combined our two enzyme databases to create a new resource (M-CSA), which contains all data for enzymes of known structure and mechanism. We will use it to de-convolute mechanisms into their individual steps and derive catalytic motifs from the 3D structures, which contribute to the chemistry. We will also use these tools to consider how enzymes transform substrates to create new products, to facilitate the design of new enzymes with new substrates and products.

For the variant analysis, we will continue our development of tools to inspect new variants. We will continue our ageing studies, exploring longevity sub-phenotypes, including their molecular signatures and identifying the small molecules that might modulate lifespan. We would also like to understand the biological mechanisms behind the epigenetic clock, including the effects of tissue turnover and the molecules involved in the epigenetic machinery of the cell.

Selected publications

Tyzack JD, et al. (2017). Understanding enzyme function evolution from a computational perspective. *Current opinion in structural biology*. doi: 10.1016/j.sbi.2017.08.003

Evers JM, et al. (2017). Structural analysis of pathogenic mutations in the DYRK1A gene in patients with developmental disorders. *Human molecular genetics*. doi: 10.1093/hmg/ddw409

Ziehm M, et al. (2017). Drug repurposing for aging research using model organisms. *Ageing cell*. doi: 10.1111/accel.12626

Beltrao group

Evolution of cellular networks

The Beltrao group studies how cellular functions have diverged during evolution, as well as how they are altered in disease. The group studies the molecular sources of phenotypic novelties, exploring how DNA changes are propagated through molecular structures and interaction networks to give rise to phenotypic variability.

We use post-translational modifications (PTMs) data from mass-spectrometry experiments to study the evolutionary dynamics and functional importance of post-translational regulatory networks. We aim to reconstruct the ancestral states of PTM regulatory networks in order to understand how some of the wondrous cellular functions that exist today were like in their primitive forms. We are also increasingly interested in understanding how these regulatory systems make decisions in present day species and how they are re-wired in the context of disease (e.g. cancer or infection). We are studying how genetic variation seen in cancer cells changes their signalling state with an aim to understand context dependent cellular vulnerabilities to drugs.

Beyond PTM regulatory networks, we are broadly interested in studying why different individuals or species diverge in their response to drugs, other environmental perturbations or additional genetic changes. For this purpose we are developing a general framework to predict the molecular consequences of DNA changes (www.mutfunc.com) and using these to guide genotype-phenotype associations.

Major achievements

After ingestion of a blood meal, the malaria parasites transition from the blood of mammals, as gametocytes, to insects where they undergo a very rapid transformation named gametocyte activation. In collaboration with the Choudhary and Billker labs at Wellcome Sanger Institute, we studied the changes in protein phosphorylation occurring within the first few seconds of this process (Invergo BM et al, 2017). A time-course profile of protein phosphorylation was generated using mass-spectrometry which resulted in the observed phosphoregulation of hundreds of proteins within the first 20 seconds of gametocyte activation.

We observed that replication and mitosis associated proteins are phosphoregulated simultaneously, which raises questions about how this process is regulated. Through computational analysis of this data, we were able to associate putative novel kinases to this signalling network and we further studied the CDPK4 and SRPK1 kinases through phosphorylation studies of the knock-out lines.

In addition to the study of protein phosphorylation, we are also studying the control of protein degradation. We analysed the extent by which gene copy number variations at the DNA level are reflected both at the level of mRNA and protein expression (Gonçalves E et al. 2017). Using tumour samples, we observed that 23% to 33% of proteins have copy number changes that are reflected in mRNA changes but are attenuated at the protein level. These attenuated proteins are enriched in stable protein complex subunits. Our observations are consistent with a model whereby the abundance of the complex sets the limit of the abundance of the subunits and that many subunits are produced in excess and degraded when unbound. We also observed that not all complex subunits are attenuated and instead some can act as “rate-limiting” as they appear to set the abundance of the full complex.

We are also interested in understanding how differences in the genomes of individuals relate to their phenotypic differences. In this context we have assembled and studied a diverse panel of around 900 strains of *E. coli* for which we collected genome sequences and condition specific growth responses in hundreds of conditions (Galardini M et al, 2017). With this collection of data, we attempted to predict the condition specific growth phenotypes for each strain from their genome sequences and what we know from gene function based on a commonly studied lab reference *E. coli* K-12 strain.

Future plans

We have shown in the past that protein phosphorylation and other PTMs can diverge quickly in evolution. This has led some to speculate that a significant fraction of these modifications may serve no purpose for organismal fitness. We will continue to work to study the extent by which phosphosites contribute to fitness and which ones in particular are more likely to be of high biological importance. In addition, we are increasingly interested in using the large genetic variation observed in cancer cells to study the processes underlying human cell biology.

Selected publications

Invergo BM, et al. (2017). Sub-minute Phosphoregulation of Cell Cycle Systems during Plasmodium Gamete Formation. *Cell Reports*. doi: 10.1016/j.celrep.2017.10.071

Gonçalves E, et al. (2017). Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Systems*. doi: 10.1016/j.cels.2017.08.013

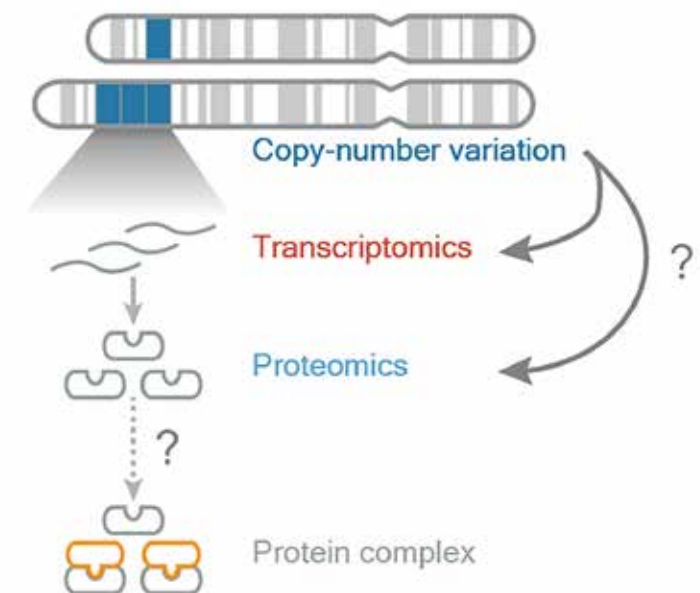
Galardini M, et al. (2017). Phenotype inference in an Escherichia coli strain panel. *Elife*. doi: 10.7554/eLife.31035



Pedro Beltrao

Research Group Leader

PhD in Biology, University of Aveiro, 2007. Postdoctoral research at the University of California San Francisco, US. Group Leader at EMBL-EBI since 2013.



Pan-cancer effects of copy-number Variation on transcript and protein

Marioni group

Computational and evolutionary genomics

The Marioni group uses computational models to understand the molecular mechanisms that underlie cell fate decisions. The group has been one of the pioneers in this field, publishing some of the earliest and most widely used computational methods, as well as using them to study cell fate decisions in a variety of contexts, focussing especially on early mammalian development.

From a methodological perspective, we develop sophisticated and rigorously tested statistical models that help us understand cell fate decisions using single-cell data. These methods range from approaches for normalising and controlling confounding factors through to modelling of cell fate decisions in four dimensions (space and time). We ensure that our code is well-documented and accessible to the wider scientific community, thereby facilitating use of these tools, as well as establishing building blocks for further methodological development.

Together with experimental colleagues, we apply these tools to understand fundamental biological questions, focussing particularly on early mammalian development where the small number of cells in the embryo and the huge variety of cell fate decisions being made are particularly amenable to single-cell genomics.

Major achievements

In 2017 we built on our previous work in the field of single-cell transcriptomics, developing a variety of computational tools for interrogating large-scale datasets. One example is a study of the utility of different normalisation strategies for single-cell RNA-sequencing data, where we demonstrated that approaches from bulk experiments lead to incorrect interpretation (Vallejos et al, 2017). Moreover, we demonstrated that a deconvolution approach developed within the lab provides a good solution to this problem (Lun et al, 2016).

In another study we noted that large-scale single-cell RNA sequencing (scRNA-seq) datasets produced in different laboratories and at different times contain batch effects that could compromise data integration and interpretation. Existing scRNA-seq analysis methods incorrectly assume that the composition of cell populations is either known, or the same, across batches. We developed a strategy for batch correction that is based on the detection of mutual nearest neighbours (MNN) in the high-dimensional expression space (Haghverdi et al, 2018). We demonstrated the superiority of our approach over existing methods using both simulated and real scRNA-seq data sets. We also showed that our MNN batch-effect correction method scales to large numbers of cells.

Turning to applications, in 2017 we explored, in collaboration with Duncan Odom's laboratory at the University of Cambridge, how aging impacts transcriptional dynamics by using single-cell RNA-sequencing to profile hundreds of naive and stimulated CD4⁺ T cells from young and old mice from two divergent species (Martinez-Jimenez et al, 2017).

Turning to cell fate decisions in early development, we led a landmark study on using single-cell RNA-seq to characterise all cell types present within a mouse embryo immediately post gastrulation. We found 20 major cell types, which frequently contain substructure, including three distinct signatures

in early foregut cells. Pseudo-space ordering of somitic progenitor cells identifies dynamic waves of transcription and candidate regulators, which we validated by molecular characterisation of spatially resolved regions of the embryo.

Finally, Dr Marioni's involvement in the Human Cell Atlas project means he will coordinate the development of a range of methods for identifying informative genes, understanding how cells cluster into groups, marker gene detection and alignment of cells along developmental axes.

Future plans

Our group will continue to develop computational tools for understanding the regulation of gene-expression levels. We will focus on methods for analysing single-cell RNA-sequencing data, which has the potential to reveal novel insights into cell fate decisions, cell-type identity and tumour development.

We will develop new computational approaches for handling single-cell RNA-sequencing data, providing robust methods for finding differentially used highly variable genes, as well as assessing the direct impact of various normalisation strategies and the efficacy of extrinsic, spike-in molecules for this purpose.



John Marioni

Research Group Leader

PhD in Applied Mathematics,
University of Cambridge, 2008.
Postdoctoral research in the
Department of Human Genetics,
University of Chicago.
At EMBL-EBI since 2010.

From the biological perspective, we will continue to use our methods to obtain insights into cell fate decisions during gastrulation – arguably the most important time in our lives, focusing on integrating data from multiple stages of early embryonic development and exploiting data from additional molecular layers, such as DNA methylation and chromatin accessibility.

Selected publications

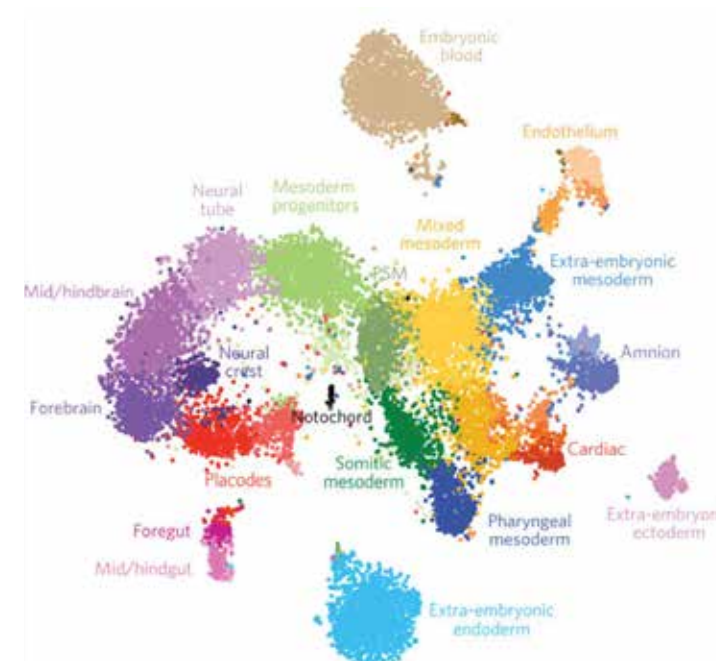
Vallejos CA, et al. (2017). Normalizing single-cell RNA-sequencing data: challenges and opportunities. *Nature Methods*. doi: 10.1038/nmeth.4292

Martinez-Jimenez CP, et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*. doi: 10.1126/science.aah4115

Lun ATL, et al. (2017). Testing for differential abundance in mass cytometry data. *Nature Methods*. doi: 10.1038/nmeth.4295

Ibarra-Soria, et al. (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature Cell Biology*. doi: 10.1038/s41556-017-0013-z

Haghverdi L, et al. (2018). Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbours. *Nature Biotechnology*. doi: 10.1038/nbt.4091



t-distributed stochastic neighbour embedding (t-SNE) plot of all cells from E8.25 mouse embryos (n=19,396 cells) computed from highly variable genes

Petsalaki group

Whole-cell signaling

The Petsalaki group studies human cell signalling with the long-term aim to create predictive and conditional whole-cell signalling models. As a first step in their studies, they are focusing on understanding the relationship between rewiring of signalling networks in different conditions, tissues and genetic backgrounds, and different resulting phenotypes.

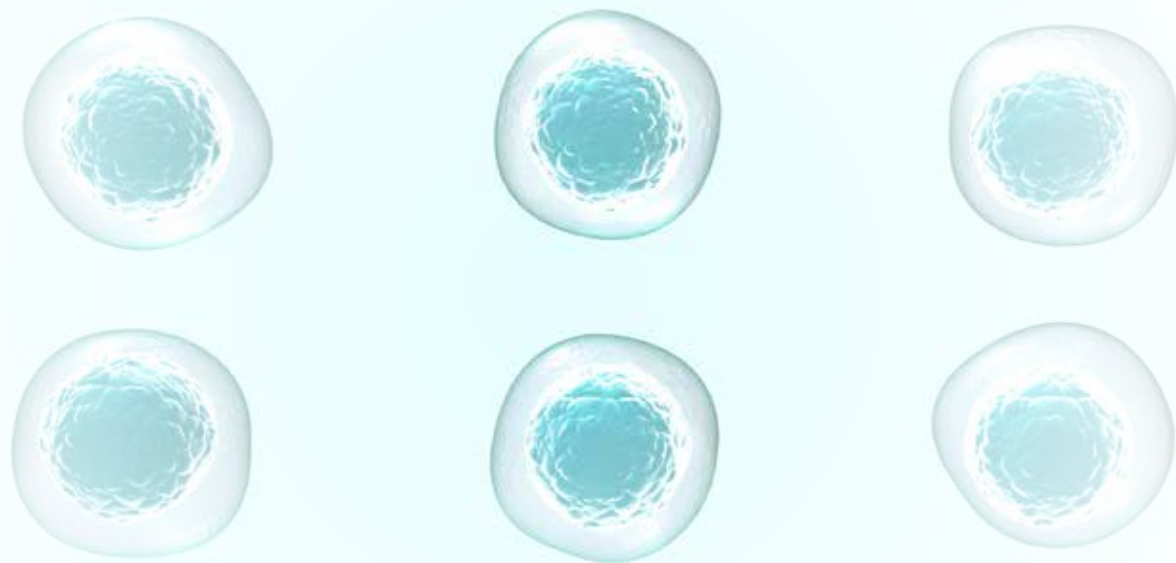
Major achievements

The group uses three major approaches to achieve its research goals. Firstly, in collaboration with the Beltrao group, we use network reconstruction approaches to derive context-specific signalling networks from publicly available phosphoproteomics datasets. In parallel we are developing a method that predicts signal propagation to transcription factors in a network given a phosphoproteomics or transcriptomics dataset.

The second approach, in collaboration with Gavin Wright at the Wellcome Sanger Institute, involves the use of CRISPR screening on cancer cell lines expressing signalling reporters to identify targets that rewire major signalling pathways. The discoveries from this study will be used to study common mechanisms of cell signalling rewiring across all cancer types.

Finally, we collaborate with Anne Claude Gavin at EMBL Heidelberg, to collect global phosphoproteomics datasets at different time points of neutrophil differentiation in cell lines that carry different genetic variants of the CSF3 receptor that is regulating this process. We will use these datasets to study the changes of the signalling networks at different time points of stimulation and how they differ according to the genetic background.

By studying network rewiring in different systems and using different approaches we aim to extract the emerging general principles of this process.



In addition to these studies, we use transcriptomics and imaging data to understand the relationship between the activation of specific transcription factors and pathways and the changes in cell morphology. In collaboration with Greg Findlay at the MRC Phosphorylation Unit in Dundee, we study signalling pathways that affect embryonic stem cell development. We also use multi-omics datasets to study the mechanisms of fatty liver disease progression in collaboration with Toni Vidal-Puig from the Metabolic Research Laboratories, MRC, University of Cambridge and Wellcome Sanger Institute. These studies will also help us understand the interplay of signalling with metabolic pathways, including lipid metabolism, and contribute to our long-term goal of creating whole cell signalling models.

Future plans

We are still in the early days of most projects in our group. We will continue to study the principles of context-specific signalling network rewiring using a combination of computational and experimental approaches. We will furthermore start developing biological modelling approaches to study the effect of perturbations on signalling networks *in silico*.

In terms of biological questions, we additionally aim to study the links between signalling, gene regulatory networks and epigenetics in stem cell differentiation, and the role of changes in lipid abundances on signalling network rewiring. Our experimental models for these studies for the near future will continue to be melanoma cell lines, mouse myeloid cell lines and embryonic stem cells and liver samples from mouse models and human patients of fatty liver disease.



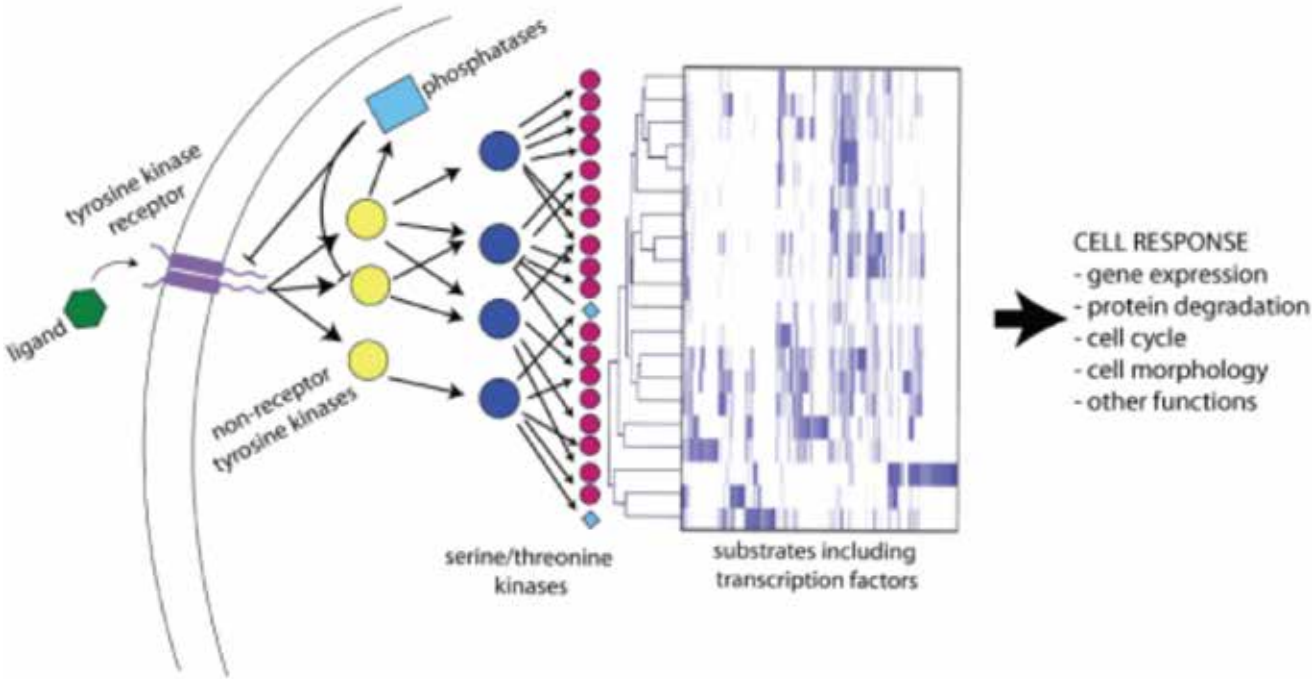
Evangelia Petsalaki
Research group leader
PhD in Structural Bioinformatics with Rob Russell, EMBL & University of Heidelberg. Post-doctoral research the Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, Toronto.
Group Leader at EMBL-EBI since February 2017.

Selected publications

Tian AL, et al. (2018). A recombinant *Fasciola gigantica* 14-3-3 epsilon protein (rFg14-3-3e) modulates various functions of goat peripheral blood mononuclear cells. *Parasites & Vectors*. doi: 10.1186/s13071-018-2745-4

Giudice G, Petsalaki E (2017). Proteomics and phosphoproteomics in precision medicine: applications and challenges. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbx141

Schematic of signal propagation from receptor tyrosine kinases to a cell's response. The clustergram shows an example of a phosphoproteomics dataset that shows the signalling state of a cell in a given condition.



Stegle group

Statistical genomics and systems genetics

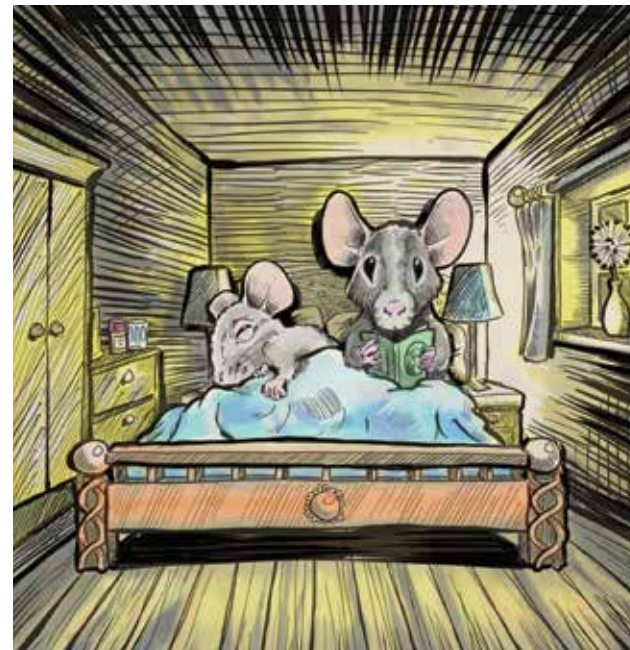
The Stegle group uses computational approaches to map genotype to phenotype on a genome-wide scale. The group seeks to understand how genetic background and environment jointly shape phenotypic traits or cause diseases, how genetic and external factors are integrated at different molecular layers, and how molecular signatures vary between individual cells.

We use statistics as our main tool to answer these questions. To make accurate inferences from high-dimensional omics datasets, it is essential to account for biological and technical noise and to propagate evidence strength between different steps in the analysis. To address these needs, we develop statistical analysis methods in the areas of gene regulation, genome wide association studies (GWAS) and causal reasoning in molecular systems. The group is also actively developing methods for exploiting data from the most recent technologies, in particular single-cell sequencing.

Major achievements

In 2017 we continued to develop statistical approaches for interrogating genotype-phenotypes relationships using single and multiple phenotypes (Casale et al, 2017). A flagship application of methods developed in the group was an initial paper as part of the HipSci project (Kilpinen et al, 2017). Together with collaborators at the Wellcome Sanger Institute, Kings College London and University of Dundee, we established the largest resource of Human Induced Pluripotent Stem cells (hiPSCs) in the UK. Data from close to 200 iPS cells lines from different donors enabled studying how genetic differences between people lead to regulatory gene expression changes in human pluripotent cells. We also identified relationships to risk factors in human cancers and other diseases.

In collaboration with the Furlong lab at EMBL Heidelberg, we also investigated regulatory variants during drosophila embryogenesis (Schor et al, 2017), one of the first investigations of genotype-phenotype associations in a developmental time course. The lab also devised new ways to study the impact of genetics beyond direct effects in an individual itself. Amelie Baud, a Wellcome fellow in the group, studied associations between genetic variants in mouse partners that exert phenotypic effects via social interactions. This paper established a new dimension of phenotypic variation that highlighted new relationships between genetic and environmental factors (Baud et al., 2017).



Research indicates that the health of individual mice is influenced by the genetic makeup of their partners.

In parallel to our efforts in population genomics, we also extended our methodological work for the analysis of single-cell readouts. Among other advances, we developed new methods for decomposing the underlying sources of variation in single-cell transcriptome studies (Buettner et al. 2017), and we developed statistical methods to enhance the analysis of single-cell epigenome variation datasets (Angermueller et al, 2017). Our DeepCpG algorithm brings principles from deep learning to single-cell biology, thereby greatly reducing technical noise and mitigating missing information in single-cell readouts.

Future plans

In 2018 we will continue to develop innovative statistical approaches to analyse data from high-throughput genetic and molecular profiling studies. Our ongoing efforts will focus on approaches for analysing new dimensions of molecular variations. These include

data from new technologies, including spatially resolved assays, but also new experimental settings using single-cell readouts to study inter-individual variation. Such efforts will enable studying genetic effects on new molecular traits, including cell type composition or subtle changes in cell-fate decisions during development and cell differentiation.



Oliver Stegle

Research Group Leader

PhD in Physics, University of Cambridge, 2009. Postdoctoral Fellow, Max Planck Institutes Tübingen, 2009–2012. At EMBL-EBI since November 2012.

Selected publications

Kilpinen H, et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*. doi: 10.1038/nature22403

Schor IE, et al. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*. doi: 10.1038/ng.3791

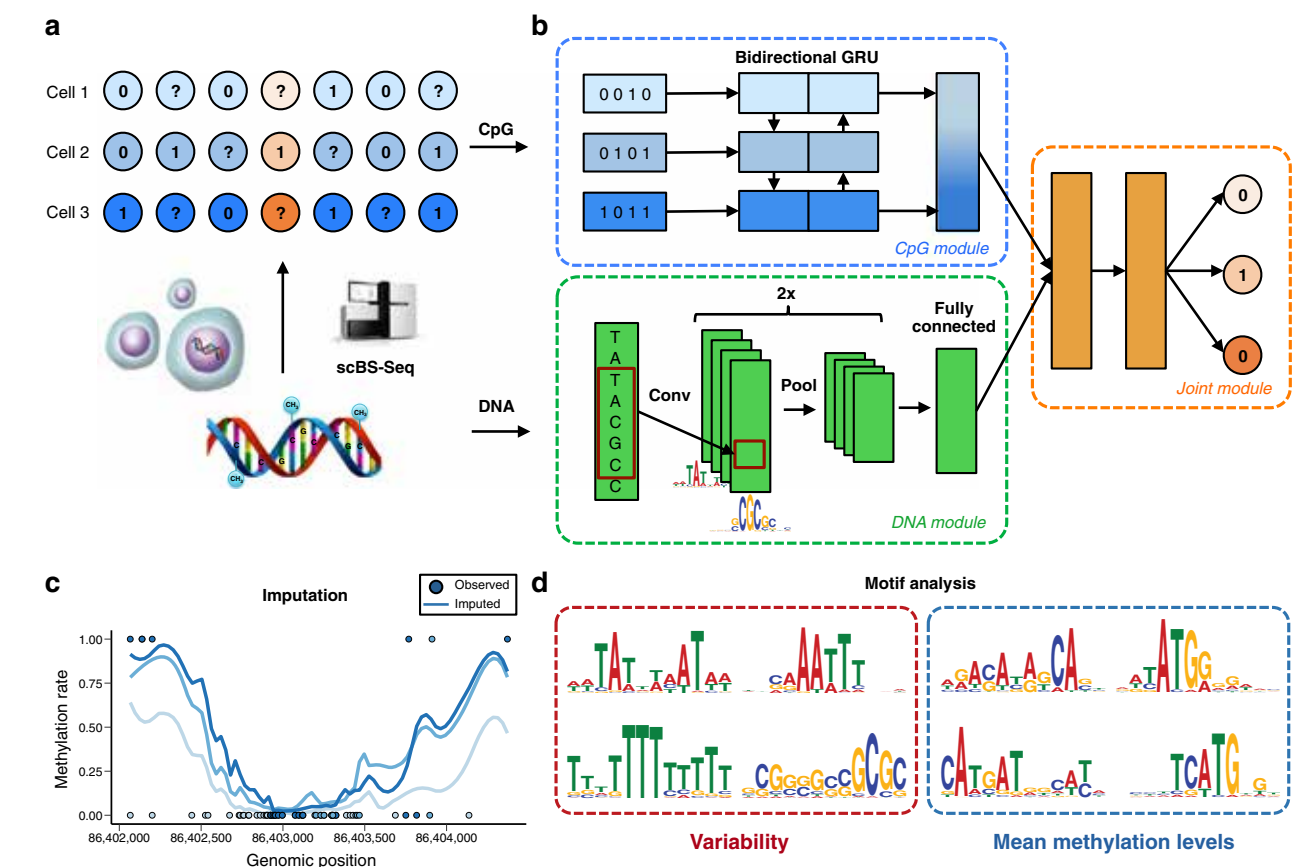
Baud A, et al. (2017). Genetic variation in the social environment contributes to health and disease. *PLOS Genetics*. doi: 10.1371/journal.pgen.1006498

Casale P, et al. (2017). Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLOS Genetics*. doi: 10.1371/journal.pgen.1006693

Buettner F, et al. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*. doi: 10.1186/s13059-017-1334-8

Angermueller C, et al. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*. doi: 10.1186/s13059-017-1189-z

Overview of the DeepCpG model for predicting single-cell DNA methylation states



Training



Astrocytes

Astrocytes perform many functions in the nervous system. They nourish other nerve cells and link many different nerve cells together through their extensive network of processes which radiate from the central cell body. They provide physical support and recycle neurotransmitter molecules such as glutamate. If brain tissue becomes damaged, it is usually astrocytes which fill the space left by other dead cells.

Training

Life-science technologies are advancing at a staggering rate, so the demand for up-to-date, comprehensive and flexible bioinformatics training is higher than ever. Training is at the heart of EMBL-EBI's mission and an important means of supporting EMBL's member states. In 2017, over 185 staff at EMBL-EBI were involved in training and scientific outreach. We reached more than 18 000 people face-to-face and many more online.

Training programme

We have expanded our training offerings to cater to our diversifying user community, which now includes early-career researchers, clinical practitioners, core facility managers, principal investigators from both academic and industrial backgrounds, and more. We have also seen a growing proportion of delegates from industry and healthcare sectors.

EMBL-EBI staff orchestrated and contributed to 340 training and scientific engagement events, supporting biomedical and life-science professionals. These included face-to-face courses at EMBL-EBI, off-site training and workshops at host institutes, demonstrations at conferences and web-based presentations and courses.

Exploring new topics

The training we offer goes beyond understanding and using the EMBL-EBI data resources and looks at the wider competencies required across the bioinformatics community. In 2017, we introduced new topics, including managerial skills, andragogy, service design and the ethical, legal and social implications of research.

Our involvement in pan-European activities to provide training for research infrastructure staff as well as end users was marked by the launch of RItrain's Executive Masters in Management of Research Infrastructures and a new webinar series for operators of research infrastructure as part of the CORBEL project.

Building capacity

In 2017 we also launched CABANA, a bioinformatics capacity-strengthening programme in Latin America, supported by the Global Challenges Research Fund. Working with nine organisations in Latin America, we aim to create a sustainable community of bioinformatics researchers and trainers contributing to world-leading research in three challenge areas: communicable disease, sustainable food production and protection of biodiversity.

Last but not least, we marked the tenth anniversary of EMBL-EBI's Training Programme by delivering our largest ever on-site programme and by orchestrating a series of events and activities throughout the year, culminating in a celebratory symposium.

2017 Training Programme highlights

11
New
online courses

38
New webinars

18 000
People reached
face-to-face

400 240
Unique IP addresses
accessed Train online

340
Training and scientific
engagement events

185
Staff involved

6
Train-the-trainer events

61
Bioinformatics
instructors trained

EMBL-EBI Training Programme in 2017



Photos from a selection of training courses and the tenth anniversary of the Training Programme

Training Programme

The EMBL-EBI Training Programme is coordinated and primarily funded centrally, and benefits from the regular input of scientific and technical experts throughout the institute. Externally-funded projects enable us to contribute more broadly to the development of training on a pan-European and global scale – both for biomolecular researchers and for those who support research by providing services and facilities for those researchers.

This arrangement sets EMBL-EBI's Training Programme apart. Our activities offer unique interactions between service developers and users, providing opportunities to gain invaluable input that can inform the evolution of existing resources and the creation of new ones.

EMBL-EBI's diversifying user community is reflected in our broad range of training offerings. Our programme, courses and materials are created in response to user demand, and cover the full spectrum of EMBL-EBI's activities.

Major achievements

The volume and variety of data generated by life scientists continues to grow; single-cell sequencing, imaging over multiple scales and the application of biomedical informatics in clinical practice are just some of the advances that fuel the ongoing need for training.

To meet the needs of different audiences we have adapted our delivery. In addition to the traditional course model based on lectures and hands-on exercises in a classroom setting, we now have e-learning courses, live and pre-recorded webinars, knowledge-exchange visits and blended learning courses, all of which contribute to broader reach and improved scalability.

In 2017 we provided a growing number of opportunities for those with limited time to get up to speed with data resources. Our Train online portal offered

eleven new online courses and 38 webinars – mostly held as live, interactive events that were captured on video and subsequently made available through YouTube and our website. In total, Train online had over 400 240 visitors in 2017, underscoring the scalability of this flexible, web-based approach. We delivered six train-the-trainer events, in the UK and Germany, training 61 bioinformatics instructors.

We continued to reach out to new audiences. New courses at EMBL-EBI included: Bioinformatics Core Facility Managers; Foundation Skills for HPC in Computational Biomolecular Research; and Data Resources and Tools for Immunologists. Online we ran an extensive webinar series on accessing EMBL-EBI services programmatically and developed a new tutorial-based course on human genetic variation.

EMBL-EBI is a partner in RItrain, an innovative training programme for managers and leaders of research infrastructure. RItrain's pilot Executive Master in Management of Research Infrastructures launched in September 2017, led by the University of Milano Bicocca. Staff exchanges and a new webinar series for technical operators of research infrastructures – our contribution to the CORBEL project – complement RItrain. In both projects our goal is to catalyse the creation of communities of practice who will continue to collaborate and learn from each other after they have participated in the formal training.

In October 2017 we launched CABANA – an exciting new project that aims to address the

slow implementation of data-driven biology in Latin America by creating a sustainable capacity-building programme. With an international consortium of ten organisations, nine in Latin America and one in the UK, the EMBL-EBI-led CABANA project will combine research secondments, workshops, train-the-trainer activities and new e-learning content to strengthen research in three challenge areas of special relevance to Latin America: communicable disease, sustainable food production and protection of biodiversity.

Evaluating the quality, reach and impact of training is a significant challenge; we have put considerable effort into developing meaningful impact measures for our face-to-face training and our approach to impact assessment is maturing.

We have developed a relational database with a web-based user interface to capture and perform basic analytics on our reach (geographical, sector and career-stage-based), perceived quality (evaluated through a survey at the end of each course) and on the ultimate impact on trainees' research (evaluated through long-term surveys). This methodology has been shaped by our involvement in the H2020-funded ELIXIR-Excelerate project. Together with the other ELIXIR nodes, we are adopting a consistent approach to impact assessment.

Satisfaction scores from our post-course surveys remain uniformly high. We have introduced feedback surveys six months and two years after each course to monitor long-term impact. We are in the process of analysing two year's worth of post-six-month surveys and the first batch of post-two-year surveys.

End-of-course surveys

- 94.1% of respondents rate the content of our courses "good" (48%) or "excellent" (46.1%)

Six-month post-course surveys

- Over 91% of course participants indicated they would recommend the course to others
- 85% had disseminated their learning to others
- 70% still use data-analysis methods covered in their course
- 13% established new collaborations with others during the course



Cath Brookbank

Head of Training

PhD in Biochemistry, University of Cambridge, 1993. Elsevier Trends, Cambridge and London, United Kingdom, 1993–2000. Nature Reviews, London, 2000–2002. At EMBL-EBI since 2002.

Last but not least, we punctuated our tenth anniversary year with a number of activities, including an online treasure hunt, free registrations to some of our courses, and a celebratory symposium and party. It was wonderful to be able to share a decade of hard-won experience with our closest colleagues and collaborators, and to thank them for their outstanding contribution to EMBL-EBI's vibrant and ever-changing training programme.

Future plans

Some of the many things we look forward to in 2018 include hosting our first cohort of secondment participants for CABANA, starting a new qualification for biocurators in collaboration with the Institute of Continuing Education at the University of Cambridge, and extending our impact work to include consideration of online learning.

Selected publications

Emery LR, Morgan SL (2017). The application of project-based learning in bioinformatics training. *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1005620

Brazas MD, et al (2017). A global perspective on bioinformatics training needs. *bioRxiv*. doi: 10.1101/098996

Paterson RRM, et al (2017). Microbiology Managers: Managerial Training in the RItrain Project. *Trends Microbiology*. doi: 10.1016/j.tim.2017.03.002

Via A, et al (2017). A new pan-European Train-the-Trainer programme for bioinformatics: pilot results on feasibility, utility and sustainability of learning. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbx112

Morgan SL, et al (2017). The ELIXIR-EXCELERATE Train-the-Trainer pilot programme: empower researchers to deliver high-quality training. *F1000Research*. doi: 10.12688/f1000research.12332

Industry and innovation

Hepatocytes

Hepatocytes make up most of the liver and are involved in a wide variety of functions. They produce many of the serum proteins of the blood, bile salts and cholesterol. They are also involved in carbohydrate and lipid metabolism and detoxify many substances in the body, including a lot of drugs.

Industry and innovation

We work closely with industry through a number of partnerships and collaborations that enable private companies to more effectively use our data resources for their research and development (R&D) activities.

Supporting companies of all sizes

The EMBL-EBI Industry Programme is a subscription-based programme for companies that make significant use of our data and resources as a core part of their R&D. Member companies primarily represent the pharmaceutical sector (most of the top 20 pharma companies), but also the agri-food, nutrition and healthcare industries. The programme provides regular quarterly strategy meetings, expert-level workshops on topics prioritised by the members, and other forms of communication including webinars and face-to-face meetings. In 2017, we welcomed Celgene as a new member of the programme.

EMBL-EBI also offers a separate (non-subscription-based) set of activities to support Small and Medium Enterprises (SMEs). These include outreach activities organised at EMBL-EBI in Hinxton, working with organisations such as OneNucleus, the Royal Society and InnovateUK Medicines Discovery Catapult. In 2017, we also collaborated with the ELIXIR SME & Innovation Forum, MinCyT in Argentina and the EMBL International Relations department for SME-focused activities.

Industry workshops

As biology becomes more data-driven, pre-competitive collaborations, open-source software and informatics standards are becoming essential to improving efficiency and reducing costs for the world's bioindustries. We organise regular quarterly strategy meetings that facilitate dialogue in these areas and encourage members to select topics of mutual interest. In 2017 we delivered 13 successful workshops on topics prioritised by the members, representing both technical areas including standards and more therapeutically-focused workshops.

EMBL-EBI's Industry Programme helps translate basic research into advances in medicine, health, and agriculture for the benefit of society.

Selected industry workshops in 2017

- ⊙ *Informatics resources to support neurodegenerative research*
- ⊙ *Ontologies in Agriculture, Food and Nutrition*
- ⊙ *Informatics and omics for Oncology Drug Resistance*
- ⊙ *Predictive Modelling of Biomarkers*
- ⊙ *The Human Microbiome: Challenges and Opportunities for Novel Therapeutics*
- ⊙ *Single-Cell RNA-sequencing*

Encouraging collaboration

The Industry Programme also serves as an interface between EMBL-EBI industry-focused initiatives and organisations including the Innovative Medicines Initiative (IMI), the Clinical Data Interchange Standards Consortium (CDISC) and the Pistoia Alliance.



The Pistoia Alliance developed a free User Experience for Life Sciences toolkit to help life science companies improve the design of their digital products.

In 2017, the Pistoia Alliance developed a free User Experience for Life Sciences (UXLS) toolkit aimed at helping companies design better digital products for the life sciences and healthcare industries. The toolkit is set to launch in 2018 and contains tips, templates, resources and case studies that will enable businesses to adopt UX principles and methods as they develop scientific software.

Last, but not least, we help EMBL-EBI researchers establish collaborations in the context of the Innovative Medicines Initiative (IMI). Recent examples include the European Bank for induced pluripotent Stem Cells (EBiSC) and EU-AIMS, a large-scale drug-discovery collaboration that brings together academic and industrial R&D with patient organisations to develop and assess novel treatment approaches for autism.

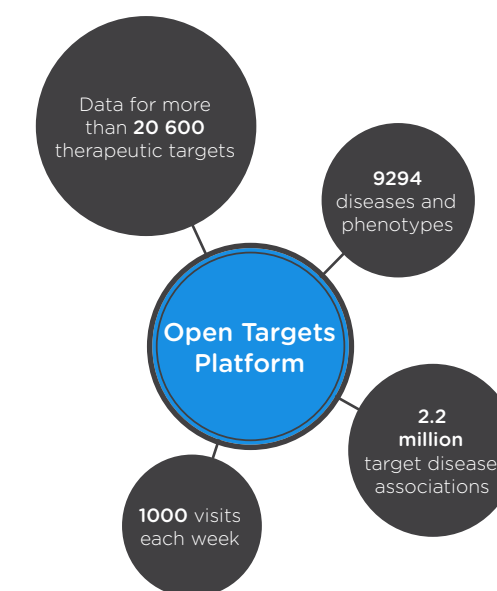
Open Targets

Open Targets is a unique pre-competitive public-private partnership that uses human genetics and genomics data for systematic drug target identification and prioritisation. Founded by EMBL-EBI, the Wellcome Sanger Institute and GSK, the collaboration has grown to include Biogen in 2016 and Takeda in 2017.

The Open Targets Platform helps wet- and dry-lab scientists to discover and prioritise evidence-based relationships between targets and diseases through a comprehensive data service built based on user-experience research. The six releases in 2017 introduced new visualisations, new ways to search and new data sources.

EMBL-EBI plays a central role in the design, development and implementation of the Open Targets Platform. Many EMBL-EBI groups partner with Open Targets in the curation and provision of data to the platform, as well as the development of new and enhanced visualisations. Additionally, EMBL-EBI teams are involved in research projects to develop new capabilities such as the application of network methods to enhance target identification and workflows to assess target tractability. EMBL-EBI's Web Development team leads user-experience research, as well as website design, development and deployment.

EMBL-EBI staff also engage with the Open Targets experimental research programme, working on approximately 40 Open Targets projects, which range from computational pipelines to oncology, induced-pluripotent stem cells and single-cell genomics. A call for new projects was initiated in 2017 with the aim of launching new Open Targets projects in 2018.



Industry Programme

The EMBL-EBI Industry Programme has been an important and vibrant part of EMBL-EBI since 1996, providing regular contact and interaction with key stakeholders and opinion leaders at major global commercial companies and informing them of the institute’s future directions.

Major achievements

EMBL-EBI’s Industry Programme provides a forum for interaction and knowledge exchange for those working at the forefront of commercial bioinformatics. In 2017, we were pleased to welcome Celgene to the Industry Programme. The addition of Celgene, with R&D Labs in California and Spain, allows us to extend further the international reach of our membership, and to continue growing our workshop programme in the US.

Industry Programme members frequently provide support for EMBL-EBI activities either through letters of support for grant applications or by providing specifications for the enhancement of EMBL-EBI services. As an example, the PDX Integrator project, funded by the National Institutes of Health – National Cancer Institute, drew heavily upon discussions and support from Industry Programme members.

Industry workshops in 2017

During 2017, with the support of our Industry Programme members, we organised the following 13 workshops (ten at EMBL-EBI and three in Cambridge, MA, hosted by AstraZeneca, Biogen and Takeda):

- ⦿ *Informatics Resources to Support Neurodegenerative Research*
- ⦿ *Expanded IUPAC Standards for Clinical Information (organised with NCBI)*
- ⦿ *Ontologies in Agriculture, Food and Nutrition*
- ⦿ *Informatics and omics for Oncology Drug Resistance*
- ⦿ *Target Tractability Assessment*
- ⦿ *Digital Biomarkers*
- ⦿ *Predictive Modelling of Biomarkers*
- ⦿ *Bio-pharmaceutical Opportunities in Proteomics*
- ⦿ *User Experiences for Life Sciences (organised with the Pistoia Alliance)*
- ⦿ *The Human Microbiome: Challenges and Opportunities for novel Therapeutics*
- ⦿ *Computational Biology approaches to Neurodegenerative Drug Discovery (hosted by Biogen)*
- ⦿ *Genome Editing in Drug Discovery and Development*
- ⦿ *Single Cell RNA-seq*

Our programme helps its partners and EMBL-EBI researchers establish collaborations in the context of the Innovative Medicines Initiative (IMI). Examples include:

- ⦿ *EBiSC: European Bank for induced pluripotent Stem Cells*
- ⦿ *EU-AIMS: A large-scale drug-discovery collaboration that brings together academic and industrial R&D with patient organisations to develop and assess novel treatment approaches for autism*
- ⦿ *EMIF: Developing a common information framework of patient-level data that will link up and facilitate access to diverse medical and research data sources*
- ⦿ *TransQST: The aim of the Translational Quantitative Systems Toxicology project is to improve the understanding of the safety of medicines. The partners are combining existing data and generating new data to support the development of tools for assessing the safety profile of drug candidates before they enter the clinical testing phase*
- ⦿ *Next Generation of Electronic Translational Safety (eTransafe NexGETS): This emerging infrastructure for preclinical and clinical data sharing supports the analysis of animal data for human safety assessment, the discovery of biomarkers and the development of predictive tools for animal and human safety*

With the support of the Industry Programme members, EMBL-EBI organised 13 workshops in 2017.



Dominic Clarke
Head of the Industry Programme
PhD in Medical Informatics, University of Wales, 1988. Imperial Cancer Research Fund, 1987–1995. United Kingdom Bioinformatics Manager, GlaxoWellcome R&D Ltd., 1995–1999. Vice President, Informatics, Pharmagene, 1999–2001. Managing Consultant, Sagentia Ltd., 2001–2009. At EMBL-EBI since 2006.

Future plans

The Industry Programme will continue to adapt and seek innovative methods of interaction with its members, commensurate with the increasingly global nature of the industries they represent and changes in the services provided by EMBL-EBI.

We will strive to maintain the global and pre-competitive nature of the programme while seeking opportunities to continue to develop the programme internationally, in accordance with the members’ wishes, and, where appropriate, bring in additional members who are major end users of EMBL-EBI data resources.

We will build on the success of our industry interactions through regular meetings and workshops held at EMBL-EBI and at member sites, seeing our interactions with industry partners becoming stronger as we work together to address shared challenges and opportunities created by big data in all the life sciences.

Selected publications

Meehan TF, et al. (2017). PDX-MI: Minimal Information for Patient-Derived Tumor Xenograft Models. *Cancer Research*. doi: 10.1158/0008-5472.CAN-17-0582

Karamanis N, et al. (2018). Designing an intuitive web application for drug discovery scientists. *Drug Discovery Today*. doi:10.1016/j.drudis.2018.01.032

Technical Services



Osteoblasts

Bone is the framework on which a vertebrate body is built, providing strength and support. Cells called osteoblasts are responsible for bone formation. They work in organised clusters to form the organic portion of bone, and then to secrete hydroxyapatite to mineralise it.

Technical Services

Our Technical Services Cluster comprises five teams, delivering a broad portfolio of IT services that support the service provision, research and administrative activities of EMBL-EBI.

Major achievements

In 2017, demands for EMBL-EBI's data resources continued to grow. Our technical services teams manage over 200 petabytes of raw storage. Our main compute farms now have over 30 000 cores to support our users in their data analysis.

As we see demand for data storage increase significantly, we are exploring how to make use of our campus and leased data centre space alongside external cloud providers to maximise our operational responsiveness while delivering the best value to our funders. We are engaging with the European Open Science Cloud to see how the next generation of e-Infrastructure activities can be established to meet the needs of the life-science community and how EMBL-EBI's open data would be integrated into this ecosystem.

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 MySQL, MongoDB, Oracle and PostgreSQL. Delphix has now been deployed across all of our data centres as a database virtualisation technology supporting Oracle and PostgreSQL. This has allowed us to quickly deploy and replicate databases to support the development and operation of our services.

The web visibility of EMBL-EBI data resources and tools depends largely on work done by our Web Production team. Approximately 1000 virtual hosts and more than 2200 distinct service endpoints distributed across two data centres existed in 2017, many producing and consuming RESTful APIs.



Our EBI Search as a service provides search functionality to eleven distinct portals, including Metagenomics, Enzyme Portal, Complex Portal, RNACentral and ENA. The main advantages are the use of a common search syntax across our data resources, as well as uniform search results navigation for the user. The EBI Search engine indexed over 1.8 billion data records from EMBL-EBI data resources during 2017.

A strong focus in 2017 was user experience design (UXD). Our Web Development team supported EMBL-EBI services and collaborators in a wide range of projects including the Human Cell Atlas, Open Targets, the Pistoia Alliance and IMPC, resulting in making life science software and tools more intuitive and easier to use.

Future plans

The Technical Services Cluster (TSC) continues to plan for a “hybrid” future where we complement the use of our leased data centre space with commercial clouds. Such a strategy will provide the most cost-effective delivery of our service to our users. It allows us to make use of hardware located in our data centres or use cloud resources when the offered service is aligned with our operational needs. The EMBL-EBI Embassy Cloud will continue to be developed to provide access to our own staff and collaborators to a cloud environment.

In 2018, we will see the evolution of our operational and service management procedures to be cloud-ready, while also undertaking a proof-of-concept use of such a cloud service. We are collaborating with European initiatives (including the European Open Science Cloud and the Helix Nebula Initiative) in addition to global cloud providers. While planning for a “cloudy” future, our investment in our data centre infrastructure continues, with a move to a new data centre space that will allow us to double the internal networking bandwidth between our three data centres to 40Gb/s. This way, we can better support the large data transfers that are part of our routine operations.

The investments we have made in building an internal software development, integration and operations team over the last few years are now providing results. A dedicated team will be established to drive this



Steven Newhouse

Head of Technical Services.
Team Leader, Technology and
Science Integration

*European Grid Infrastructure (EGI.eu),
2008-2013. Microsoft, 2007-2008.
Open Middleware Infrastructure
Institute UK, 2005-2007. Member
and Chair, Open Grid Forum Board of
Directors, 2008-2012. Sun Lecturer
in e-Science, Imperial College
London Department of Computing,
2003-2005. Imperial College London
e-Science Centre, 1989-2005.
Head of Technical Services at
EMBL-EBI since 2013.*

activity, consolidating much internal and external software development within the cluster and supporting engagement with external projects such as the Human Cell Atlas.

The upcoming introduction of the General Data Protection Regulation (GDPR) has resulted in an EMBL internal policy on data protection set to be implemented in 2018. This will lead to changes across EMBL-EBI's services – including the main website, which will be supported by the TSC cluster. We will be supporting the ongoing work to harmonise EMBL's web interface to deliver a common user experience.

We are also considering the next evolution of the hosting environment for our data and services to provide increased self-service management and to be aligned with an industry “DevOps” model. Already deployable across two data centres to provide resilience through high availability, we plan to bring into production a container-based model for hosting services that could be deployed across both our own data centres and into external cloud infrastructures for both disaster recovery and for geolocation to improve user experience.

(Left) Over 60 researchers working in computational biology gathered at EMBL-EBI to take part in the first coding challenge organised by the Human Cell Atlas.

Technical Services

Team achievements



Steven Newhouse

Technology & Science Integration team

- ⦿ Release of the Authentication and Authorization Profile (AAP) Service, which aggregates user identities and manages access to EMBL-EBI's different data archives as part of the Unified Submission Interface project
- ⦿ Development of the EBI Cloud Portal to support the EMBL Identity Provider to enable access by EMBL staff to internal and external cloud resources
- ⦿ Adoption of the AAP Service to manage permissions that give users access to the various software packages available within the EBI Cloud Portal
- ⦿ Continued use and engagement with commercial cloud providers within the Helix Nebula Science Cloud project and directly with commercial cloud providers
- ⦿ Exploring how scientific workflows, such as those from marine metagenomics, can be ported to run efficiently on cloud-based resources.
- ⦿ As part of our engagement with the ELIXIR Compute Platform, defining, developing and testing the Reference Data Set Distribution Service for maintaining copies of data set on remote clouds and clusters
- ⦿ Developed the Resource Usage Portal, which collects the usage by individuals, teams and services across EMBL-EBI of the clusters, clouds, storage, virtual machines and databases provided by the Technical Services teams
- ⦿ Continued development and operation of the FIRE archiving service, currently at 15PB and growing at an average 500TB a month
- ⦿ Prototyping a new user interface for EBI Tools that simplifies the interface for new users and allows tools to be easily chained together to build workflows



Jonathan Hickford

Web Development team

- ⦿ Provided User Experience expertise (user research, prototypes of the data ingest service and evaluations of the metadata definitions) to the Human Cell Atlas (HCA) project
- ⦿ Provided UX and web development services for the Open Targets platform introducing new milestone features for batch search and data visualisation for non-geneticists
- ⦿ User research, design and development activities in conjunction with the Mouse Bioinformatics team to develop the wireframes and themes for the PDX Finder project
- ⦿ Contribution to the UX for Life Science Toolkit, a Pistoia Alliance project consisting of patterns, practices, templates, guides and case studies to apply UX techniques in life science projects
- ⦿ Produced new websites and applications for the IMEx Consortium, CABANA Latin American capacity building and BioExcel projects
- ⦿ Provided UX and web development services for the EMBL-EBI Universal Submissions Interface and Tools Workspace projects
- ⦿ Continued to drive adoption and improvements of the EMBL-EBI Visual Framework, the common styles and templates used by EMBL-EBI web resources
- ⦿ Provided User Experience Research consultation to teams and groups across the institute



Rodrigo Lopez

Web Production team

- ⦿ Completed the consolidation of services involving 1000 Web virtual machines running some 2200 service endpoints. This resulted in improved service reliability and failover between our London and Hinxton data centres
- ⦿ Introduction of additional automation to improve ticket response times and focus on edgecases. The aim is to empower users so they have more control of the infrastructure on which their services run
- ⦿ Increasing number of teams adopted the EBI Search RESTful API, integrating search into their systems and customising views of the search results in novel and more informative ways
- ⦿ Steady usage of the Job Dispatcher framework. More than 40 000 sequence libraries are available to search over the RESTful and SOAP APIs. Usage is from around the globe but importantly, from collaborators such as UniProt, Ensembl Genomes, ENA, etc.
- ⦿ Adoption of Elastic Stack technologies to integrate data from the various service teams in relation to web and download traffic, allowing us to inform our teams about real-time usage of compute resources



Petteri Jokinen

Systems Infrastructure team

- ⦿ Migrated 13PB of data from old to new storage systems
- ⦿ Increased capacity of object store used by the main Sequence Archive system by 10PB
- ⦿ Introduced Foreman to reduce external dependencies and provide improved overall machine management
- ⦿ Introduced Singularity as an alternative containerisation technology for use on the internal clusters
- ⦿ Completed an optimisation project to allow users to run more concurrent jobs
- ⦿ Significantly improved network reliability and performance across all data centres
- ⦿ Doubled core network capacity enabling growth for the next 18 months



Andy Cafferkey

Systems Applications team

- ⦿ Optimisation and tuning of Ensembl's newly migrated large MySQL databases
- ⦿ Doubled the number of tenants in Embassy Cloud to 60
- ⦿ Working with the cross-TSC FitSM workgroup, the team implemented changes to the helpdesk software to support a FitSM workflow
- ⦿ Conclusion of the Oracle migration project, which established a path to migrate suitable EMBL-EBI databases from Oracle to Open Source database managements systems
- ⦿ New VPN clients on all client devices, introducing software tokens (LinOTP) for authentication, which replaced the hardware RSA token keys

Technology & Science Integration

The continued growth of the stored data and diverse analyses taking place at EMBL-EBI provides an ongoing need to assess and deploy new technology and services that support the scientific community. The Technology and Science Integration (TSI) team builds strong collaborations with technology innovators and service providers to help shape and adopt new technologies within the institute. As a result of this growth, a new group – Software Development and Operations – is now being established from within the TSI group, to focus on these collaborative software activities.

Major achievements

Software and services

The bulk of TSI's activities now focus on the development of software to meet our internal or collaborators' use cases, and the operation of the resulting software for the relevant internal or external user groups.

In 2017, we released the Authentication and Authorization Profile (AAP) Service that aggregates identities (primarily from ELIXIR) and manages access to EMBL-EBI's different data archives as part of the Unified Submission Interface project. The EBI Cloud Portal has adopted the AAP Service to manage permissions that give users access to the various software packages available within the Portal, and to control onto which clouds the user can deploy the software.

EMBL-EBI offers web and programmatic access by the public to hundreds of data analysis tools for our data resources. An internal project has explored how the user interface to the EBI Tools portfolio can be improved – especially for inexperienced users. We have already seen how the prototype is demonstrating improvement in the flexibility and usability of this analysis environment. This new environment includes a simplified interface to the tools, the ability to guide the user to other popular tools following the user's initial analysis, the ability for a user to upload data and download results, and for the user to specify workflows linking our tools together.

The TSI team also manage the development and operation on many internal projects. The File Replication (FIRE) Service helps ingest data from our major data archives to ensure that the data is reliably placed in our distributed object store and long-term tape archive to provide a resilient storage infrastructure encompassing over 50PB of storage.

The Resource Usage Portal is helping to account for the activity of different EMBL-EBI teams across our IT resources encompassing storage, compute, virtual machines, etc. This portal will help plan our IT provision and align it with our strategic priorities.

Multi-cloud analysis environment

Consistently executing a workflow on data distributed across different sites is becoming a reality as organisations explore the infrastructure needed to support personalised medicine. Organisations such as the Global Alliance for Genomics and Health (GA4GH) are establishing the standards for building such analysis infrastructure. We are collaborating with the GA4GH's Cloud WorkStream to contribute to and implement the specifications required to build such an analysis environment. An implementation of the GA4GH's Task Execution Service specification is being developed through an ELIXIR Implementation Study that will be deployed across multiple clouds in ELIXIR in 2018.

Hybrid cloud strategy

Over the last decade EMBL-EBI has had a strategy of using leased data centre space to house our hardware. This has provided a professional hosting environment for our public services that has improved our service uptime and reliability. Our strategy for the next decade is to make use of internal and leased data centre space alongside the use of external cloud providers that allows us to maximise our operational responsiveness while delivering the best value to our funders.

To achieve this long-term strategic aim we are exploring how to deliver this hybrid cloud strategy. We are considering how to make our scientific workloads



Steven Newhouse

Head of Technical Services.
Team Leader, Technology and Science Integration

European Grid Infrastructure (EGI.eu), 2008-2013. Microsoft, 2007-2008. Open Middleware Infrastructure Institute UK, 2005-2007. Member and Chair, Open Grid Forum Board of Directors, 2008-2012. Sun Lecturer in e-Science, Imperial College London Department of Computing, 2003-2005. Imperial College London e-Science Centre, 1989-2005. Head of Technical Services at EMBL-EBI since 2013.

portable across our internal clusters and our internal and external cloud providers. Our public services are hosted on our internal VMware infrastructure. In the future, we want to use a platform that will help us deploy our public services to internal and external cloud resources to give us greater operational flexibility for running geo-dispersed services and disaster recovery scenarios.

We are collaborating with the major global, European and national cloud providers to deliver the services that we need. This includes exploring with other large science labs, through the Helix Nebula Science Cloud, how we procure cloud resources, and engaging with the European Open Science Cloud to see how the next generation of e-Infrastructure activities can be established to meet the needs of the life-science community.

Web Development

The Web Development team designs, develops and maintains the internal and external websites relating to both EMBL-EBI’s core activities and affiliated websites. The team is a central consultancy for web development and User Experience Design (UXD) at the institute.

Our team maintains the global EMBL-EBI website, its underlying content database, the EMBL-EBI Intranet and training portals. We help develop and support several ancillary web portals and services, including projects such as CABANA, BioMedBridges, the International Nucleotide Sequence Database Collaboration and HUGO Gene Nomenclature Committee. We also provide front-end user experience (UX), web design and development for the Open Targets platform. Our team supports web developers throughout EMBL-EBI by providing web guidelines, templates, style sheets and training, as well as support in Drupal, JavaScript and other key web technologies. We also offer considerable expertise in UXD, an area of strategic importance for EMBL-EBI services.

Major achievements

Human Cell Atlas

The Web Development team has supported the Human Cell Atlas (HCA) project since July 2017. Specific activities have been interviews with stakeholders and the research project participants, designing and producing a HCA submissions prototype, and contributing UX recommendations through retrospectives and discussions. This work was presented at the autumn meeting in Palo Alto, raising awareness of the UX activities for the upcoming quarters.

Following this meeting, focus has been on UX user research and evaluation of metadata activities with the content teams, including evaluation of the metadata spreadsheet with potential HCA project members. Design work with the infrastructure team has continued, integrating the prototype for the Data Coordination Platform data ingest service.

Open Targets

User experience design and web development has focused on researching, designing and building new features for the Open Targets platform.

Milestones this year include:

- ⦿ *Designing and releasing a “batch search” functionality, which allows users to search the platform for a list of genes and inspect the diseases, pathways and drugs associated with them*
- ⦿ *Better supporting safety experts by improving the way we display protein homology and mouse phenotypes for their safety reviews*
- ⦿ *Visualising the baseline expression of a target more clearly and using the same data to filter the targets associated with a disease by tissue expression*
- ⦿ *Exploring ways to present genetics and functional genomics evidence to non-geneticists in an intuitive way*

Industry Programme and Pistoia Alliance

Our team has contributed to a Pistoia Alliance project called the “UX for Life Science Toolkit”, a set of case studies, guides and templates for common UX processes, tailored for usage in life sciences. We provided samples of our template and documentation, and helped curate these into best practice examples.



The Pistoia Alliance organised several user experience (UX) workshops, which will result in the launch of a UX Toolkit for the life sciences.

PDX Finder

We supported the mouse informatics team through UX, design and web development consultancy and capacity for the PDX Finder project. We conducted end-user research, designed product wireframes and engaged an external graphic designer. We are developing a web-based visual framework, derived from the EMBL-EBI Visual Framework for this project and will supply this for integration with the scientific service. We will also be developing and hosting a WordPress instance for the content aspects of the service.

Universal Submissions Interface

In 2017 Web Development iterated on the Universal Submissions Interface (USI) prototypes, incorporating continuous feedback from the community and curators by holding usability sessions. Through this feedback, the prototype has moved from paper versions through to a web-based prototype built on the EMBL-EBI Visual Framework. The Web Development team has provided development capacity, and is starting to build the front-end interface based on the wireframes and back-end APIs in conjunction with the USI team.

EMBL-EBI Visual Framework

We released two new updates for the Visual Framework. Based on input from the Web Guidelines Committee, these updates improved the presentation of data and simplified how teams use the framework in the software stack of their choice.

Future plans

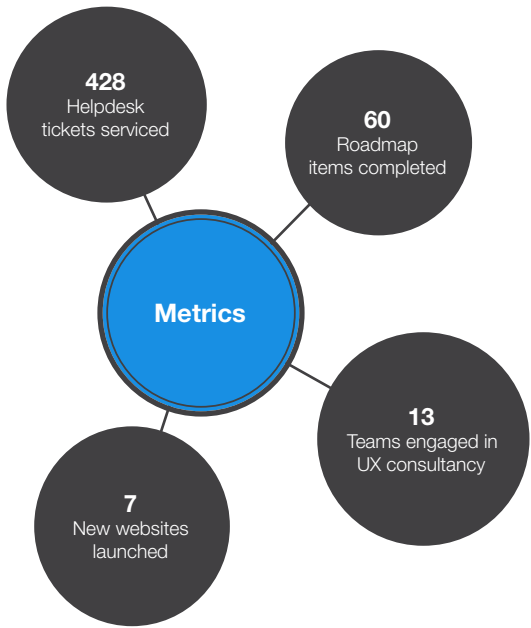
We plan continued improvements to our content database, and internal websites so that we can provide a single source of information on events, people, tools and more to be used on sites across EMBL-EBI and other EMBL sites. In particular, we look to help EMBL-EBI meet best practices outlined under the General Data Protection Regulation (GDPR), and to empower more EMBL-EBI employees to be able to self service for common tasks. We also aim to expand our collaborations with colleagues at other EMBL sites, in both the technical and content teams.



Jonathan Hickford

Team Leader, Web Development
*MSci Physics, University of Bristol 2006.
Team Leader at EMBL-EBI since 2016.*

Metrics



Selected publications

Karamanis N, et al. (2018). Designing an intuitive web application for drug discovery scientists. *Drug Discovery Today*. doi: 10.1016/j.drudis.2018.01.032

Cham JA, Costa K (2017). UX Design: Maximising the value of scientific software in life science R&D. *Drug Discovery World*. doi: N/A

Koscielny G, et al. (2017). Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*. doi: 10.1093/nar/gkw1055

Web Production

The Web Production team (WP) provides the EMBL-EBI search engine, tools under the Job Dispatcher (JD) Framework and web server administration. It comprises a mix of full stack software and system engineers that provide access to core services and support to all the teams and groups at EMBL-EBI.

Our software engineers are specialists in Java, Perl and Python and specialise in web services technologies for search and bioinformatics tools in local and cloud environments. The system engineers, or Web Administrators, specialise in deployment and management of web application servers, web server custom configuration, and web traffic management for the deployment of EMBL-EBI's public services.

Major achievements

The web visibility of EMBL-EBI data resources and tools depends largely on work done by our team. Approximately 1000 virtual hosts and more than 2200 distinct service endpoints exist in 2017, many producing and consuming RESTful APIs.

In 2017, EBI Search as a service provided search functionality to eleven distinct portals, including EBI Metagenomics, Enzyme Portal, Complex Portal, RNACentral and ENA. The main advantages are using a common search syntax across services, as well as uniform search results navigation for users. The EBI Search engine indexed over 1.8 billion data records from EMBL-EBI data resources during 2017.

The Job Dispatcher framework, through which users run NCBI Blast+, HMMER3, InterProScan and Clustal Omega services, handled more than 120 million job requests (compared to 130 million jobs in 2016). The decrease is due to a reduction of traffic from the Asia Pacific region and the installation of an instance of the JD dispatcher in China. The EBI Search and Job Dispatcher tools are amongst EMBL-EBI's top ten most used services. A new service method developed during 2017 is the adoption of Common Workflow Language service descriptions that enable users to design and deploy analytical workflows that can easily be built into third party environments.

Other team projects in 2017 included glue projects such as EBI Tools, UK Cloud Pilot, HMMER3, as well as THOR developments that allow users to claim datasets via EBI Search using their ORCIDs and participation in ELIXIR Exceleerate, Ensembl migration from the Wellcome Sanger Institute, etc.

Increased usage

The top five most used services run via the Job Dispatcher framework are InterProScan, NCBI Blast+, Infernal_Scan, Clustal Omega and Needle, a popular pairwise sequence alignment tool.

Sequence libraries for sequence similarity searches are updated on a daily basis. Sequence similarity searches on up-to-date libraries are essential to our users, especially data curators, biotech labs working on the identification of active biomedical compounds in newly characterised species, and patent examiners. The number of sequence libraries served via the sequence similarity services is 47 000. These include the ENA, UniProt and Ensembl Genomes, with the latter contributing to most.

The Elastic Search based resource usage reporting framework we announced in 2015 has become the main reporting system for web and download usage. We have further enhanced and enriched the reports to make it possible to provide up-to-date data to management committees.

Outreach, training and support

The team is involved in many internal training activities focused on teaching users to develop and use web services or integrate these into their own workflows and pipelines. We participated in 22 distinct training and outreach activities and produced two webinars during 2017. Our helpdesks have a constant stream of requests from over the world regarding the use of tools and provide solutions to resolve queries.

Future plans

In 2018 we will focus on reducing the number of tickets and ticket time. This implies empowering users to do more themselves, by giving them what they need, while we focus on edge cases. There is no doubt that the continuous growth of staff and data at EMBL-EBI will present challenges. The use of container-based technology for deploying web services is high on the agenda.



Rodrigo Lopez

Team Leader, Web Production

MSc Veterinary Medicine, Oslo Veterinaerhoyskole, 1984. MSc Molecular Biology and Toxicology and Informatics, University of Oslo, 1987. At EMBL-EBI since 1995.

Selected publications

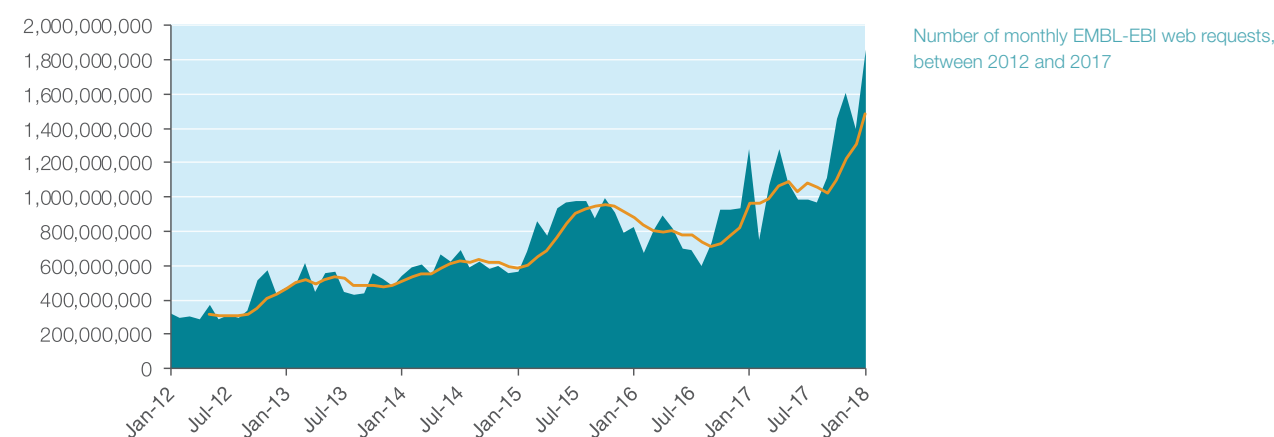
Perez-Riverol Y, et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology*. doi: 10.1038/nbt.3790

Park YM, et al. (2017). The EBI search engine: EBI search as a service-making biological data accessible for all. *Nucleic Acids Research*. doi: 10.1093/nar/gkx359

Chojnacki S, et al. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research*. doi: 10.1093/nar/gkx273

Pearson WR, et al. (2017). Query-seeded iterative sequence similarity searching improves selectivity 5-20-fold. *Nucleic Acids Research*. doi: 10.1093/nar/gkw1207

EMBL-EBI web requests per month 2012 - 2017



Systems Infrastructure

The Systems Infrastructure Team manages EMBL-EBI's compute servers, storage, data centres, networking and Internet connection. The team works closely with all project groups, maintaining and planning their specific infrastructures, and plays a key role in managing the technical frameworks supported by the UK Government's Large Facilities Capital Fund.



EMBL-EBI data centre in Hinxton, United Kingdom

Major achievements

Storage

2017 was the year of data migrations, taking up a significant part of the storage team's time. We replaced our aging scale-out NAS estate with a new solution. During the year we migrated 13PB of data. We introduced another object storage vendor into the Sequence Archive system, increasing its capacity by 10PB.

These two programmes illustrate how we seek the best of breed solutions to help EMBL-EBI provide biological data to the global life science community. Our work enables us to get the most robust, performant and cost-effective solution available during our procurement processes.

Compute

We have introduced Foreman as a provisioning system and a central aggregation of information for the Katello and Ansible tools. Moving to a modern and community-supported tool, like Foreman, allows us to have better control over the installation of the OS and a better definition of the infrastructure and the roles played by each machine we manage. The distributed nature of Foreman also benefits our multi-location infrastructure. Several functionalities can be offloaded to external proxies running locally in the different data centres and reducing dependencies from the WAN. We plan to move away from our current tool in favour of Foreman during 2018.

We started providing content updates (RPMs) through the lifecycle management software Katello. We can now apply bug/security fixes and updates in a controlled process where the machines are first updated/tested in an isolated environment and then put in production when the new content is known to work correctly. Katello is part of Foreman and is fully integrated with it.

We also introduced an alternative containerisation technology, Singularity. We are promoting Singularity as the primary containerisation technology in our clusters, but we will continue supporting Docker because it is not currently clear which technology will be better suited for our life-science application environments.

We continue to visualise the resource usage by providing improved and

simplified web-based cluster usage statistics for users. We also developed and deployed an in-house web-based traffic light system in order to give a high-level resource summary of all of the cluster attached storage.

The team's continual improvement process cycle on EMBL-EBI's various compute clusters has proven beneficial again, for example with the installation of LSF 10 to all clusters servicing our web tools. Another example is the successful completion of an optimisation project aimed at memory usage within our clusters. This has made a massively positive impact on EMBL-EBI's cluster resources so that users are now able to run more concurrent jobs.

Networking

During the first quarter of 2017 we switched over 200 10G ports in the EMBL-EBI Hemel Hempstead data centre, with only minimal disruption to operations. Following the completion of this project, we observed significant improvements in performance and reliability.

We also made capacity improvements to switching in the Hinxton data centre. We doubled core capacity, providing growth for around 18 months into the future, with capability for 100Gb/s core-edge links when this becomes necessary.

We have now designed and built a network access control system, which is being rolled out across the institute in 2018. This improves the security of EMBL-EBI's desktop network; in particular it provides a measure of protection against threats from malware on staff and visitor personal devices.

Future plans

We will keep supporting the data growth to the best of our abilities and continue to migrate the approximately 6PB of remaining data (research, external services and general purpose data) into the newly purchased systems.

We will finally retire all the remaining legacy NAS systems, including the PDBe staging area, which are in total approximately 500TB.

We also aim to finish the implementation of the redesigned storage and network architecture for Hinxton and Hemel Hempstead web production, reducing the load on our firewalls and potentially improving IO performance using direct non-routable network connections.

On the back of the use of LSF 10 within the web-production clusters, the team will roll out this



Petteri Jokinen

Team Leader,
Systems & Networking

*MSc in Computer Science 1990,
Helsinki University. At EMBL-EBI since
1996.*

version to all remaining compute clusters.

We aim to complete the full implementation of Foreman for builds instead of ABACUS, Katello for all of our cluster life cycle management and Ansible instead of Puppet on all of our clusters.

During 2018, we will increase the capacity of EMBL-EBI's network links between Hinxton and Hemel Hempstead from 20Gb/s to 40Gb/s.

A major project for 2018 is to replace a large proportion of the Hinxton data centre switching. A significant side effect of this is to bring switching across the infrastructure into relative homogeneity. This allows the use of very sophisticated telemetry across the estate, improving visibility, time-to-fix and performance monitoring.

EMBL-EBI will migrate all equipment out of our Flint Cross data replication site during the first half of 2018, resulting in significant improvements in network bandwidth available for object storage operations backing the sequence archives.

We will roll out network access control over the entire desktop access network. In addition to this, we will unify access controls across wireless, wired and VPN connections, making staff and visitor experience more uniform.

Systems Applications

The Systems Applications team manages the higher layers of EMBL-EBI's IT infrastructure, which includes virtual machine (VM) management, virtualisation and private clouds, database services, desktop systems audiovisual and telephones. The team supports EMBL-EBI staff in their daily computer-based activities and works closely with service teams and research groups.

Major achievements

Virtualisation and cloud

Our team added an S3 object store interface to Embassy Cloud. Tenants are using this as both a new primary storage tier and as a backup facility for their data and configurations. The S3 object store is sized to hold the Pan Prostate Cancer project primary data expected in 2018.

We have also implemented a standardised method to provide secure access to internal databases from Embassy Cloud, which increases the internal EMBL-EBI resources available to tenants, and is an efficient example of the benefits Embassy can provide to tenants. This has allowed us to increase the number of tenants from 30 to 60 without a substantial increase in the current support staff.

We believe that the main blocker to greater cloud adoption is the lack of skills and experience within the user community to modify their workload for cloud deployment. To assist users, the Systems Applications team has been working with service and research teams across EMBL-EBI on their Embassy tenancies, offering practical advice and support in person and through the ticketing system. In this way, we have seen notable successes with PRIDE, EBI Metagenomics and PhenoMeNal.

Within the eMedLab Consortium, EMBL-EBI was actively involved in the system redesign and migration to a newer Openstack version. Our experience with the Embassy Cloud has allowed us to provide guidance in this project, and the result is a much simpler set up, where usability and access have been improved to make a better user experience.

We have continued to contribute to the HelixNebula Science Cloud project by participating in the writing, presentation and evaluation of the tender bids and now working with the winning bidders to ensure their designs match the requirements of the three EMBL use cases – PanCancer, Image Data Repository and ELIXIR.

Databases

Continuous improvement activities are regularly conducted each year by the Database Team to guarantee a stable and state-of-the-art large database back-end. In 2017 we conducted operating system upgrades across all database platforms to RHEL 7 and Oracle Linux 7 where possible, or to the highest Linux version 6 compliant to the DBMS version hosted.

Our team also conducted major projects during 2017 in specific areas including Delphix, MySQL, MongoDB, and Oracle migrations.

We conducted Delphix upgrades of the appliances versions, upgrades and technology refresh of the storage back-end and, of special note, the extension of Delphix functionality to PostgreSQL, equivalent to the functionality available for Oracle at EMBL-EBI.

Following the Ensembl data resource migration from the Wellcome Sanger Institute to EMBL-EBI, we worked closely with all the Ensembl teams to tune the configuration of their MySQL instances and the VM specs in order to achieve optimal running of their production and web pipelines.

We have also worked on a new MongoDB facility. The new system has three main advantages over the previous one: higher storage specs, better workload isolation (separating each project in a dedicated virtual cluster), and simplified multi-tenancy.

Finally, in July 2017 the Oracle Migration project was successfully concluded. Seeking to migrate suitable EMBL-EBI databases from Oracle to alternative suitable open-source database management systems, we managed to migrate 38 instances and achieve our target. As a result we acquired extensive technical knowledge that, together with the migration method we have developed, will provide confidence for any possible future migrations. The Oracle migration activity is now part of the business-as-usual service offered by the Database Team and is available for any team interested.

Desktop support

In 2017, the Desktop team implemented a new process of handling support queries; each day a member of the team monitors all tickets in the desktop queue and triages all tickets, applying appropriate tags and immediately responding to critical tickets.

As well as providing desktop support to EMBL-EBI users, the team has successfully delivered many IT



Andy Cafferkey

Team Leader, Systems Applications

Senior Systems Administrator, Cambridge Positioning Systems, 2000-2003. Pi Group, Ford Motor Company, 2003-2005. At EMBL-EBI since 2005, Technical Team Leader since 2015.

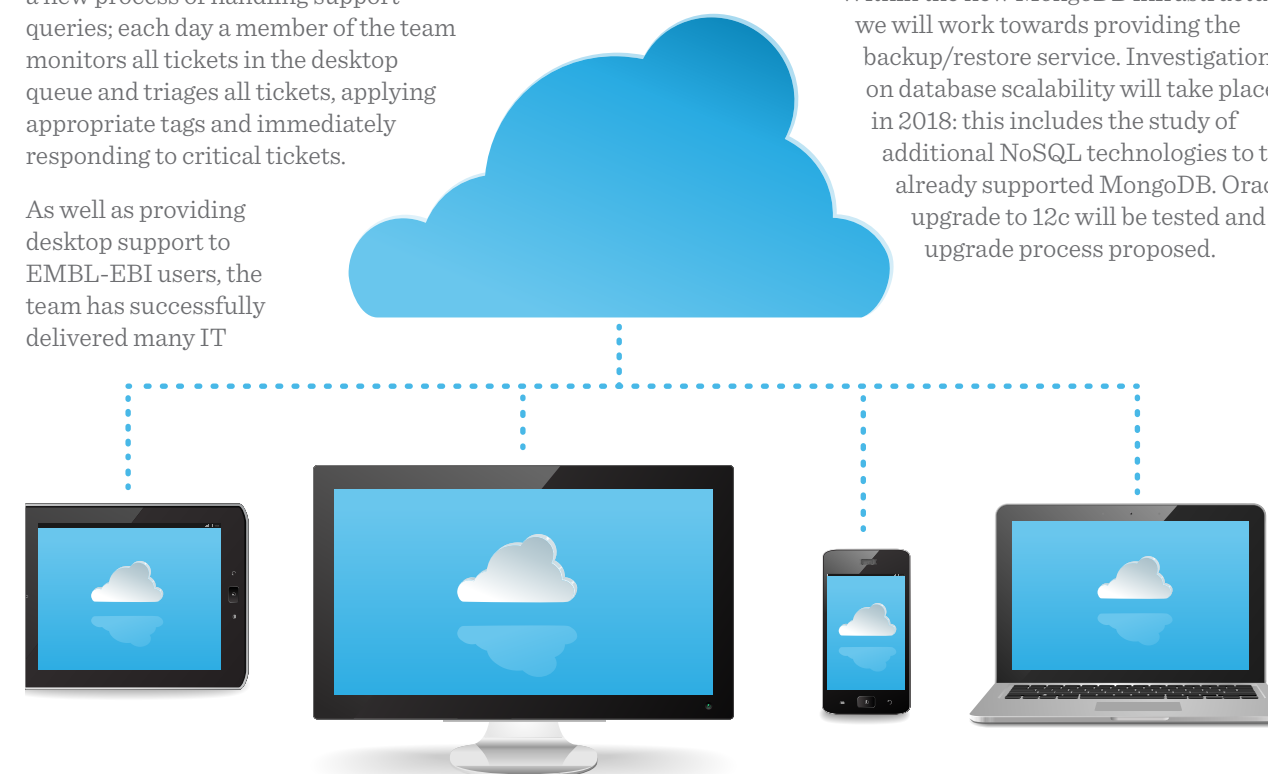
projects, including deployment of new macOSX (High Sierra), site-wide operation to protect all Windows based client machines from WannaCry ransomware, roll out of new VPN client on all devices, and installation of new printers in both EMBL-EBI buildings.

Future plans

In 2018 the team will work with EMBL-EBI research groups on the Pan Prostate Cancer project analysis, which will be conducted in Embassy Cloud. Additionally, we will continue to assist tenants in their use of Embassy Cloud to help close the skills gap we see as a barrier to greater cloud adoption.

The audiovisual infrastructure in EMBL-EBI will be replaced and extended to include soft video conferencing facilities in meeting rooms to support Skype, Lync, GoToMeeting and other software VC solutions.

Within the new MongoDB infrastructure, we will work towards providing the backup/restore service. Investigations on database scalability will take place in 2018: this includes the study of additional NoSQL technologies to the already supported MongoDB. Oracle upgrade to 12c will be tested and an upgrade process proposed.



Administration and External Relations

Schwann cells

Schwann cells are a type of cell in the nervous system that enable the other cells to function optimally. They come in two types. Myelinating Schwann cells wrap around the axon of the neuron, forming insulation and accelerating the speed of nerve impulses. Non-myelinating Schwann cells also wrap around the axon, providing nutrients and ensuring its well-being.

Administration

The EMBL-EBI Administration team facilitates the work of the institute by contributing to the EMBL-wide implementation of efficient administrative processes, which enable the effective deployment and development of resources within a complex regulatory environment.

Major achievements

Strategic Project Management Office

The Strategic Project Management Office (SPMO) team offers essential support for EMBL-EBI’s capital investment programmes and wide change initiatives.

Highlights in 2017 included the UK Cabinet Office review of EMBL-EBI’s Large Facilities Capital Fund programme, looking into the operational readiness and the benefit realisation of the investment. Finding that the programme had been successfully delivered and had already secured many of the promised benefits, the assessors awarded us a very rare ‘Green’ assessment rating. The assessors also congratulated the project team, in which the SPMO members played a central part.

The office is currently developing the onward long-term capital investment case including the emerging central EMBL-EBI Bioimaging Archiving initiative. Our team has provided support for imaging community engagement, internal strategy development and technology investment.

SPMO has assisted in the development and use of pre-procurement guidance to support more transparent compliance to EMBL financial regulations for more demanding procurement exercises. We have also worked with the wider EMBL procurement teams to ensure coherent guidance across the organisation.

We have further developed the institute’s risk management process with a formal risk register and review process adopted during 2017. This has also paved the way for expanding the business continuity planning for EMBL-EBI’s activities as a whole. Work is ongoing and should be completed in 2018.

Finance

Together with the Research Office, we embarked on an overhaul of our travel and event management process. By introducing new software and a new Travel Management Company, we increased the efficiency of the progress and significantly reduced the costs.

We have successfully supported our diverse and increasing funding portfolio, which now incorporates new funders such as the Chan Zuckerberg Institute.

The Accounts and Purchasing Team has seen a steady increase in productivity, up by approximately 8.5% compared to 2016 in terms of the number of orders and invoices processed.

Human Resources

The flexible working arrangements for EMBL-EBI staff, introduced as a pilot scheme in 2015, have now been fully implemented. This includes provision for regular working from home arrangements and “keeping in touch” days for staff on maternity leave.

The Human Resources team managed a 10% increase in recruitment, a record level of recruitment activity, during 2017. This increase reflects both success in attracting external funds and fluctuations inherent to the built-in staff turnover scheme. The emerging “Cambridge biotech corridor”, combined with an improving economy, also made it measurably easier for our staff to find new jobs at the end of their contracts. We have piloted Agile recruitment exercises aiming to recruit the best talent to the Service Teams at institute, rather than team level.

Work is now ongoing on the introduction of a new, EMBL-wide, online recruitment tool and the EMBL-EBI HR team is fully involved in the scoping, design and testing of this tool.

Facilities Management and Health and Safety

We have been actively managing the increased demand for space within EMBL-EBI buildings, including relocating teams to make the best use of our facilities. The expansion of existing teams and the incorporation of incoming teams, has led to the design and rebuild of several areas. However, we have now reached the point where temporary accommodation has to be considered for onward growth.

We have recently replaced the multi-function print, scan and copy devices to achieve greater efficiencies and cost-savings. We are working with the Technical Services Desktop Team on an ambitious program to renew and improve our audiovisual facilities.

We have introduced a Health and Safety Checklist for use by Group and Team Leaders, and provided support to a huge range of institute and campus initiatives (e.g. implications for remote working under EMBL-EBI’s flexible working policy, and receiving training as campus develops its mental health policies).



Mark Green

Head of Administration

EMBL-EBI Administration Fellow of the Chartered Institute of Internal Auditors. At EMBL since 1997; joint appointment with EMBL-EBI from 1999 and fully at EMBL-EBI since 2003.

EMBL-EBI’s success is highly dependant on compute activities so workplace ergonomics are essential to ensure staff health and wellbeing. We are committed to maintaining a target of over 90% of personnel being assessed for workplace ergonomics.

Future plans

We are looking forward to the full implementation of EMBL-EBI’s Business Continuity plan and the reinvigoration of the Project Management Network.

External Relations

EMBL-EBI's strong relationships with funders, policymakers, key opinion formers and collaborators within and beyond Europe are founded on effective communication. The External Relations team develops and disseminates news about the institute's research and services through our website and social media channels. We also produce informative publications and raise the profile of bioinformatics in the broadcast and print media.

Major achievements

In 2017 we hosted many inbound visits to the Hinxton campus from a broad range of stakeholders from Government and industry. We hosted inbound scientific and ministerial delegations from Italy, Japan, Portugal, Finland, Singapore and China. Some of our most prominent visitors included the Ambassador of France to the United Kingdom, the Technology and Science Attaché to the Embassy of the Kingdom of the Netherlands and the Technology Attaché to the Austrian Embassy. We also hosted three visits by UK MPs and two visits by UK Government Ministers.

The External Relations team worked with colleagues from EMBL's Alumni Relations office to deliver the annual EMBL in the UK meeting at Lady Margaret Hall, Oxford in May 2017. This event provided an opportunity for UK-based alumni to present their current work and to network and exchange ideas with other alumni and their guests.

We supported our scientists and other staff in their efforts to make their science better understood at conferences and events across the world. In 2017 we assumed responsibility for coordinating EMBL-EBI's presence at key scientific conferences and exhibitions: ISMB/ECCB in Prague and Nature Jobs Fair in London.

After careful consideration and research, we decided our presence at these events should promote three major topics: career opportunities, the diverse range of services and tools, and our highly comprehensive training programme. By pulling resources from across different teams together, we managed to create a clear and complete picture of our institute.

More than 40 EMBL alumni and their guests gathered in the University of Oxford's Lady Margaret Hall for an afternoon of networking.



In terms of media relations, 2017 was a year of significant growth in the coverage of EMBL-EBI and the work of its scientists in print and broadcast media in 18 countries. Our team generated 270 press clippings across online, print and broadcast outlets.

We achieved significant reach with a general audience by appearing in high-profile national media such as *BBC Radio 4*, *Le Monde*, *The Atlantic*, *The Sunday Times* and *The Huffington Post*. Our press activity also targeted a scientific audience, and resulted in EMBL-EBI being featured in international specialist publications including *The Conversation*, *Scienza&tecnica*, *Naked Genetics*, *Genetic Engineering News* and *New Scientist*. For the first time, we focused on highlighting our infrastructure and technical capabilities with a more technical audience, and our efforts resulted in clippings on major technology websites including, *Motherboard*, *Tech Times*, *Wired* and *Mashable*.

In 2017 the External Relations team used its social media channels to grow and nurture the bioinformatics community and to further the reach of EMBL-EBI's missions. We ran effective campaigns to raise awareness of Open Access, women in STEM and the Science is Global initiative.

Through strategic and consistent social media activity, we achieved a 29% increase in our Twitter follower base, now standing at 28 000 followers – and a 23% increase in our Facebook fan base – now standing at 13 300.



Followers
28 000



Followers
13 300



Lindsey Crosswell

Head of External Relations

BA Hons, London University. BP plc, Government and Public Affairs Manager, 1995–2003. Head of External Relations, Chatham House, Royal Institute of International Affairs 2000–2003 (secondment), Director of Development, Oundle School 2004–2008. At EMBL-EBI since 2011.

Future plans

In 2018, we will add a videographer to our team, allowing us to focus more on video storytelling. We aim to produce a range of core video assets for the institute, to help us promote key areas, such as impact, recruitment, research, and to highlight EMBL-EBI's interdisciplinary, international and collaborative spirit.

We will also begin planning for EMBL-EBI's upcoming 25th anniversary in 2019, which we hope to celebrate with a symposium.



Facts and figures

A stylized illustration of neurons. The central neuron is large and detailed, with a star-shaped soma and numerous long, thin dendrites. Its axon is long and segmented, glowing with a bright blue light. Other neurons are visible in the background, some smaller and more distant. The background is dark and textured, with swirling patterns and a bright, glowing area in the center, suggesting a complex network of neural activity.

Neurons

Neurons carry electrical signals in the nervous system. They have a cell body called a soma and a long projection called an axon, which acts like an electrical wire. A human brain can have over 100 billion neurons.

Data growth and usage

What is a request?

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

What is a job?

A job is a program, e.g., BLAST, run by users to analyse and/or compare data from our data resources through sequence searching or by submitting their own data for comparison or manipulation using our computational tools.

Data usage

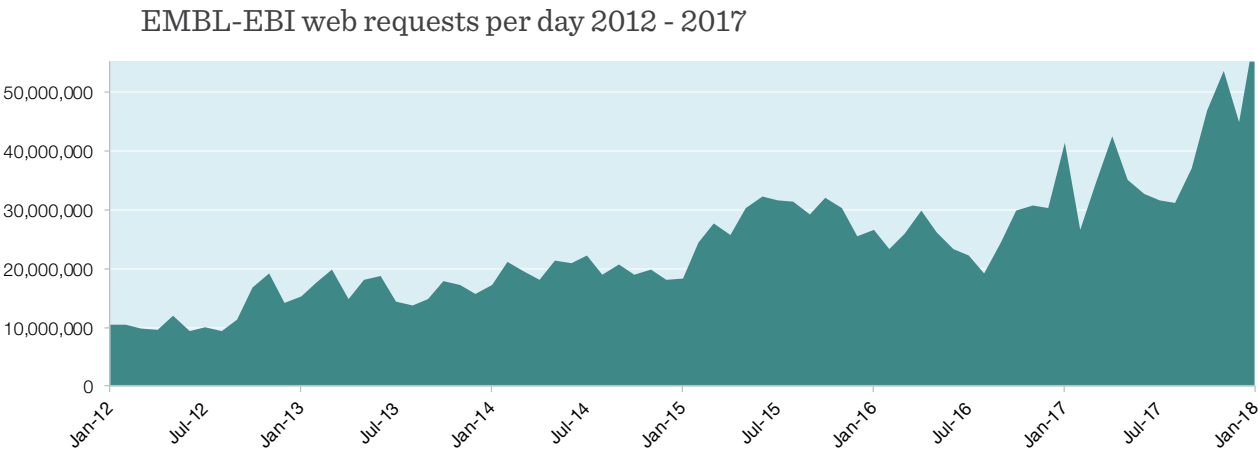
On an average day at the end of 2017 we saw just under 38 million requests to our websites. This is a significant increase from the 27 million requests per day in 2016 and shows the expanding usage of our data resources from year to year. Scientists all over the world are accessing our services, with the heaviest usage in 2017 coming from the USA (29%), China (17%), the UK(10%), France (8%) and Germany (6%).

During 2017, we ran 140 million jobs for our users working in both industry and academic settings. Most of this usage was via RESTful APIs (88%), with 8% from web interfaces and 4% from SOAP web services.

Extracting information on the number of users based on web logs is very difficult. Many users access our pages from multiple IP addresses and, conversely, it is not uncommon for one IP address to represent a whole organisation. However, from our records, we see that in 2017 there were on average 1.17 billion requests to our sites from 3.3 million unique hosts, each month.

Data growth by data resource

In 2017 our data resources continued their steady growth as the cost of generating data continued to fall. This has a dramatic impact on EMBL-EBI databases as it enables researchers to generate more data for submission to our archival data resources. To address the research community's data needs, EMBL-EBI continues to develop and implement innovative data-storage methods. The graph on the right illustrates the growth of our largest data resources.



Retired tools

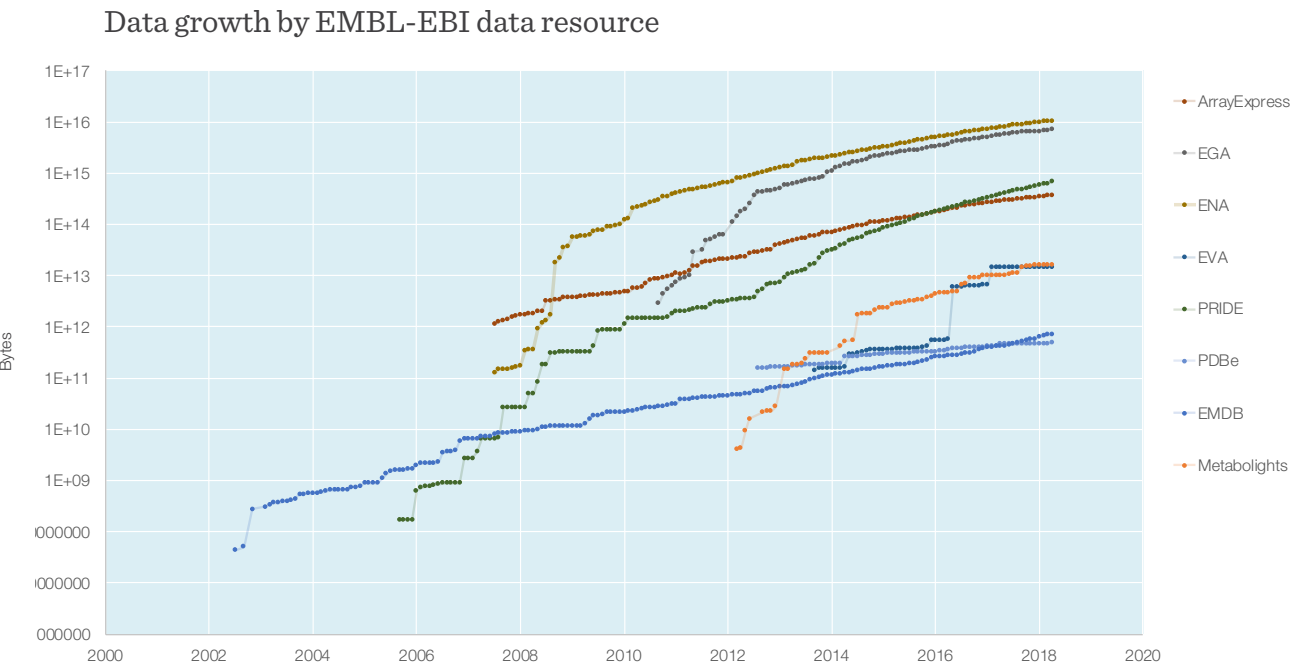
During 2017 we conducted a user survey to evaluate the usability and usefulness of our tools. As part of our regular usage review we identified some tools that, due to changes in technology and data usage, are seeing little or no usage, and will be retired. These include:

- ⦿ *promoterwise and wise2DBA*
- ⦿ *Small Molecule Search (SMS) in Reactome*
- ⦿ *ChEMBL NTD SMS*
- ⦿ *GPCR SARfari Blast search*
- ⦿ *GPCR SARfari SMS*
- ⦿ *Kinase SARfari BLAST search*
- ⦿ *Kinase SARfari SMS*.

Most popular tools

In 2017 the top 10 heaviest used sequence analysis tools were:

- ⦿ *iprscan5* (79.71%)
- ⦿ *BLAST+* (7.33%)
- ⦿ *Clustal Omega* (5.43%)
- ⦿ *Infernal Scan* (4.12%)
- ⦿ *Phobiu* (1.37%)
- ⦿ *pfamscan* (1%)
- ⦿ *muscle* (0.73%).



Growth of data resources at EMBL-EBI 2002-2017.

Data resources shown:
ArrayExpress, European Genome-Phenome Archive (EGA), European Nucleotide Archive (ENA), European Variation Archive (EVA), Proteomics Identifications (PRIDE), Protein Data Bank in Europe (PDBe), Electron Microscopy Data Bank (EMDb), and MetaboLights.

The y-axis is logarithmic and growth in all data types is exponential. Doubling times range from 12 to 24 months. Growth in ENA and EGA, which store nucleotide sequences, reflect continued improvements in sequencing technology, as well as lower costs for sequencing.

Of particular interest is growth in proteomics data shown in PRIDE, showing growth in non-sequencing data types for our service resources.

Data growth at a glance



Scientific collaborations

EMBL-EBI works with research communities throughout the world to establish standards, exchange information, improve methods for analysis and share best practice for the curation of complex biological information. Our highly collaborative research programme benefits from strong, productive partnerships with a large network of academic peers throughout the world.

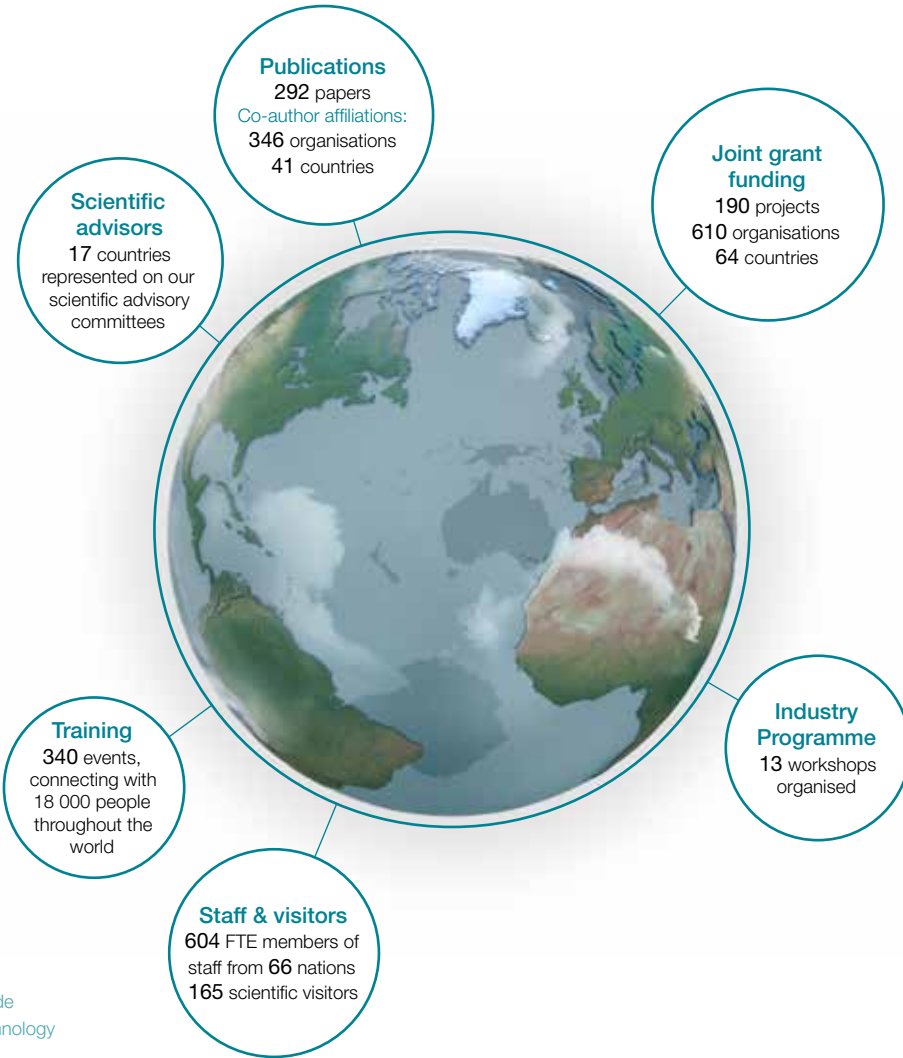
Joint grant funding

In 2017, EMBL-EBI shared joint grant funding with researchers and institutes in 64 countries throughout the world – most notably in the United Kingdom, Germany, Spain and France, but also with colleagues in countries with smaller research communities such as Colombia. Of the 190 grants received, 39 were awarded exclusively to EMBL-EBI.

These figures are potentially underestimated, as not every grant lists all partners involved.

Joint publications

Most of our 292 articles published online in 2017 were co-authored with colleagues at other institutes throughout the world, including other EMBL sites. Our most productive partnerships were with people at institutes in the United Kingdom, United States, Germany, Switzerland, Canada and France, and our collaborations extended well beyond Europe to Mexico, Brazil and South Korea.

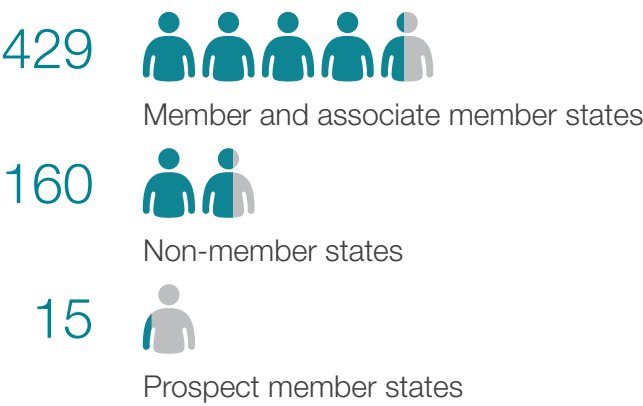


Our impact is global and our community is truly international. Many of our achievements are made possible by collaborations with science and technology professionals throughout the world.

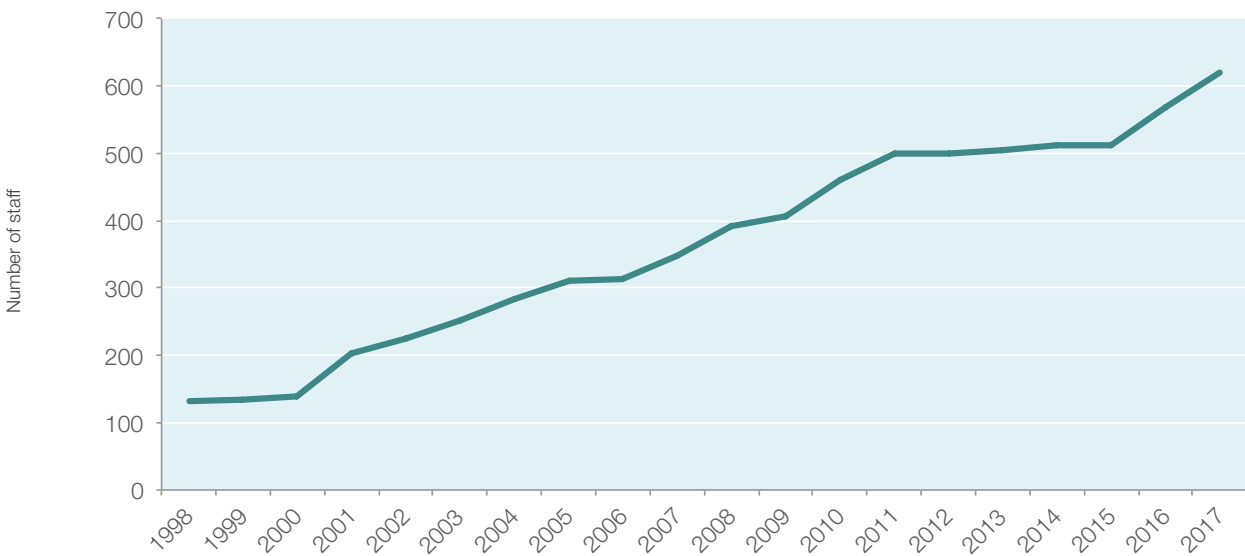
Staff and alumni

EMBL-EBI is proud to report that in 2017, its staff represented 66 nationalities (compared to 64 in 2016). There were 604 FTE* members of staff in 2017, including 17 FTE trainees. EMBL-EBI hosted 165 scientific visitors for longer than one month.

EMBL-EBI Personnel nationalities in 2017 in FTE*



Staff growth at EMBL-EBI, 1998-2017



*FTE: Full-time equivalent

New leadership

In 2017 we welcomed two new group leaders to support and guide the development of our public data resources.

Evangelia Petsalaki joined EMBL-EBI from the Lunenfeld-Tanenbaum Research Institute to lead research on deciphering human cell signalling to understand health and disease.

Zamin Iqbal joined us from the University of Oxford's Bioinformatics and Pathogen Genomics Group. In his new role, he leads the Computational Microbial Genomics Group working on genetic variation in microbes.

New research group leader **Virginie Uhlmann** and EMBL-EBI's new Head of Administration and Operations **Rachel Curran**, recruited in 2017, will begin their work in 2018.

Alumni profile

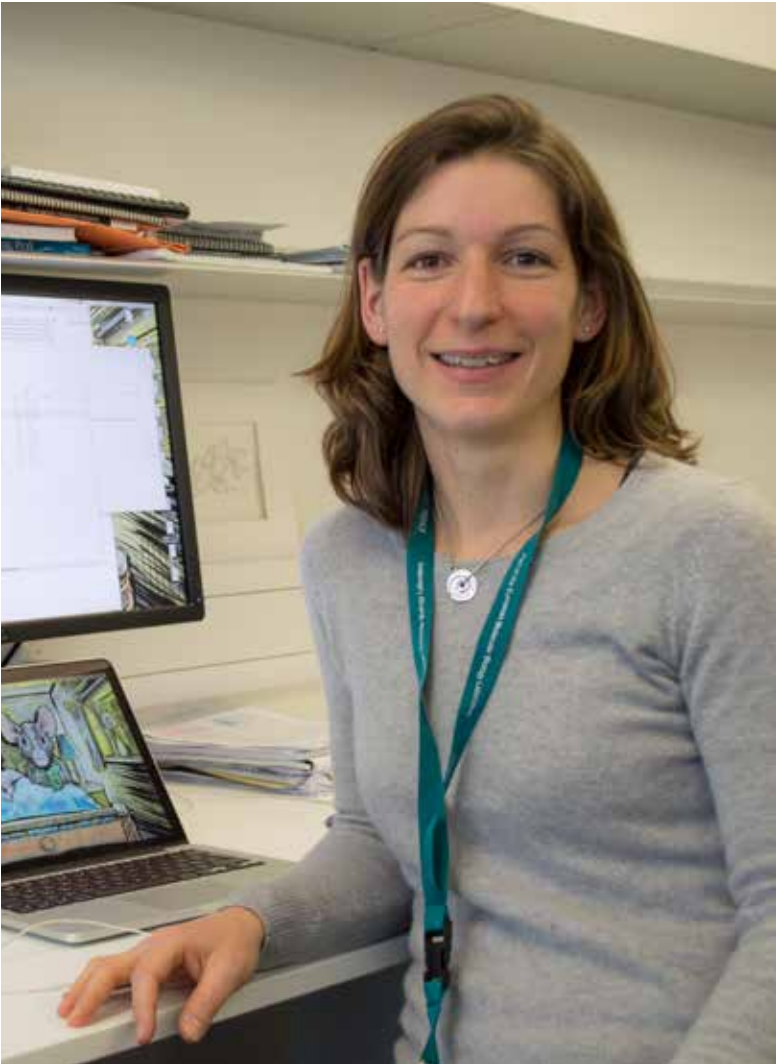
Amelie Baud, postdoctoral research fellow in Oliver Stegle's Group

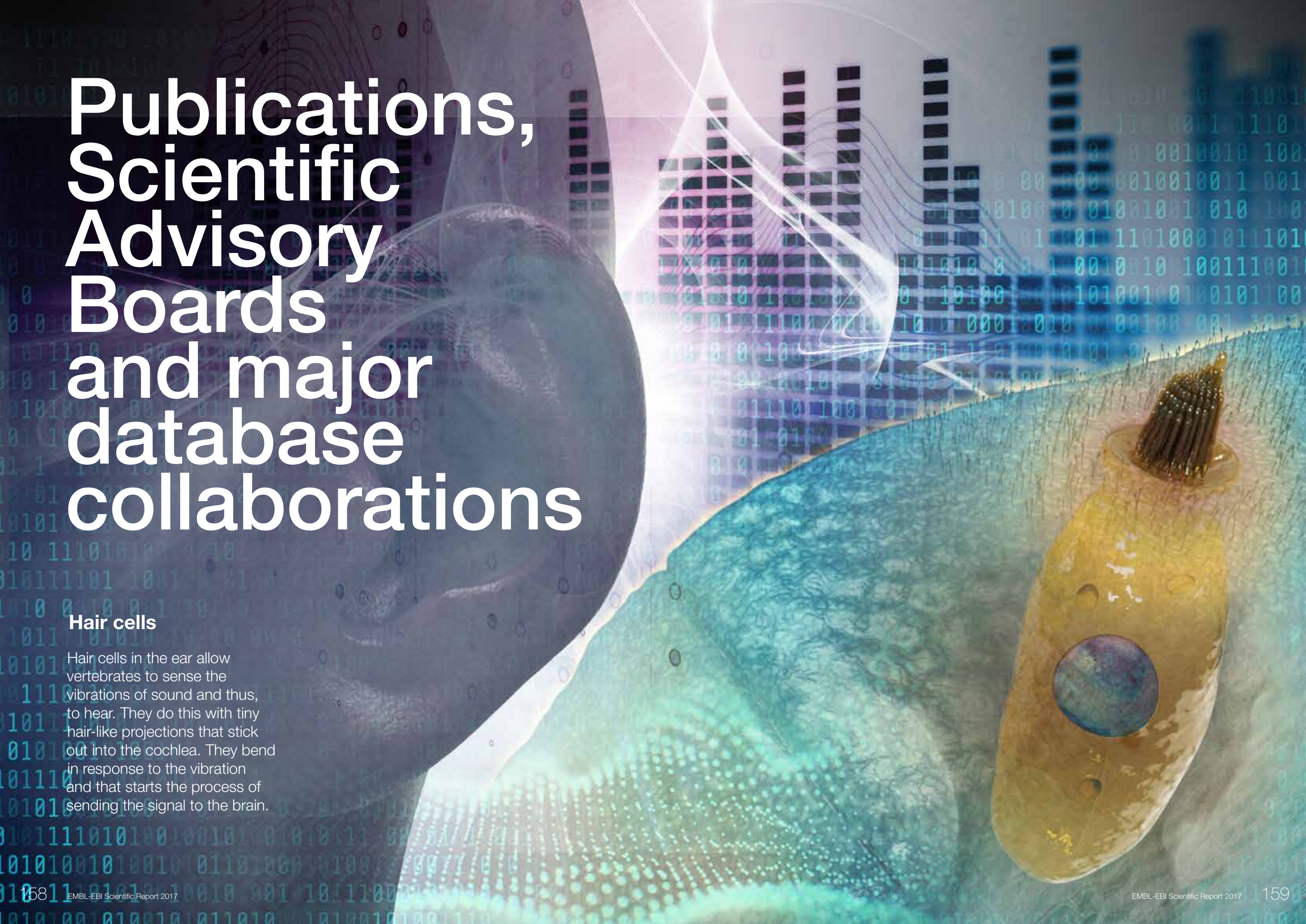
As a Sir Henry Wellcome fellow at EMBL-EBI, Amelie focused on researching social genetic effects. Using mixed models, she showed that healing and anxiety are influenced by the genetics of one's social partners. In September 2017, Amelie joined the Palmer lab at the University of California San Diego, as a visiting researcher.

"My background is in biology with some maths, but I describe myself as a statistical geneticist. What I do now is an extension of my PhD research, where I focused on direct genetic effects, exploring how our own genes affect our traits.

I like the dynamics at EMBL-EBI. There is a good emphasis on work-life balance, which I've always thought of as prerequisite for happiness and well-being. From flexible working hours and health insurance to an on-site nursery, there are a lots of benefits and opportunities here.

Learning also plays a big part for me. I like that there are daily science talks on campus, and excellent training courses, both technical and to improve soft skills. Doing science is made easy at EMBL-EBI. The computational power we have here is amazing and it comes with excellent administrative support."





Publications, Scientific Advisory Boards and major database collaborations

Hair cells

Hair cells in the ear allow vertebrates to sense the vibrations of sound and thus, to hear. They do this with tiny hair-like projections that stick out into the cochlea. They bend in response to the vibration and that starts the process of sending the signal to the brain.

Publications

EMBL is proud to be a member of the ORCID Foundation, the public, open registry of unique researcher identifiers that helps researchers take credit for their work. This list, based on ORCID IDs and affiliation data extracted from Web of Science, represents EMBL-EBI articles published online in 2017, ordered by doi number.

001. Peterson RE, Cai N, Bigdeli TB, et al. (2017). The Genetic Architecture of Major Depressive Disorder in Han Chinese Women. *JAMA Psychiatry* 74(2):162-168. doi:10.1001/jamapsychiatry.2016.3578
002. Rostom R, Svensson V, Teichmann SA, et al. (2017). Computational approaches for interpreting scRNA-seq data. *FEBS Lett* 591(15):2213-2222. doi:10.1002/1873-3468.12684
003. Arshad Q, Bonsu A, Lobo R, et al. (2017) Biased numerical cognition impairs economic decision-making in Parkinson's disease. *Ann Clin Transl Neurol* 4(10):739-748. doi:10.1002/acn3.449
004. Sundaram V, Wang T (2017). Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells: Re-Visiting the "Gene-Battery" Model. *Bioessays*. doi:10.1002/bies.201700155
005. Jarnuczak AF, Vizcaino JA (2017). Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets. *Curr Protoc Bioinformatics* 59:13.31.1-13.31.12. doi:10.1002/cpbi.30
006. Prakash A, Jeffries M, Bateman A, et al. (2017). The HMMER Web Server for Protein Sequence Similarity Search. *Curr Protoc Bioinformatics* 60:3.15.1-3.15.23. doi:10.1002/cpbi.40
007. Burgess S, Zuber V, Valdes-Marquez E, et al. (2017). Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genet. Epidemiol.* doi:10.1002/gepi.22077
008. Carraro M, Minervini G, Giollo M, et al. (2017). Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. *Hum Mutat* 38(9):1042-1050. doi:10.1002/humu.23235
009. Suri M, Evers JMG, Laskowski RA, et al. (2017). Protein structure and phenotypic analysis of pathogenic and population missense variants in STXBP1. *Mol Genet Genomic Med* 5(5):495-507. doi:10.1002/mgg3.304
010. Perez-Riverol Y, Tennent T, Koch M, et al. (2017). OLS Client and OLS Dialog: Open source tools to annotate public omics datasets. *Proteomics* 17(19). doi:10.1002/pmic.201700244
011. Laskowski RA, Jab ńska J, Pravda L, et al. (2017). PDBsum: Structural summaries of PDB entries. *Protein Sci* 27(1):129-134. doi:10.1002/pro.3289
012. Barradas-Bautista D, Moal IH, Fernández-Recio J (2017). A systematic analysis of scoring functions in rigid-body protein docking: the delicate balance between the predictive rate improvement and the risk of overtraining. *Proteins* 85(7):1287-1297. doi:10.1002/prot.25289
013. Rifaoglu AS, Doğan T, Saraç ÖS, et al. (2017). Large-scale Automated Function Prediction of Protein Sequences and an Experimental Case Study Validation on PTEN Transcript Variants. *Proteins* 86(2):135-151. doi:10.1002/prot.25416
014. Lensink MF, Velankar S, Baek M, et al. (2017). The challenge of modeling protein assemblies: The CASP12-CAPRI experiment. *Proteins*. doi:10.1002/prot.25419
015. Lloret-Villas A, Varusai TM, Juty N, et al. (2017). The Impact of Mathematical Modeling in Understanding the Mechanisms Underlying Neurodegeneration: Evolving Dimensions and Future Directions. *CPT Pharmacometrics Syst Pharmacol* 6(2):73-86. doi:10.1002/psp4.12155
016. Bizzotto R, Comets E, Smith G, et al. (2017). PharmML in Action – An interoperable language for modelling and simulation. *CPT Pharmacometrics Syst Pharmacol* 6(10):651-665. doi:10.1002/psp4.12213
017. Smith MK, Moodie SL, Bizzotto R, et al. (2017). Model Description Language (MDL): A Standard for Modeling and Simulation. *CPT Pharmacometrics Syst Pharmacol* 6(10):647-650. doi:10.1002/psp4.12222
018. Traynard P, Tobalina L, Eduati F, et al. (2017). Logic modeling in quantitative systems pharmacology. *CPT Pharmacometrics Syst Pharmacol* 6(8):499-511. doi:10.1002/psp4.12225
019. Kim JH, Kurtz A, Yuan BZ, et al. (2017). Report of the International Stem Cell Banking Initiative Workshop Activity: Current Hurdles and Progress in Seed-Stock Banking of Human Pluripotent Stem Cells. *Stem Cells Transl Med* 6(11):1956-1962. doi:10.1002/sctm.17-0144
020. Pundir S, Martin MJ, O'Donovan C (2017). UniProt Protein Knowledgebase. *Methods Mol Biol* 1558:41-55. doi:10.1007/978-1-4939-6783-4_2
021. Rawlings ND (2017). Using the MEROPS Database for Investigation of Lysosomal Peptidases, Their Inhibitors, and Substrates. *Methods Mol Biol* 1594:213-226. doi:10.1007/978-1-4939-6934-0_14
022. Burley SK, Berman HM, Kleywegt GJ, et al. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol Biol* 1607:627-641. doi:10.1007/978-1-4939-7000-1_26
023. Laskowski RA (2017). The ProFunc Function Prediction Server. *Methods Mol Biol* 1611:75-95. doi: 10.1007/978-1-4939-7015-5_7
024. Saidi R, Boudelloua I, Martin MJ, et al. (2017). Rule Mining Techniques to Predict Prokaryotic Metabolic Pathways. *Methods Mol Biol* 1613:311-331. doi:10.1007/978-1-4939-7027-8_12
025. Seaton DD (2017). ODE-Based Modeling of Complex Regulatory Circuits. *Methods Mol Biol* 1629:317-330. doi:10.1007/978-1-4939-7125-1_20
026. Kuepfer L, Clayton O, Thiel C, et al. (2017). A model-based assay design to reproduce in vivo patterns of acute drug-induced toxicity. *Arch Toxicol* 92(1):553-555. doi:10.1007/s00204-017-2041-7
027. Riesgo A, Burke EA, Laumer C, et al. (2017). Genetic variation and geographic differentiation in the marine triclad *Bdelloura candida* (Platyhelminthes, Tricladida, Maricola), ectocommensal on the American horseshoe crab *Limulus polyphemus*. *Mar Biol* 164(5):1-14. doi:10.1007/s00227-017-3132-y
028. Kagami LP, das Neves GM, da Silva AWS, et al. (2017). LiGRO: a graphical user interface for protein-ligand molecular dynamics. *J Mol Model* 23(11): doi:10.1007/s00894-017-3475-9
029. Buettner F, Jay K, Wischnewski H, et al. (2017). Non-targeted metabolomic approach reveals two distinct types of metabolic responses to telomerase dysfunction in *S. cerevisiae*. *Metabolomics* 13. doi:10.1007/s11306-017-1195-x
030. Spicer R, Salek RM, Moreno P, et al. (2017). Navigating freely-available software tools for metabolomics analysis. *Metabolomics* 13(9). doi:10.1007/s11306-017-1242-7
031. Tachmazidou I, Süveges D, Min JL, et al. (2017). Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am J Hum Genet* 100(6):865-884. doi:10.1016/j.ajhg.2017.04.014
032. Haug K, Salek RM, Steinbeck C (2017). Global open data management in metabolomics. *Curr Opin Chem Biol* 36:58-63. doi:10.1016/j.cbpa.2016.12.024
033. Yates LR, Knappskog S, Wedge D, et al. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* 32(2):169-184.e7. doi:10.1016/j.ccell.2017.07.005
034. Cabezas-Wallscheid N, Buettner F, Sommerkamp P, et al. (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell* 169(5):807-823. doi:10.1016/j.cell.2017.04.018
035. Jiménez-Sánchez A, Memon D, Pourpe S, et al. (2017). Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient. *Cell* 170(5):927-938. doi:10.1016/j.cell.2017.07.025
036. Martincorena I, Raine KM, Gerstung M, et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* 171(5):1029-1041. doi:10.1016/j.cell.2017.09.042
037. Mohammed H, Hernando-Herraez I, Savino A, et al. (2017). Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep* 20(5):1215-1228. doi:10.1016/j.celrep.2017.07.009
038. Liu Y, González-Porta M, Santos S, et al. (2017). Impact of Alternative Splicing on the Human Proteome. *Cell Rep* 20(5):1229-1241. doi:10.1016/j.celrep.2017.07.025
039. Roumeliotis TI, Williams SP, Goncalves E, et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. *Cell Rep* 20(9):2201-2214. doi: 10.1016/j.celrep.2017.08.010
040. Bolukbasi E, Khericha M, Regan JC, et al. (2017). Intestinal Fork Head Regulates Nutrient Absorption and Promotes Longevity. *Cell Rep* 21(3):641-653. doi:10.1016/j.celrep.2017.09.042
041. Invergo BM, Brochet M, Yu L, et al. (2017). Sub-minute Phosphoregulation of Cell Cycle Systems during Plasmodium Gamete Formation. *Cell Rep* 21(7):2017-2029. doi:10.1016/j.celrep.2017.10.071
042. Koo BM, Kritikos G, Farelli JD, et al. (2017). Construction and Analysis of Two Genome-Scale Deletion Libraries for *Bacillus subtilis*. *Cell Syst* (16). doi:10.1016/j.cels.2016.12.013
043. Gonçalves E, Fragoulis A, Garcia-Alonso L, et al. (2017). Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Syst* 5(4):386-398. doi:10.1016/j.cels.2017.08.013
044. Mehmet Gönen, Barbara A. Weir, Glenn S. Cowley, et al. (2017). A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell Lines. *Cell Syst* 5(5):485-497. doi:10.1016/j.cels.2017.09.004
045. Blattmann P, Henriques D, Zimmermann M, et al. (2017). Systems Pharmacology Dissection of Cholesterol Regulation Reveals Determinants of Large Pharmacodynamic Variability between Cell Lines. *Cell Syst*. doi:10.1016/j.cels.2017.11.002
046. Tang J, Tanoli ZU, Ravikumar B, et al. (2017). Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. *Cell Chem Biol*. doi:10.1016/j.chembiol.2017.11.009
047. Antoranz A, Sakellaropoulos T, Saez-Rodriguez J, et al. (2017). Mechanism-based biomarker discovery. *Drug Discov Today* 22(8):1209-1215. doi:10.1016/j.drudis.2017.04.013
048. Khaladkar M, Koscielny G, Hasan S, et al. (2017). Uncovering novel repositioning opportunities using the Open Targets platform. *Drug Discov Today* 22(12):30189 doi:10.1016/j.drudis.2017.09.007
049. Salgado R, Moore H, Martens JWM, et al. (2017). Societal challenges of precision medicine: Bringing order to chaos. *Eur J Cancer* 84:325-334. doi:10.1016/j.ejca.2017.07.028
050. Puzyn T, Jeliakova N, Sarimveis H, et al. (2017). Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology. *Food Chem Toxicol*. doi:10.1016/j.fct.2017.09.037
051. Natarajan KN, Teichmann SA, Kolodziejczyk AA (2017). Single cell transcriptomics of pluripotent stem cells: reprogramming and differentiation. *Curr Opin Genet Dev* 46:66-76. doi:10.1016/j.gde.2017.06.003
052. Pfeuffer J, Sachsenberg T, Alka O, et al. (2017). OpenMS – A platform for reproducible analysis of mass spectrometry data. *J Biotechnol* S0168-1656(17):30251 doi:10.1016/j.jbiotec.2017.05.016
053. Cern A, Marcus D, Tropsha A, et al. (2017). New drug candidates for liposomal delivery identified by computer modeling of liposomes' remote loading and leakage. *J Control Release* 252:18-27. doi:10.1016/j.jconrel.2017.02.015
054. Valasatava Y, Rosato A, Furnham N, et al. (2017). To what extent do structural changes in catalytic metal sites affect enzyme function? *J Inorg Biochem*. 179:40-53. doi:10.1016/j.jinorgbio.2017.11.002
055. Joseph AP, Lagerstedt I, Patwardhan A, et al. (2017). Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J Struct Biol* 199(1):12-26. doi:10.1016/j.jsb.2017.05.007
056. Bresson S, Tuck A, Staneva D, et al. (2017). Nuclear RNA Decay Pathways Aid Rapid Remodeling of Gene Expression in Yeast. *Mol Cell* 65(5):787-800.e5. doi:10.1016/j.molcel.2017.01.005
057. Ivanova I, Much C, Di Giacomo M, et al. (2017). The RNA m(6)A Reader YTHDF2 Is Essential for the Post-transcriptional Regulation of the Maternal Transcriptome and Oocyte Competence. *Mol Cell* 67(6):1059-1067.e4. doi:10.1016/j.molcel.2017.08.003
058. Ayad LAK, Barton C, Pissis SP (2017). A faster and more accurate heuristic for cyclic edit distance computation. *Pattern Recogn Lett*. 88:81-87. doi:10.1016/j.patrec.2017.01.018
059. Tyzack JD, Furnham N, Sillitoe I, et al. (2017). Understanding enzyme function evolution from a computational perspective. *Curr Opin Struct Biol*. 47:131-139. doi:10.1016/j.sbi.2017.08.003
060. Thornton J, Orengo C (2017). Editorial overview: Catalysis and regulation. *Curr Opin Struct Biol* 47. doi:10.1016/j.sbi.2017.11.005
061. De Sousa PA, Steeg R, Wächter E, et al. (2017). Rapid establishment of the European Bank for induced Pluripotent Stem Cells (EBiSC) – the Hot Start experience. *Stem Cell Res* 20:105-114. doi:10.1016/j.scr.2017.03.002
062. González S, Volkova N, Beer P, et al. (2017). Immuno-oncology from the perspective of somatic evolution. *Semin Cancer Biol*. doi:10.1016/j.semcancer.2017.12.001
063. Ambrosi TH, Scialdone A, Graja A, et al. (2017). Adipocyte Accumulation in the Bone Marrow during Obesity and Aging Impairs Stem Cell-Based Hematopoietic and Bone Regeneration. *Cell Stem Cell*. doi:10.1016/j.stem.2017.02.009
064. Welby E, Lakowski J, Di Foggia V, et al. (2017). Isolation and Comparative Transcriptome Analysis of Human Fetal and iPSC-Derived Cone Photoreceptor Cells. *Stem Cell Reports* 9(6):1898-1915. doi:10.1016/j.stemcr.2017.10.018
065. Young JY, Westbrook JD, Feng Z, et al. (2017). OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. *Structure* 25(3):536-545. doi:10.1016/j.str.2017.01.004
066. Gore S, Sanz García E, Hendrickx PMS, et al. (2017). Validation of Structures in the Protein Data Bank. *Structure* 25(12):1916-1927. doi:10.1016/j.str.2017.10.009
067. Martens L, Vizcaino JA (2017). A Golden Age for Working with Public Proteomics Data *Trends Biochem Sci* (17). doi:10.1016/j.tibs.2017.01.001
068. Paterson RR, Lima N, Brooksbank C, et al. (2017). Microbiology Managers: Managerial Training in the Rltrain Project. *Trends Microbiol* S0966-842X(17):30054-30059. doi:10.1016/j.tim.2017.03.002
069. Gonçalves E, Sciacovelli M, Costa ASH, et al. (2017). Post-translational regulation of metabolism in fumarate hydratase deficient cancer cells. *Metab Eng* 45:149-157. doi:10.1016/j.ymben.2017.11.011
070. Bittremieux W, Walzer M, Tenzer S, et al. (2017). The Human Proteome Organization-Proteomics Standards Initiative Quality Control Working Group: Making Quality Control More Accessible for Biological Mass Spectrometry. *Anal Chem* doi:10.1021/acs.analchem.6b04310
071. Verras A, Waller CL, Gedeck P, et al. (2017). Shared Consensus Machine Learning Models for Predicting Blood Stage Malaria Inhibition. *J Chem Inf Model* 57(3):445-453. doi:10.1021/acs.jcim.6b00572

072. Tresadern G, Trabanco AA, Pérez-Benito L, et al. (2017). Identification of Allosteric Modulators of Metabotropic Glutamate 7 Receptor Using Proteochemometric Modeling. *J Chem Inf Model* 57(12):2976-2985. doi:10.1021/acs.jcim.7b00338

073. Hayes TW, Moal IH (2017). Modeling Protein Conformational Transition Pathways Using Collective Motions and the LASSO Method. *J Chem Theory Comput* 13(3):1401-1410. doi:10.1021/acs.jctc.6b01110

074. Terfve C, Sabidó E, Wu Y, et al. (2017). System-Wide Quantitative Proteomics of the Metabolic Syndrome in Mice: Genotypic and Dietary Effects. *J Proteome Res* 16(2):831-841. doi:10.1021/acs.jproteome.6b00815

075. Imamura H, Wagih O, Niinae T, et al. (2017). Identifications of Putative PKA Substrates with Quantitative Phosphoproteomics and Primary-Sequence-Based Scoring. *J Proteome Res* 16(4):1825-1830. doi:10.1021/acs.jproteome.7b00087

076. Deutsch EW, Orchard S, Binz PA, et al. (2017). Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J Proteome Res* 16(12):4288-4298. doi:10.1021/acs.jproteome.7b00370

077. Guruceaga E, Garin-Muga A, Prieto G, et al. (2017). Enhanced Missing Proteins Detection in NCI60 Cell Lines Using an Integrative Search Engine Approach. *J Proteome Res* 16(12):4374-4390. doi:10.1021/acs.jproteome.7b00388

078. Valteau S, Studer RA, Häse F, et al. (2017). Absence of Selection for Quantum Coherence in the Fenna-Matthews-Olson Complex: A Combined Evolutionary and Excitonic Study. *ACS Cent Sci* 3(10):1086-1095. doi:10.1021/acscentsci.7b00269

079. Anderson WP, Global Life Science Data Resources Working Group. (2017). Data management: A global coalition to sustain core data. *Nature* 543(7644):179 doi:10.1038/543179a

080. Bolli N, Biancon G, Moarri M, et al. (2017). Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups. *Leukemia*. doi:10.1038/leu.2017.344

081. Ju YS, Martincorena I, Gerstung M, et al. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* 543(7647):714-718. doi:10.1038/nature21703

082. Mascher M, Gundlach H, Himmelbach A, et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544(7651):426-433. doi:10.1038/nature22043

083. Kilpinen H, Goncalves A, Leha A, et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*. doi:10.1038/nature22403

084. Morgan M, Much C, DiGiacomo M, et al. (2017). mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature* 548(7667):347-351. doi:10.1038/nature23318

085. Tan MH, Li Q, Shanmugam R, et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550(7675):249-254. doi:10.1038/nature24041

086. Yang J, Ryan DJ, Wang W, et al. (2017). Establishment of mouse expanded potential stem cells. *Nature* 550(7676):393-397. doi:10.1038/nature24052

087. Tukiainen T, Villani AC, Yen A, et al. (2017). Landscape of X chromosome inactivation across human tissues. *Nature* 550(7675):244-248. doi:10.1038/nature24265

088. Li X, Kim Y, Sang EKT, et al. (2017). The impact of rare variation on gene expression across tissues. *Nature* 550(7675):239-243. doi:10.1038/nature24267

089. Aguet F, Brown AA, Castel SE, et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550:204-213. doi:10.1038/nature24277

090. Schwarzer W, Abdennur N, Goloborodko A, et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 551(7678):51-56. doi:10.1038/nature24281

091. Shahbazi MN, Scialdone A, Skorupska N, et al. (2017). Pluripotent state transitions coordinate morphogenesis in mouse and human embryos. *Nature*. doi:10.1038/nature24675

092. Perez-Riverol Y, Bai M, da Veiga Leprevost F, et al. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nat Biotechnol* 35(5):406-409. doi:10.1038/nbt.3790

093. Bowers RM, Kyrpides NC, Stepanauskas R, et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35(8):725-731. doi:10.1038/nbt.3893

094. Donati G, Rognoni E, Hiratsuka T, et al. (2017). Wounding induces dedifferentiation of epidermal Gata6(+) cells and acquisition of stem cell properties. *Nat Cell Biol*. doi:10.1038/ncb3532

095. Iraci N, Gaude E, Leonardi T, et al. (2017). Extracellular vesicles are independent metabolic units with asparaginase activity. *Nat Chem Biol*. doi:10.1038/nchembio.2422

096. Adema CM, Hillier LW, Jones CS, et al. (2017). Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat Commun* 8. doi:10.1038/ncomms15451

097. Zhao Y, Gilliat AF, Ziehm M, et al. (2017). Two forms of death in ageing *Caenorhabditis elegans*. *Nat Commun* 8:1-8. doi:10.1038/ncomms15458

098. Karp NA, Mason J, Beaudet AL, et al. (2017). Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat Commun* 8. doi:10.1038/ncomms15475

099. Xue YL, Mezzavilla M, Haber M, et al. (2017). Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* 8:15927. doi:10.1038/ncomms15927

100. Petersen R, Lambourne JJ, Javierre BM, et al. (2017). Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nat Commun* 8. doi:10.1038/ncomms16058

101. Gerstung M, Papaemmanuil E, Martincorena I, et al. (2017). Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet* 49(3):332-340. doi:10.1038/ng.3756

102. Glodzik D, Morganella S, Davies H, et al. (2017). A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet* 49(3):341-348. doi:10.1038/ng.3771

103. Schor IE, Degner JF, Harnett D, et al. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat Genet* 49(4):550-558. doi:10.1038/ng.3791

104. Speed D, Cai N, Johnson MR, et al. (2017). Reevaluation of SNP heritability in complex human traits. *Nat Genet* 49(7):986-992. doi:10.1038/ng.3865

105. Meehan TF, Conte N, West DB, et al. (2017). Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat Genet* doi:10.1038/ng.3901

106. Jasinska AJ, Zelaya I, Service SK, et al. (2017). Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. *Nat Genet* 49(12):1714-1721. doi:10.1038/ng.3959

107. Davies H, Glodzik D, Morganella S, et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* 23(4):517-525. doi:10.1038/nm.4292

108. Buggenthin F, Buettner F, Hoppe PS, et al. (2017). Prospective identification of hematopoietic lineage choice by deep learning. *Nat Methods* 14(4):403-406. doi:10.1038/nmeth.4182

109. Perez-Riverol Y, Vizcaino JA (2017). Synthetic human proteomes for accelerating protein research. *Nat Methods* 14(3):240-242. doi:10.1038/nmeth.4191

110. Svensson V, Natarajan KN, Ly LH, et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14(4):381-387. doi:10.1038/nmeth.4220

111. Kiselev VY, Kirschner K, Schaub MT, et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 14(5):483-486. doi:10.1038/nmeth.4236

112. Vallejos CA, Risso D, Scialdone A, et al. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 14(6):565-571. doi:10.1038/nmeth.4292

113. Lun ATL, Richard AC, Marioni JC (2017). Testing for differential abundance in mass cytometry data. *Nat Methods* 14(7):707-709. doi:10.1038/nmeth.4295

114. Williams E, Moore J, Li SW, et al. (2017). The Image Data Resource: A Bioimage Data Integration and Publication Platform. *Nat Methods* 14(8):775-781. doi:10.1038/nmeth.4326

115. Liechti R, George N, Götz L, et al. (2017). SourceData: a semantic platform for curating and searching figures. *Nat Methods* 14(11):1021-1022. doi:10.1038/nmeth.4471

116. Sanz F, Pognan F, Steger-Hartmann T, et al. (2017). Legacy data sharing to improve drug safety assessment: the eTOX project. *Nat Rev Drug Discov*. doi:10.1038/nrd.2017.177

117. Thornton JM, Valencia A, Schwede T (2017). Anna Tramontano 1957-2017. *Nat Struct Mol Biol* 24(5):431-432. doi:10.1038/nsmb.3410

118. Vasilakaitė L, Vitsios D, Berrens RV, et al. (2017). A MILI-independent piRNA biogenesis pathway empowers partial germline reprogramming. *Nat Struct Mol Biol* 24(7):604-606. doi:10.1038/nsmb.3413

119. Kar G, Kim JK, Kolodziejczyk AA, et al. (2017). Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *Nat Commun* 8(1):36 doi:10.1038/s41467-017-00052-2

120. Bowl MR, Simon MM, Ingham NJ, et al. (2017). A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. *Nat Commun* 8(1): doi:10.1038/s41467-017-00595-4

121. Wu Q, Ferry QRV, Baeumlér TA, et al. (2017). In situ functional dissection of RNA cis-regulatory elements by multiplex CRISPR-Cas9 genome engineering. *Nat Commun* 8(1): doi:10.1038/s41467-017-00686-2

122. Wong ES, Schmitt BM, Kazachenka A, et al. (2017). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat Commun* 8(1). doi:10.1038/s41467-017-01037-x

123. Schulz H, Ruppert AK, Herms S, et al. (2017). Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat Commun* 8(1). doi:10.1038/s41467-017-01818-4

124. Bach K, Pensa S, Grzelak M, et al. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun* 8(1). doi:10.1038/s41467-017-02001-5

125. Berthelot C, Villar D, Horvath JE, et al. (2017). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* 2(1):152-163. doi:10.1038/s41559-017-0377-2

126. Schwartzentruber J, Fokolou S, Kilpinen H, et al. (2017). Molecular and functional variation in iPSC-derived sensory neurons. *Nat Genet*. doi:10.1038/s41588-017-0005-8

127. Mahlich Y, Reeb J, Hecht M, et al. (2017). Common sequence variants affect molecular function more than rare variants? *Sci Rep* 7(1):1-13. doi:10.1038/s41598-017-01054-2

128. Hendricks AE, Bochukova EG, Marenne G, et al. (2017). Rare Variant Analysis of Human and Rodent Obesity Genes in Individuals with Severe Childhood Obesity. *Sci Rep* 7(1). doi:10.1038/s41598-017-03054-8

129. Contreras-Martos S, Plai A, Kosol S, et al. (2017). Linking functions: an additional role for an intrinsically disordered linker domain in the transcriptional coactivator CBP. *Sci Rep* 7(1):4676. doi:10.1038/s41598-017-04611-x

130. Steinberg J, Ritchie GRS, Roumeliotis TI, et al. (2017). Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. *Sci Rep* 7(1). doi:10.1038/s41598-017-09335-6

131. Moya-García A, Adeyelu T, Kruger FA, et al. (2017). Structural and Functional View of Polypharmacology. *Sci Rep* 7(1). doi:10.1038/s41598-017-10012-x

132. Cai N, Bigdeli TB, Kretschmar WW, et al. (2017). 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci Data* 4:170011. doi:10.1038/sdata.2017.11

133. Spicer RA, Salek R, Steinbeck C (2017). Compliance with minimum information guidelines in public metabolomics repositories. *Sci Data* 4. doi:10.1038/sdata.2017.137

134. Spicer RA, Salek R, Steinbeck C (2017). Comment: A decade after the metabolomics standards initiative it's time for a revision. *Sci Data* 4. doi:10.1038/sdata.2017.138

135. Beier S, Himmelbach A, Colmsee C, et al. (2017). Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci Data* 4:1-24. doi:10.1038/sdata.2017.44

136. Alberti A, Poulain J, Engelen S, et al. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* 4(170093). doi:10.1038/sdata.2017.93

137. Carroni M, De March M, Medagli B, et al. (2017). New insights into the GINS complex explain the controversy between existing structural models. *Sci Rep* 7:40188. doi:10.1038/srep40188

138. Arseneault M, Monlong J, Vasudev NS, et al. (2017). Loss of chromosome Y leads to down regulation of KDM5D and KDM6C epigenetic modifiers in clear cell renal cell carcinoma. *Sci Rep* 7(44876):1-8. doi:10.1038/srep44876

139. Zaru R, Magrane M, O'Donovan C, et al. (2017). From the research laboratory to the database: the *Caenorhabditis elegans* kinome in UniProtKB. *Biochem J* 474(4):493-515. doi:10.1042/bcj20160991

140. Hegazy AN, West NR, Stubbington MJT, et al. (2017). Circulating and Tissue-resident CD4+ T Cells With Reactivity to Intestinal Microbiota Are Abundant in Healthy Individuals and Function is Altered During Inflammation. *Gastroenterology*. doi:10.1053/j.gastro.2017.07.047

141. Howell KJ, Kraicz J, Nayak KM, et al. (2017). DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology*. doi:10.1053/j.gastro.2017.10.007

142. Varshney A, Scott LJ, Welch RP, et al. (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci USA* 114(9):2301-2306. doi:10.1073/pnas.1621192114

143. Vizcaino JA, Mayer G, Perkins SR, et al. (2017). The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Mol Cell Proteomics*. doi:10.1074/mcp.M117.068429

144. Gabrys HS, Buettner F, Sterzing F, et al. (2017). Parotid gland mean dose as a xerostomia predictor in low-dose domains. *Acta Oncol* 56(9):1197-1203. doi:10.1080/0284186X.2017.1324209

145. Nowotka MM, Gaulton A, Mendez D, et al. (2017). Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opin Drug Discov* 12(8):1-11. doi:10.1080/17460441.2017.1339032

146. Pathan M, Keerthikumar S, Chisanga D, et al. (2017). A novel community driven software for functional enrichment analysis of extracellular vesicles data. *J Extracell Vesicles* 6(1). doi:10.1080/2013078.2017.1321455

147. Thomas DC, Clare S, Sowerby JM, et al. (2017). Eros is a novel transmembrane protein that controls the phagocyte respiratory burst and is essential for innate immunity. *J Exp Med* 214(4):1111-1128. doi:10.1084/jem.20161382

148. Giudice G, Petsalaki E (2017). Proteomics and phosphoproteomics in precision medicine: applications and challenges. *Brief Bioinform*. doi:10.1093/bib/bbx141

149. McCarthy DJ, Campbell KR, Lun AT, et al. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33(8):1179-1186. doi:10.1093/bioinformatics/btw777

150. Psomopoulos FE, Vitsios DM, Baichoo S, et al. (2017). BioPAXviz: a cytoscape application for the visual exploration of metabolic pathway evolution. *Bioinformatics* 33(9):1418-1420. doi:10.1093/bioinformatics/btw813

151. Moal IH, Barradas-Bautista D, Jiménez-García B, et al. (2017). IRePPA: information retrieval based integration of biophysical models for protein assembly selection. *Bioinformatics* 33(12):1806-1813. doi:10.1093/bioinformatics/btx068

152. Hernandez-Armenta C, Ochoa D, Gonçalves E, et al. (2017). Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*. doi:10.1093/bioinformatics/btx082

153. Watkins X, Garcia LJ, Pundir S, et al. (2017). ProtVista: visualization of protein sequence annotations. *Bioinformatics*:1-2. doi:10.1093/bioinformatics/btx120

154. Petryszak R, Fonseca NA, Füllgrabe A, et al. (2017). The RNASeq-er API – a gateway to systematically updated analysis of public RNA-Seq data. *Bioinformatics*. doi:10.1093/bioinformatics/btx143

155. Larralde M, Lawson TN, Weber RJ, et al. (2017). mzML2ISA & nmrML2ISA: generating enriched ISA-Tab metadata files from metabolomics XML data. *Bioinformatics*. doi:10.1093/bioinformatics/btx169

156. Leprevost FD, Grüning BA, Alves Aflitos S, et al. (2017). BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics* 1-3. doi:10.1093/bioinformatics/btx192

157. Poux S, Arighi CN, Magrane M, et al. (2017). On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*. doi:10.1093/bioinformatics/btx439

158. Sidiropoulos K, Viteri G, Sevilla C, et al. (2017). Reactome enhanced pathway visualization. *Bioinformatics* 33(21):3461-3467. doi:10.1093/bioinformatics/btx441

159. Wagih O (2017). ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*. doi:10.1093/bioinformatics/btx469

160. Combe CW, Sivade MD, Hermjakob H, et al. (2017). ComplexViewer: visualization of curated macromolecular complexes. *Bioinformatics*. doi:10.1093/bioinformatics/btx497

161. Torrente A, Brazma A (2017). clustComp, a Bioconductor package for the comparison of clustering results. *Bioinformatics* 33(24):4001-4003. doi:10.1093/bioinformatics/btx532

162. Forslund K, Pereira C, Capella-Gutierrez S, et al. (2017). Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*. doi:10.1093/bioinformatics/btx542

163. Biffi C, de Marvao A, Attard M, et al. (2017). Three-dimensional Cardiovascular Imaging-Genetics: A Mass Univariate Framework. doi:10.1093/bioinformatics/btx552

164. Cokelaer T, Chen E, Iorio F, et al. (2017). GDSCTools for Mining Pharmacogenomic Interactions in Cancer. *Bioinformatics*. doi:10.1093/bioinformatics/btx744

165. Fabregat A, Sidiropoulos K, Viteri G, et al. (2017). Reactome diagram viewer: Data structures and strategies to boost performance. *Bioinformatics*. doi:10.1093/bioinformatics/btx752

166. Lun ATL, Marioni JC (2017). Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* 18(3):451-464. doi:10.1093/biostatistics/kxw055

167. Ruffier M, Kähäri A, Komorowska M, et al. (2017). Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database (Oxford)* 2017(1):1-11. doi:10.1093/database/bax020

168. Boddy AM, Harrison PW, Montgomery SH, et al. (2017). Evidence of a Conserved Molecular Response to Selection for Increased Brain Size in Primates. *Genome Biol Evol* 9(3):700-713. doi:10.1093/gbe/evx028

169. Taskent RO, Alioglu ND, Fer E, et al. (2017). Variation and Functional Impact of Neanderthal Ancestry in Western Asia. *Genome Biol Evol* 9(12):3516-3524. doi:10.1093/gbe/evx216

170. Zheng-Bradley X, Streeter I, Fairley S, et al. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* 6(7):1-8. doi:10.1093/gigascience/gix038

171. Hoopen PT, Finn RD, Bongo LA, et al. (2017). The metagenomic data life-cycle: standards and best practices. *Gigascience* 6(8):1-11. doi:10.1093/gigascience/gix047

172. Salek RM, Conesa P, Cochrane K, et al. (2017). Automated assembly of species metabolomes through data submission into a public repository. *Gigascience* 6(8):1-4. doi:10.1093/gigascience/gix062

173. Evers JM, Laskowski RA, Bertolli M, et al. (2017). Structural analysis of pathogenic mutations in the DYRK1A gene in patients with developmental disorders. *Hum Mol Genet* 26(3):519-526. doi:10.1093/hmg/ddw409

174. Raisaro JL, Tramèr F, Ji Z, et al. (2017). Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *J Am Med Inform Assoc* 1-8. doi:10.1093/jamia/ocw167

175. Gelová Z, Ten Hoopen P, Novák O, et al. (2017). Antibody-mediated modulation of cytokinins in tobacco: organ-specific changes in cytokinin homeostasis. *J Exp Bot*. doi:10.1093/jxb/erx426

176. Levchenko M, Gou Y, Graef F, et al. (2017). Europe PMC in 2017. *Nucleic Acids Res*. doi:10.1093/nar/gkx1005

177. Kersey PJ, Allen JE, Allot A, et al. (2017). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res*. doi:10.1093/nar/gkx1011

178. Ribeiro AJM, Holliday GL, Furnham N, et al. (2017). Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res*. doi:10.1093/nar/gkx1012

179. Glont M, Nguyen TVN, Graesslin M, et al. (2017). BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res*. 46(D1):D1248-D1253. doi:10.1093/nar/gkx1023

180. Pujar S, O'Leary NA, Farrell CM, et al. (2017). Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res*. doi:10.1093/nar/gkx1031

181. Kalvari I, Argasinska J, Quinones-Olvera N, et al. (2017). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*. doi:10.1093/nar/gkx1038

182. Mir S, Alhroub Y, Anyango S, et al. (2017). PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res*. doi:10.1093/nar/gkx1070

183. Gouw M, Michael S, Sámano-Sánchez H, et al. (2017). The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res*. 46(D1):D428-D434. doi:10.1093/nar/gkx1077

184. Karsch-Mizrachi I, Takagi T, Cochrane G (2017). The international nucleotide sequence database collaboration. *Nucleic Acids Res*. 46(D1):D48-D51. doi:10.1093/nar/gkx1097

185. Zerbino DR, Achuthan P, Akanni W, et al. (2017). Ensembl 2018. *Nucleic Acids Res*. doi:10.1093/nar/gkx1098

186. Tello-Ruiz MK, Naithani S, Stein JC, et al. (2017). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res*. 46(D1):D1181-D1189. doi:10.1093/nar/gkx1111

187. Silvester N, Alako B, Amid C, et al. (2017). The European Nucleotide Archive in 2017. *Nucleic Acids Res*. doi:10.1093/nar/gkx1125

188. Fabregat A, Jupe S, Matthews L, et al. (2017). The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 46(D1):D649-D655. doi:10.1093/nar/gkx1132

189. Rawlings ND, Barrett AJ, Thomas PD, et al. (2017). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. doi:10.1093/nar/gkx1134

190. Cook CE, Bergman MT, Cochrane G, et al. (2017). The European Bioinformatics Institute in 2017: data coordination and integration. *Nucleic Acids Res*. 46(D1):D21-D29. doi:10.1093/nar/gkx1154

191. Papatheodorou I, Fonseca NA, Keays M, et al. (2017). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res*. doi:10.1093/nar/gkx1158

192. Nightingale A, Antunes R, Alpi E, et al. (2017). The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res*. 45(W1):W539-W544. doi:10.1093/nar/gkx237

193. Chojnacki S, Cowley A, Lee J, et al. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res*. 45:W550-W553. doi:10.1093/nar/gkx273

194. Bertelli C, Laird MR, Williams KP, et al. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res*. 45:W30-W35. doi:10.1093/nar/gkx343

195. Park YM, Squizzato S, Buso N, et al. (2017). The EBI search engine: EBI search as a service-making biological data accessible for all. *Nucleic Acids Res*. 45:W545-W549. doi:10.1093/nar/gkx359

196. Shao W, Pedrioli PGA, Wolski W, et al. (2017). The SysMHC Atlas project. *Nucleic Acids Res*. doi:10.1093/nar/gkx664

197. Martin-Herranz DE, Ribeiro AJM, Krueger F, et al. (2017). cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches. *Nucleic Acids Res*. 45(20):11559-11569. doi:10.1093/nar/gkx814

198. Vitsios DM, Kentepozidou E, Quintais L, et al. (2017). Mirnova: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res*. 45(21). doi:10.1093/nar/gkx836

199. Varadi M, De Baets G, Vranken WF, et al. (2017). AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res*. doi:10.1093/nar/gkx950

200. Sarkans U, Gostev M, Athar A, et al. (2017). The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res*. doi:10.1093/nar/gkx965

201. Mitchell AL, Scheremetjew M, Denise H, et al. (2017). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res*. doi:10.1093/nar/gkx967

202. Lee RYN, Howe KL, Harris TW, et al. (2017). WormBase 2017: molting into a new stage. *Nucleic Acids Res*. doi:10.1093/nar/gkx998

203. Pundir S, Onwubiko J, Zaru R, et al. (2017). An update on the Enzyme Portal: an integrative approach for exploring enzyme knowledge. *Protein Eng Des Sel* 30(3):245-251. doi:10.1093/protein/gzx008

204. Klopstein S, Massingham T, Goldman N (2017). More on the Best Evolutionary Rate for Phylogenetic Analysis. *Syst Biol* 66(5):769-785. doi:10.1093/sysbio/syx051

205. Kılınç GM, Koptekin D, Atakuman Ç, et al. (2017). Archaeogenomic analysis of the first steps of Neolithization in Anatolia and the Aegean. *Proc Biol Sci* 284(1867). doi:10.1098/rspb.2017.2064

206. Carmona SJ, Teichmann SA, Ferreira L, et al. (2017). Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. *Genome Res*. doi:10.1101/gr.207704.116

207. Brammell JS, Petljak M, Martincorena I, et al. (2017). Genome-wide chemical mutagenesis screens allow unbiased saturation of the cancer genome and identification of drug resistance mutations. *Genome Res* 27(4):613-625. doi:10.1101/gr.213546.116

208. Schneider VA, Graves-Lindsay T, Howe K, et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 27(5):849-864. doi:10.1101/gr.213611.116

209. Löytynoja A, Goldman N (2017). Short template switch events explain mutation clusters in the human genome. *Genome Res* 27(6):1039-1049. doi:10.1101/gr.214973.116

210. Saha A, Kim Y, Gewirtz ADH, et al. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res* 27(11):1843-1858. doi:10.1101/gr.216721.116

211. Yang F, Wang JB, Pierce BL, et al. (2017). Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res* 27(11):1859-1871. doi:10.1101/gr.216754.116

212. Clavijo BJ, Venturini L, Schudoma C, et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res* 27(5):885-896. doi:10.1101/gr.217117.116

213. Lun ATL, Calero-Nieto FJ, Haim-Vilmovsky L, et al. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res* 27(11):1795-1806. doi:10.1101/gr.222877.117

214. Patwardhan A (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr D Struct Biol* 73(Pt 6):503-508. doi:10.1107/S2059798317004181

215. Ziehm M, Kaur S, Ivanov DK, et al. (2017). Drug repurposing for aging research using model organisms. *Aging Cell*. doi:10.1111/acer.12626

216. Soler M, Macias-Sanchez E, Martin-Galvez D, et al. (2017). Complex feeding behaviour by magpies in nests with great spotted cuckoo nestlings. *J Avian Biol* 48:1406-1413. doi:10.1111/jav.01473

217. Berney C, Ciuprina A, Bender S, et al. (2017). UniEuk: Time to Speak a Common Language in Protistology! *J Eukaryot Microbiol* 64(3):407-411. doi:10.1111/jeu.12414

218. Martinez-Jimenez CP, Eling N, Chen HC, et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 355(6332):1433-1436. doi:10.1126/science.aah4115

219. Kelsey G, Stegle O, Reik W (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science* 358(6359):69-75. doi:10.1126/science.aan6826

220. Lönnberg T, Svensson V, James KR, et al. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Th fate bifurcation in malaria. *Sci Immunol* 2(9):eaal2192-eaal2192. doi:10.1126/sciimmunol.aal2192

221. Heninger AK, Eugster A, Kuehn D, et al. (2017). A divergent population of autoantigen-responsive CD4+ T cells in infants prior to cell autoimmunity. *Sci Transl Med* 9(378). doi:10.1126/scitranslmed.aaf8848

222. Finan C, Gaulton A, Kruger FA, et al. (2017). The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* 9(383). doi:10.1126/scitranslmed.aag1166

223. Peng K, Jin L, Niu YD, et al. (2017). Condensed tannins affect bacterial and fungal microbiomes and mycotoxin production during ensiling and upon aerobic exposure. *Appl Environ Microbiol*. doi:10.1128/AEM.02274-17

224. Chen Y, Farrer RA, Giamberardino C, et al. (2017). Microevolution of Serial Clinical Isolates of *Cryptococcus neoformans* var. *grubii* and *C. gattii*. *MBio* 8(2). doi:10.1128/mbio.00166-17

225. Ohmann C, Banzi R, Canham S, et al. (2017). Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* 7(12). doi:10.1136/bmjopen-2017-018647

226. Kraiczyn J, Nayak KM, Howell KJ, et al. (2017). DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. *Gut*. doi:10.1136/gutjnl-2017-314817

227. Marioni JC, Arendt D (2017). How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu Rev Cell Dev Biol*. doi:10.1146/annurev-cellbio-100616-060818

228. Meehan TF, Conte N, Goldstein T, et al. (2017). PDX-MI: Minimal Information for Patient-Derived Tumor Xenograft Models. *Cancer Res* 77(21):e62-e66. doi:10.1158/0008-5472.CAN-17-0582

229. Garcia-Alonso LM, Iorio F, Matchan A, et al. (2017). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res*. doi:10.1158/0008-5472.CAN-17-1679

230. Eduati F, Doldàn-Martelli V, Klinger B, et al. (2017). Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models. *Cancer Res* 1-34. doi:10.1158/0008-5472.can-17-0078

231. Quispel-Janssen JM, Badhai J, Schunselaar L, et al. (2017). Comprehensive pharmacogenomic profiling of malignant pleural mesothelioma identifies a subgroup sensitive to FGFR inhibition. *Clin Cancer Res*. doi:10.1158/1078-0432.CCR-17-1172

232. Harrison PW, Montgomery SH (2017). Genetics of Cerebellar and Neocortical Expansion in Anthropoid Primates: A Comparative Approach. *Brain Behav Evol* 89(4):274-285. doi:10.1159/000477432

233. Fabregat A, Sidiropoulos K, Viteri G, et al. (2017). Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18(1):142. doi:10.1186/s12859-017-1559-2

234. Ong E, Sarntivijai S, Jupp S, et al. (2017). Comparison, alignment, and synchronization of cell line information between CLO and EFO. *BMC Bioinformatics* 18(Suppl 17): Paper presented at the 2017 International Conference on Biomedical Ontology (ICBO-2017), Newcastle, UK, 13 September 2017. doi:10.1186/s12859-017-1979-z
235. Richter S, Helm C, Meunier FA, et al. (2017). Comparative analyses of glycerotoxin expression unveil a novel structural organization of the bloodworm venom system. *BMC Evol Biol* 17(1):1-19. doi:10.1186/s12862-017-0904-4
236. Zuber V, Bettella F, Witoelar A, et al. (2017). Bromodomain protein 4 discriminates tissue-specific super-enhancers containing disease-specific susceptibility loci in prostate and breast cancer. *BMC Genomics* 18(1):270-281. doi:10.1186/s12864-017-3620-y
237. Wong YC, Teh HF, Mebus K, et al. (2017). Differential gene expression at different stages of mesocarp development in high- and low-yielding oil palm. *BMC Genomics* 18(1):470. doi:10.1186/s12864-017-3855-7
238. Griffiths JA, Scialdone A, Marioni JC (2017). Mosaic autosomal aneuploidies are detectable from single-cell RNAseq data. *BMC Genomics* 18(1). doi:10.1186/s12864-017-4253-x
239. Martín-Gálvez D, Dunoyer de Segonzac D, Ma MCJ, et al. (2017). Genome variation and conserved regulation identify genomic regions responsible for strain specific phenotypes in rat. *BMC Genomics* 18(1). doi:10.1186/s12864-017-4351-9
240. Brandizi M, Melnichuk O, Bild R, et al. (2017). Orchestrating differential data access for translational research: a pilot implementation. *BMC Med Inform Decis Mak* 17(1):30. doi:10.1186/s12911-017-0424-6
241. Ferrero E, Dunham I, Sanseau P (2017). In silico prediction of novel therapeutic targets using gene-disease association data. *J Transl Med* 15(1). doi:10.1186/s12967-017-1285-6
242. Ecker S, Chen L, Pancaldi V, et al. (2017). Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol* 18(1):18. doi:10.1186/s13059-017-1156-8
243. Cheung WA, Shao X, Morin A, et al. (2017). Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol* 18(1):50. doi:10.1186/s13059-017-1173-7
244. Angermueller C, Lee HJ, Reik W, et al. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* 18(1):1-13. doi:10.1186/s13059-017-1189-z
245. Vieira Braga FA, Teichmann SA, Stubbington MJ (2017). Are cells from a snowman realistic? Cryopreserved tissues as a source for single-cell RNA-sequencing experiments. *Genome Biol* 18(1). doi:10.1186/s13059-017-1192-4
246. Stubbs TM, Bonder MJ, Stark AK, et al. (2017). Multi-tissue DNA methylation age predictor in mouse. *Genome Biol* 18(1):1-14. doi:10.1186/s13059-017-1203-5
247. Buettner F, Pratanwanich N, McCarthy DJ, et al. (2017). f-sclVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 18(1):212. doi:10.1186/s13059-017-1334-8
248. Steward CA, Parker APJ, Minassian BA, et al. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med* 9(1). doi:10.1186/s13073-017-0441-1
249. Lenselink EB, Ten Dijke N, Bongers B, et al. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform* 9(1). doi:10.1186/s13321-017-0232-0
250. Osumi-Sutherland D, Courtot M, Balhoff JP, et al. (2017). Dead simple OWL design patterns. *J Biomed Semantics* 8(1). doi:10.1186/s13326-017-0126-0
251. Kafkas Ş, Dunham I, McEntyre J (2017). Literature evidence in open targets – a target validation platform. *J Biomed Semantics* 8(1):20. doi:10.1186/s13326-017-0131-3
252. Roncaglia P, van Dam TJP, Christie KR, et al. (2017). The Gene Ontology of eukaryotic cilia and flagella. *Cilia* 6. doi:10.1186/s13630-017-0054-8
253. Baud A, Flint J (2017). Identifying genes for neurobehavioural traits in rodents: progress and pitfalls. *Dis Model Mech* 10(4):373-383. doi:10.1242/dmm.027789
254. Gutierrez-Vazquez C, Enright AJ, Rodríguez-Galán A, et al. (2017). 3'uridylation controls mature microRNA turnover during CD4 T cell activation. *RNA*. doi:10.1261/rna.060095.116
255. Gonzalez G, Hardwick S, Maslen SL, et al. (2017). Structure of the *Escherichia coli* ProQ RNA chaperone protein. *RNA* 23(5):696-711. doi: 10.1261/rna.060343.116
256. Panni S, Prakash A, Bateman A, et al. (2017). Yeast non-coding RNA interaction network. *RNA*. doi:10.1261/rna.060996.117
257. Robertsen EM, Denise H, Mitchell A, et al. (2017). ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services. *F1000Res* 6. doi:10.12688/f1000research.10443.1
258. Jain M, Tyson JR, Loose M, et al. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res* 6. doi:10.12688/f1000research.11354.1
259. Jiménez RC, Kuzak M, Alhamdoosh M, et al. (2017). Four simple recommendations to encourage best practices in research software. *F1000Res* 6. doi:10.12688/f1000research.11407.1
260. Vizcaíno JA, Walzer M, Jiménez RC, et al. (2017). A community proposal to integrate proteomics activities in ELIXIR. *F1000Res* 6. doi:10.12688/f1000research.11751.1
261. Zhang C, Bijlard J, Staiger C, et al. (2017). Systematically linking tranSMART, Galaxy and EGA for reusing human translational research data. *F1000Res* 6. doi:10.12688/f1000research.12168.1
262. Morgan SL, Palagi PM, Fernandes PL, et al. (2017). The ELIXIR-EXCELERATE Train-the-Trainer pilot programme: empower researchers to deliver high-quality training. *F1000Res* 6. doi:10.12688/f1000research.12332.1
263. Van Rijswijk M, Beirnaert C, Caron C, et al. (2017). The future of metabolomics in ELIXIR. *F1000Res* 6. doi:10.12688/f1000research.12342.1
264. Van Rijswijk M, Beirnaert C, Caron C, et al. (2017). The future of metabolomics in ELIXIR. *F1000Res* 6. doi:10.12688/f1000research.12342.2
265. McMurtry JA, Juty N, Blomberg N, et al. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol* 15(6). doi:10.1371/journal.pbio.2001414
266. Thiel C, Cordes H, Fabbri L, et al. (2017). A Comparative Analysis of Drug-Induced Hepatotoxicity in Clinically Relevant Situations. *PLoS Comput Biol* 13(2):e1005280. doi:10.1371/journal.pcbi.1005280
267. Gonçalves E, Raguz Nakic Z, Zampieri M, et al. (2017). Systematic Analysis of Transcriptional and Post-transcriptional Regulation of Metabolism in Yeast. *PLoS Comput Biol* 13(1):e1005297. doi:10.1371/journal.pcbi.1005297
268. Henriques D, Villaverde AF, Rocha M, et al. (2017). Data-driven reverse engineering of signaling pathways using ensembles of dynamic models. *PLoS Comput Biol*. 13(2):e1005379. doi:10.1371/journal.pcbi.1005379
269. Emery LR, Morgan SL (2017). The application of project-based learning in bioinformatics training. *PLoS Comput Biol* 13(8). doi:10.1371/journal.pcbi.1005620
270. Baud A, Mulligan MK, Casale FP, et al. (2017). Genetic Variation in the Social Environment Contributes to Health and Disease. *PLoS Genet* 13(1):e1006498. doi:10.1371/journal.pgen.1006498
271. Kerr F, Sofola-Adesakin O, Ivanov DK, et al. (2017). Direct Keap1-Nrf2 disruption as a potential therapeutic target for Alzheimer's disease. *PLoS Genet* 13(3):e1006593. doi:10.1371/journal.pgen.1006593
272. Tong P, Monahan J, Prendergast JG (2017). Shared regulatory sites are abundant in the human genome and shed light on genome evolution and disease pleiotropy. *PLoS Genet* 13(3). doi:10.1371/journal.pgen.1006673
273. Casale FP, Horta D, Rakitsch B, et al. (2017). Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS Genet* 13(4):1-27. doi:10.1371/journal.pgen.1006693
274. Protasio AV, van Dongen S, Collins J, et al. (2017). MiR-277/4989 regulate transcriptional landscape during juvenile to adult transition in the parasitic helminth *Schistosoma mansoni*. *PLoS Negl Trop Dis* 11(5). doi:10.1371/journal.pntd.0005559
275. Hulo C, Masson P, de Castro E, et al. (2017). The ins and outs of eukaryotic viruses: Knowledge base and ontology of a viral infection. *PLoS ONE* 12(2):1-17. doi:10.1371/journal.pone.0171746
276. Penas DR, Henriques D, González P, et al. (2017). A parallel metaheuristic for large mixed-integer dynamic optimization problems, with applications in computational biology. *PLoS ONE* 12(8). doi:10.1371/journal.pone.0182186
277. Perez-Riverol Y, Kuhn M, Vizcaíno JA, et al. (2017). Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS ONE* 12(12). doi:10.1371/journal.pone.0189875
278. Ferreira JD, Inácio B, Salek RM, et al. (2017). Assessing Public Metabolomics Metadata, Towards Improving Quality. *J Integr Bioinform* 14(4). doi:10.1515/jib-2017-0054
279. Davis MP, Carrieri C, Saini HK, et al. (2017). Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. *EMBO Rep* 18(5):1-17. doi:10.15252/embr.201744059
280. Vamathevan J, Birney E (2017). A Review of Recent Advances in Translational Bioinformatics: Bridges from Biology to Medicine. *Yearb Med Inform* 26(1):178-187. doi:10.15265/IY-2017-017
281. Imprialou M, Kahles A, Steffen JG, et al. (2017). Genomic Rearrangements in Arabidopsis Considered as Quantitative Traits. *Genetics* 205(4):1425-1441. doi:10.1534/genetics.116.192823
282. Morgan AP, Gatti DM, Najarian ML, et al. (2017). Structural Variation Shapes the Landscape of Recombination in Mouse. *Genetics* 206(2):603-619. doi:10.1534/genetics.116.197988
283. Cavaliere D, Di Paola M, Rizzetto L, et al. (2017). Genomic and Phenotypic Variation in Morphogenetic Networks of Two *Candida albicans* Isolates Subtends Their Different Pathogenic Potential. *Front Immunol* 8:1997. doi:10.3389/fimmu.2017.01997
284. Tkavc R, Matrosova VY, Grichenko OE, et al. (2017). Prospects for Fungal Bioremediation of Acidic Radioactive Waste Sites: Characterization and Genome Sequence of *Rhodotorula taiwanensis* MD1149. *Front Microbiol* 8. doi:10.3389/fmicb.2017.02528
285. Mugumbate G, Mendes V, Blaszczyk M, et al. (2017). Target Identification of *Mycobacterium tuberculosis* Phenotypic Hits Using a Concerted Chemogenomic, Biophysical, and Structural Approach. *Front Pharmacol* 8:681. doi:10.3389/fphar.2017.00681
286. Patwardhan A, Brandt R, Butcher SJ, et al. (2017). Building bridges between cellular and molecular structural biology. *eLife* 6:e25835. doi:10.7554/eLife.25835
287. Nikolic MZ, Carltg O, Jeng Q, et al. (2017). Human embryonic lung epithelial tips are multipotent progenitors that can be expanded in vitro as long-term self-renewing organoids. *eLife* 6:e26575. doi:10.7554/eLife.26575.001
288. Regev A, Teichmann SA, Lander ES, et al. (2017). Science Forum: The Human Cell Atlas. *eLife* 6:e270416. doi:10.7554/eLife.27041
289. Putzbach W, Gao QQ, Patel M, et al. (2017). Many si/shRNAs can kill cancer cells by targeting multiple survival genes through an off-target mechanism. *eLife* 6:e29702. doi:10.7554/eLife.29702
290. White RJ, Collins JE, Sealy IM, et al. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. *eLife* 6:e30860. doi:10.7554/eLife.30860
291. Galardini M, Koumoutsis A, Herrera-Dominguez L, et al. (2017). Phenotype inference in an *Escherichia coli* strain panel. *eLife* 6:e31035. doi:10.7554/eLife.31035
292. Huang Y, Kim JK, Do DV, et al. (2017). STELLA modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *eLife* 6:e22345. doi:10.7554/elife.22345

Scientific Advisory Committees

BioModels Scientific Advisory Committee

- ⊙ Carole Goble, University of Manchester, United Kingdom
- ⊙ Thomas Lemberger, Nature Publishing Group/EMBO
- ⊙ Pedro Mendes, University of Manchester, United Kingdom
- ⊙ Wolfgang Mueller, HTS, Germany
- ⊙ Philippe Sanseau, GSK, United Kingdom

Chemistry Services Scientific Advisory Committee

- ⊙ Jildau Bouwman, TNO Innovation for Life, the Netherlands
- ⊙ Steve Bryant, NIH, United States
- ⊙ Edgar Jacoby, Novartis, Switzerland
- ⊙ Gabi Kastenmueller, Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Munchen, Germany
- ⊙ Tudor Oprea, University of New Mexico, United States
- ⊙ Alfonso Valencia, CNIO, Spain
- ⊙ Val Gillet, University of Sheffield, United Kingdom
- ⊙ Michel Dumontier, Maastricht University, the Netherlands

Ensembl Scientific Advisory Board

- ⊙ Wendy Bickmore, MRC Institute of Genetics and Molecular Medicine at the University of Edinburgh, United Kingdom
- ⊙ Deanna Church, 10X Genomics, United States
- ⊙ Federica Di Palma, Earlham Institute, United Kingdom
- ⊙ Mark Diekhans, Center for Biomolecular Science & Engineering, University of California Santa Cruz, United States
- ⊙ Nils Gehlenborg, Harvard Medical School, United States
- ⊙ Martien Groenen, Wageningen University, the Netherlands
- ⊙ Erich Jarvis, Howard Hughes Medical Institute, The Rockefeller University, United States
- ⊙ Felicity Jones, Friedrich Miescher Laboratory, Germany
- ⊙ Heidi Rehm, BWH and Harvard Medical School, Partners Laboratory for Molecular Medicine, United States of America

Ensembl Genomes Scientific Advisory Board

- ⊙ Dick de Ridder, University of Wageningen, the Netherlands
- ⊙ Anne-Francoise Adam-Blondin, URGI, INRA, France
- ⊙ Mario Caccamo, NIAB, United Kingdom
- ⊙ Alexander Goessman, Justus Leibig University, Germany
- ⊙ Inge Jonassen, University of Bergen, Norway
- ⊙ Mara Lawinczak, Wellcome Sanger Institute, United Kingdom
- ⊙ Claudine Medigue, Genoscope, France
- ⊙ Jason Stajich, University of Riverside, United States

European Genome-Phenome Archive (EGA) Scientific Advisory Board

- ⊙ Gil McVean, University of Oxford, United Kingdom
- ⊙ Teri Manolio, National Human Genome Research Institute, United States
- ⊙ Joaquín Dopazo, Centro de Investigación Príncipe Felipe, Spain
- ⊙ Dixie Baker, Martin-Blanc and Associates, United States
- ⊙ Anne Cambon-Thomsen, Institut National de la Santé et de la Recherche Médicale (INSERM), France
- ⊙ Jan-Willem Boiten, Dutch Techcentre for Life Sciences, the Netherlands

European Nucleotide Archive Scientific Advisory Board

- ⊙ Mark Blaxter, University of Edinburgh, United Kingdom
- ⊙ Alvis Brazma, EMBL-EBI, United Kingdom
- ⊙ Fiona Brinkman, Simon Fraser University, Canada
- ⊙ Antoine Danchin, CNRS, Institut Pasteur, France
- ⊙ Frank Oliver Glöckner, Max Planck Institute for Marine Microbiology, Germany
- ⊙ Tim Hubbard, King's College London, United Kingdom
- ⊙ Macha Nikolski, CNRS Bordeaux (CBiB), France
- ⊙ Babis Savakis, University of Crete and Alexander Fleming BSRC, Greece
- ⊙ Martin Vingron, Max-Planck Institute for Molecular Genetics, Germany
- ⊙ Patrick Wincker, Genoscope, France

Expression Atlas and ArrayExpress

- ⊙ Roderic Guigo Serra, Centre de Regulació Genòmica, Spain
- ⊙ Ruedi Aebersold, ETH, Switzerland
- ⊙ Jurg Bahler, University College London, United Kingdom
- ⊙ Angela Brooks, University of California Santa Cruz, United States
- ⊙ Kathryn Lilley, University of Cambridge, United Kingdom
- ⊙ Zemin Zhang, Peking University, China
- ⊙ Oliver Stegle, EMBL-EBI, United Kingdom
- ⊙ Wolfgang Huber, EMBL Heidelberg, Germany

GWAS Catalog Scientific Advisory Board

- ⊙ Nancy Cox, Vanderbilt University, United States
- ⊙ Josh Denny, Vanderbilt University, United States
- ⊙ Mike Feolo, National Center for Biotechnology Information, United States
- ⊙ Marylyn Ritchie, Pennsylvania State University, United States
- ⊙ Alexis Battle, Johns Hopkins University, United States
- ⊙ Ines Barroso, Wellcome Sanger Institute, United Kingdom

IntAct and Complex Portal

- ⊙ Pascal Braun, Technische Universität München, Germany
- ⊙ Alex Jones, University of Warwick, United Kingdom
- ⊙ Giovanni Cesareni, University of Rome Tor Vergata, Italy
- ⊙ Willem Ouwehand, University of Cambridge NHS Blood and Transplant Centre, United Kingdom
- ⊙ Peter Woollard, GSK, United Kingdom
- ⊙ Evangelia Petsalaki, EMBL-EBI, United Kingdom

International Genome Sample Resource (IGSR) Scientific Advisory Board

- ⊙ Eimear Kerry, Icahn School of Medicine at Mount Sinai, United States
- ⊙ Jan Korbel, EMBL Heidelberg, Germany
- ⊙ Richard Durbin, Wellcome Sanger Institute, United Kingdom

- ⊙ Piero Carninci, Riken, Japan
- ⊙ Jane Kaye, University of Oxford, United Kingdom
- ⊙ Cisca Wijmenga, University of Groningen, the Netherlands

International Nucleotide Sequence Database Collaboration (INSDC) International Advisory Committee

- ⊙ Mark Blaxter, University of Edinburgh, United Kingdom
- ⊙ Antoine Danchin, CNRS, Institut Pasteur, France
- ⊙ Tim Hubbard, King's College London, United Kingdom
- ⊙ Babis Savakis, University of Crete and IMBB-FORTH, Greece
- ⊙ Jean Weissenbach, Genoscope, France

InterPro/Pfam Scientific Advisory Board

- ⊙ Patrick Aloy, Institute for Research in Biomedicine, Spain
- ⊙ Michael Galperin, National Center for Biotechnology Information, United States
- ⊙ Nicola Mulder, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa
- ⊙ Sean Munro, MRC Laboratory of Molecular Biology, United Kingdom
- ⊙ Jennifer Potts, University of York, United Kingdom
- ⊙ Alfonso Valencia, Barcelona Supercomputing Centre, Spain

Europe PMC and Literature Services Cluster Scientific Advisory Board

- ⊙ Terry Attwood, University of Manchester, United Kingdom
- ⊙ Theo Bloom, British Medical Journal, United Kingdom
- ⊙ Jan Brasse, Göttingen State and University Library, Germany
- ⊙ Martin Fenner, DataCite, Germany
- ⊙ Jenny Malloy, University of Cambridge and ContentMine, United Kingdom
- ⊙ Patrick Ruch, University of Applied Sciences, Switzerland
- ⊙ Frank Uhlmann, The Francis Crick Institute, United Kingdom

Metagenomics (EBI Metagenomics) Scientific Advisory Board

- ◉ Mark Blaxter, University of Edinburgh, United Kingdom
- ◉ Chris Bowler, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), France
- ◉ Alice McHardy, Helmholtz-Zentrum für Infektionsforschung GmbH, Germany
- ◉ Eric Pelletier, Genoscope, France
- ◉ Katie Pollard, Gladstone Institutes UCSF, United States
- ◉ Phil Poole, University of Oxford, United Kingdom

Open Targets

- ◉ Bissan Al-Lazikani, The Institute of Cancer Research, United Kingdom
- ◉ Søren Brunak, Center for Biological Sequence Analysis - Department of Systems Biology, Technical University of Denmark, Denmark
- ◉ Robert Graham, OMNI Human Genetics, Genentech, United States
- ◉ Berent Prakken, Wilhelmina Children's Hospital UMC Utrecht, the Netherlands
- ◉ Robert Vries, Hubrecht Organoid Technology Foundation, HUB, the Netherlands

Proteomics Identifications Database (PRIDE)

- ◉ Ruedi Aebersold, Swiss Federal Institute of Technology, ETH, Switzerland
- ◉ Kathryn Lilley, University of Cambridge, United Kingdom
- ◉ Juri Rappsilber, University of Edinburgh, United Kingdom
- ◉ Pedro R. Cutillas, Queen Mary University, United Kingdom
- ◉ Hans Vissers, Waters Corporation, United Kingdom
- ◉ Jurgen Cox, Max Planck Institute of Biochemistry, Germany

Reactome Scientific Advisory Committee

- ◉ Russ Altman, Stanford University, United States
- ◉ Gary Bader, University of Toronto, Canada
- ◉ Fiona Brinkman, Simon Fraser University, Canada
- ◉ Melissa Haendel, Oregon Health and Science University, United States

- ◉ John Overington, Medicines Discovery Catapult, United Kingdom
- ◉ Jill Mesirov, Broad Institute of MIT and Harvard, United States
- ◉ Bill Pearson, University of Virginia, United States
- ◉ Brian Shoichet, University of California San Francisco, United States
- ◉ Josh Stuart, University of California, Santa Cruz, United States

RNA group (RNAcentral and Rfam) Scientific Advisory Board

- ◉ Sean Eddy, Harvard University, United States
- ◉ Eric Westhof, University of Strasbourg, France
- ◉ John Rinn, University of Colorado Boulder, United States
- ◉ Michele Meyer, Boston College, United States
- ◉ Michaela Zavolan, University of Basel, Switzerland
- ◉ Manja Marz, Friedrich Schiller University Jena, Germany

Technical Services Cluster Scientific Advisory Board

- ◉ Ewan Birney, EMBL-EBI, United Kingdom
- ◉ Rolf Apweiler, EMBL-EBI, United Kingdom
- ◉ Rupert Lueck, EMBL Heidelberg, Germany
- ◉ Anton Enright, EMBL-EBI, United Kingdom
- ◉ David Fergusson, University of Edinburgh Information Services Group, United Kingdom
- ◉ Nick Goldman, EMBL-EBI, United Kingdom
- ◉ Helen Parkinson, EMBL-EBI, United Kingdom
- ◉ Ugis Sarkans, EMBL-EBI, United Kingdom
- ◉ Andy Yates, EMBL-EBI, United Kingdom

External Training Programme Advisory Group

- ◉ Alex Bateman, EMBL-EBI, United Kingdom
- ◉ Rob Finn, EMBL-EBI, United Kingdom
- ◉ Helen Firth, Cambridge University Hospitals Trust, United Kingdom
- ◉ Mark Forster, Daresbury Laboratory, United Kingdom
- ◉ Nick Goldman, EMBL-EBI, United Kingdom
- ◉ Paul Kersey, EMBL-EBI, United Kingdom
- ◉ Andy Yates, EMBL-EBI, United Kingdom

- ◉ Gos Micklem, University of Cambridge, United Kingdom
- ◉ Nicky Mulder, University of Cape Town, South Africa
- ◉ Patricia Palagi, Swiss Institute of Bioinformatics, Switzerland
- ◉ Chris Ponting, University of Edinburgh, United Kingdom
- ◉ Rochelle Tractenberg, Georgetown University, United States
- ◉ Andrew White, Unilever, United Kingdom

UniProt: The Universal Protein Resource Scientific Advisory Board

- ◉ Russ Altman, Stanford University, United States
- ◉ Carol Bult, The Jackson Laboratory, Mouse Genome Informatics, United States
- ◉ Martin Ebeling, Roche Pharmaceutical Research and Early Development, Switzerland
- ◉ Fuchu He, Beijing Proteome Research Center, China
- ◉ Maricel Kann, University of Maryland, United States
- ◉ Edward Marcotte, University of Texas at Austin, United States
- ◉ Maryann E. Martone, National Center for Microscopy and Imaging Research, University of California, United States
- ◉ Lynne Regan, Yale University, United States
- ◉ Peter Robinson, The Jackson Laboratory for Genomic Medicine, United States
- ◉ Philippe Sanseau, GSK, United Kingdom
- ◉ Paul Thomas, University of Southern California, United States

Worldwide Protein Data Bank (wwPDB) Advisory Committee

- ◉ R. Andrew Byrd, National Institutes of Health, United States
- ◉ Paul Adams, Lawrence Berkeley Laboratory, United States
- ◉ Manju Bansal, Indian Institute of Science, India
- ◉ David Brown, University of Kent, United Kingdom
- ◉ Sarah Butcher, University of Helsinki, Finland
- ◉ Wah Chiu, Stanford University, United States
- ◉ Jianping Ding, Shanghai Institutes for Biological Sciences, China
- ◉ Arthur Edison, University of Georgia, United States
- ◉ Tsuyoshi Inoue, Osaka University, Japan
- ◉ Gaetano Montelione, Rutgers University, United States

- ◉ Edward N. Baker, University of Auckland, New Zealand
- ◉ Cynthia Wolberger, Howard Hughes Medical Institute, United States
- ◉ Kei Yura, Ochanomizu University, Japan

Molecular and Cellular Structure Cluster Scientific Advisory Committee

- ◉ Sarah Butcher, University of Helsinki, Finland
- ◉ Alexandre Bonvin, Utrecht University, the Netherlands
- ◉ David Brown, University of Kent, United Kingdom
- ◉ Lucy Collinson, The Francis Crick Institute, United Kingdom
- ◉ Manuela Helmer Citterich, University of Rome Tor Vergata, Italy
- ◉ Susan Lea, University of Oxford, United Kingdom
- ◉ Michael Nilges, Institut Pasteur, France
- ◉ Arwen Pearson, University of Hamburg, Germany

Major database collaborations

ArrayExpress

- ⦿ *Gene Expression Omnibus, National Center for Biotechnology Information, Bethesda, United States*

BioModels database

- ⦿ *California Institute of Technology, United States*
- ⦿ *Database of Quantitative Cellular Signalling, National Center for Biological Sciences, India*
- ⦿ *JWS Online, Stellenbosch University, South Africa*
- ⦿ *Physiome Model Repository, Auckland Bioengineering Institute, New Zealand*
- ⦿ *The Virtual Cell, University of Connecticut Health Center, United States*

ChEBI

- ⦿ *ChemIdPlus, National Library of Medicine, United States*
- ⦿ *DrugBank, University of Alberta, Canada*
- ⦿ *Immune Epitope Database (IEDB) at La Jolla Institute for Allergy and Immunology, United States*
- ⦿ *KEGG Compound, Kyoto University Bioinformatics Centre, Japan*
- ⦿ *OBI Ontology Consortium*
- ⦿ *PubChem, National Institutes of Health, United States*
- ⦿ *UniPathways, Swiss Institute of Bioinformatics, Switzerland*

ChEMBL

- ⦿ *BindingDB, University of California San Diego, United States*
- ⦿ *CanSAR, Institute of Cancer Research, London, United Kingdom*
- ⦿ *PubChem, NCBI, National Institutes of Health, United States*
- ⦿ *UCL Institute of Cardiovascular Science, United Kingdom*
- ⦿ *Illuminating the Druggable Genome Knowledge Management Centre, led out of University of New Mexico*

Ensembl

Here we list a selection of collaborations representing genome sequencing centres, groups providing genomics information resources and major international projects. There are many others.

- ⦿ *Baylor College of Medicine, United States*
- ⦿ *Broad Institute, United States*
- ⦿ *The Genome Institute, Washington University in St. Louis, United States*
- ⦿ *The Roslin Institute, United Kingdom*
- ⦿ *Wellcome Sanger Institute, United Kingdom*
- ⦿ *University of California Santa Cruz, United States*
- ⦿ *National Center for Biotechnology Information, United States*
- ⦿ *Mouse Genome Informatics at the Jackson Laboratory, United States*
- ⦿ *Rat Genome Database at the Medical College of Wisconsin, United States*
- ⦿ *Genotype-Tissue Expression (GTEx) Consortium*
- ⦿ *Genome Reference Consortium*
- ⦿ *Global Alliance for Genomics and Health*
- ⦿ *Genome 10K/ Vertebrate Genomes Project*

Ensembl Genomes

- ⦿ *Gramene at Cold Spring Harbor Laboratory, United States*
- ⦿ *PomBase with University College London and the University of Cambridge, United Kingdom*
- ⦿ *PhytoPath with Rothamsted Research, United Kingdom*
- ⦿ *VectorBase: a collaboration with University of Notre Dame, United States; Harvard University, United States; Institute of Molecular Biology and Biochemistry, Greece; University of New Mexico, United States; and Imperial College London, United Kingdom*
- ⦿ *Microme, a European collaboration with 14 partners*
- ⦿ *transPLANT, a European project with 11 partners*
- ⦿ *WormBase, a collaboration with the California Institute of Technology and Washington University, United States; Ontario Institute for Cancer Research, Canada; Wellcome Sanger Institute and Oxford University, United Kingdom*

European Genome-phoneme Archive (EGA)

- ⦿ *Database of Genotypes and Phenotypes (dbGaP), National Center for Biotechnology Information, United States*
- ⦿ *Japanese Genome-phenome Archive (JGA), DNA Data Bank of Japan, Japan*

The European Nucleotide Archive (ENA)

The ENA is part of the International Nucleotide Sequence Database Collaboration. Other partners include:

- ⦿ *National Center for Biotechnology Information, United States (GenBank, Trace Archive and Sequence Read Archive)*
- ⦿ *National Institute of Genetics, Japan (DNA DataBank of Japan, Trace Archive and Sequence Read Archive)*

Other ENA collaborations

- ⦿ *Catalogue of Life*
- ⦿ *Genomics Standards Consortium*
- ⦿ *Expression Atlas*
- ⦿ *Oregon State University, United States*
- ⦿ *Cold Spring Harbor Laboratory, United States*
- ⦿ *Wellcome Sanger Institute, United Kingdom*
- ⦿ *Gene Expression Omnibus, National Center for Biotechnology Information, United States*

Gene Ontology Consortium

- ⦿ *Agbase, Mississippi State University, United States*
- ⦿ *The Arabidopsis Information Resource, Carnegie Institution of Washington, United States*
- ⦿ *Berkeley Bioinformatics and Ontology Project, United States*
- ⦿ *British Heart Foundation, University College London, United Kingdom*
- ⦿ *Candida Genome Database, Stanford University, United States*
- ⦿ *DictyBase at Northwestern University, United States*
- ⦿ *EcoliWiki*
- ⦿ *FlyBase at the University of Cambridge, United Kingdom*
- ⦿ *GeneDB S. pombe and GeneDB for protozoa at the Wellcome Trust Sanger Institute, United Kingdom*

- ⦿ *Gramene at Cornell University, United States*
- ⦿ *Institute for Genome Sciences, University of Maryland, United States*
- ⦿ *The J. Craig Venter Institute, United States*
- ⦿ *Mouse Genome Informatics, The Jackson Laboratory, United States*
- ⦿ *Muscle TRAIT, University of Padua, Italy*
- ⦿ *Plant-Association Microbe Gene Ontology, Virginia Polytechnic Institute and State University, United States*
- ⦿ *Rat Genome Database at the Medical College of Wisconsin, United States*
- ⦿ *Reactome at Cold Spring Harbor Laboratory, United States*
- ⦿ *Saccharomyces Genome Database, Stanford University, United States*
- ⦿ *WormBase at California Institute of Technology, United States*
- ⦿ *The Zebrafish Information Network at the University of Oregon, United States*

IMEx Consortium

- ⦿ *Centro Nacional de Biotecnologia, Spain*
- ⦿ *DIP at the University of California, United States*
- ⦿ *University College London, United Kingdom*
- ⦿ *HPIDB at Mississippi State University, United States*
- ⦿ *I2D at Ontario Institute for Cancer Research, Canada*
- ⦿ *InnateDB at Simon Fraser University, Canada*
- ⦿ *MBInfo at National University of Singapore, Singapore*
- ⦿ *MINT at University Tor Vergata, Italy*
- ⦿ *Molecular Connections, India*
- ⦿ *UniProt/Swiss Institute of Bioinformatics, Switzerland*

InterPro

- ⦿ *CATH-Gene3D at University College London, United Kingdom*
- ⦿ *HAMAP at the Swiss Institute of Bioinformatics, Switzerland*
- ⦿ *PANTHER at University of Southern California, United States*
- ⦿ *Pfam at EMBL-EBI, United Kingdom*
- ⦿ *PIRSF at the Protein Information Resource, Georgetown University Medical Centre, United States*
- ⦿ *PRINTS at the University of Manchester, United Kingdom*

- ⦿ *ProDom at INRA and CNRS, France*
- ⦿ *PROSITE at the Swiss Institute of Bioinformatics, Switzerland*
- ⦿ *SMART at EMBL, Germany*
- ⦿ *SUPERFAMILY at the Laboratory of Molecular Biology, University of Cambridge, United Kingdom, United States*
- ⦿ *SLDF at the University of California, United States*
- ⦿ *CDD at the National Center for Biotechnology Information, United States*
- ⦿ *MobiDB at University of Padua, Italy*

MetaboLights

MetaboLights is part of the MetabolomeXchange (<http://www.metabolomexchange.org>). Other partners include:

- ⦿ *Metabolomics Workbench at UCSD, United States*
- ⦿ *Metabolomic Repository Bordeaux, France*
- ⦿ *Metabolonote, Japan*
- ⦿ *Leiden University, the Netherlands*

Protein Data Bank in Europe

PDBe is a partner in the World Wide Protein Data Bank (wwPDB). Other partners include:

- ⦿ *BioMagResBank, University of Wisconsin, Madison, United States*
- ⦿ *PDBj at Osaka University, Japan*
- ⦿ *Research Collaboratory for Structural Bioinformatics, United States*

PRIDE

PRIDE is a partner in the international ProteomeXchange Consortium of proteomics repositories. Other partners include:

- ⦿ *PeptideAtlas, Institute for Systems Biology, United States*
- ⦿ *MassIVE, University of California San Diego (UCSD), United States*
- ⦿ *jPOST, various institutions, Japan*
- ⦿ *iProX, Phoenix Center, China*

RNAcentral

- ⦿ *dictyBase, Northwestern University, United States*
- ⦿ *Greengenes Consortium, a collaboration with University of Colorado, United States; University of Queensland, Australia; Second Genome Inc, United States*
- ⦿ *GtRNADB, University of California Santa Cruz, United States*
- ⦿ *FlyBase, University of Cambridge, United Kingdom*
- ⦿ *LNCipedia, Ghent University, Belgium*
- ⦿ *lncRNADB, Garvan Institute of Medical Research, Australia*
- ⦿ *miRBase, University of Manchester, United Kingdom*
- ⦿ *Mouse Genome Informatics, the Jackson Laboratory, United States*
- ⦿ *Modomics, International Institute of Molecular and Cell Biology, Poland*
- ⦿ *NONCODE, Institute of Biophysics at Chinese Academy of Sciences, China*
- ⦿ *Pombase, University College London and the University of Cambridge, United Kingdom*
- ⦿ *Ribosomal Database Project, Michigan State University, United States*
- ⦿ *RefSeq, National Center for Biotechnology Information, United States*
- ⦿ *Saccharomyces Genome Database, Stanford University, United States*
- ⦿ *SILVA, Max Planck Institute for Marine Microbiology, Germany*
- ⦿ *snOPY, University of Miyazaki, Japan*
- ⦿ *SRPDB, University of Texas Health Science Center, United States*

- ⦿ *The Arabidopsis Information Resource (TAIR), Phoenix Bioinformatics Corporation, United States*
- ⦿ *tmRNA Website, Sandia National Laboratories, United States*
- ⦿ *Wormbase, a collaboration with the California Institute of Technology and Washington University, United States; Ontario Institute for Cancer Research, CA; Wellcome Sanger Institute and Oxford University, United Kingdom*

Europe PMC

Europe PubMed Central is part of PubMed Central International. Other database partners include:

- ⦿ *PubMed Central, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, United States*
- ⦿ *PubMed Central Canada, Canada*

Reactome

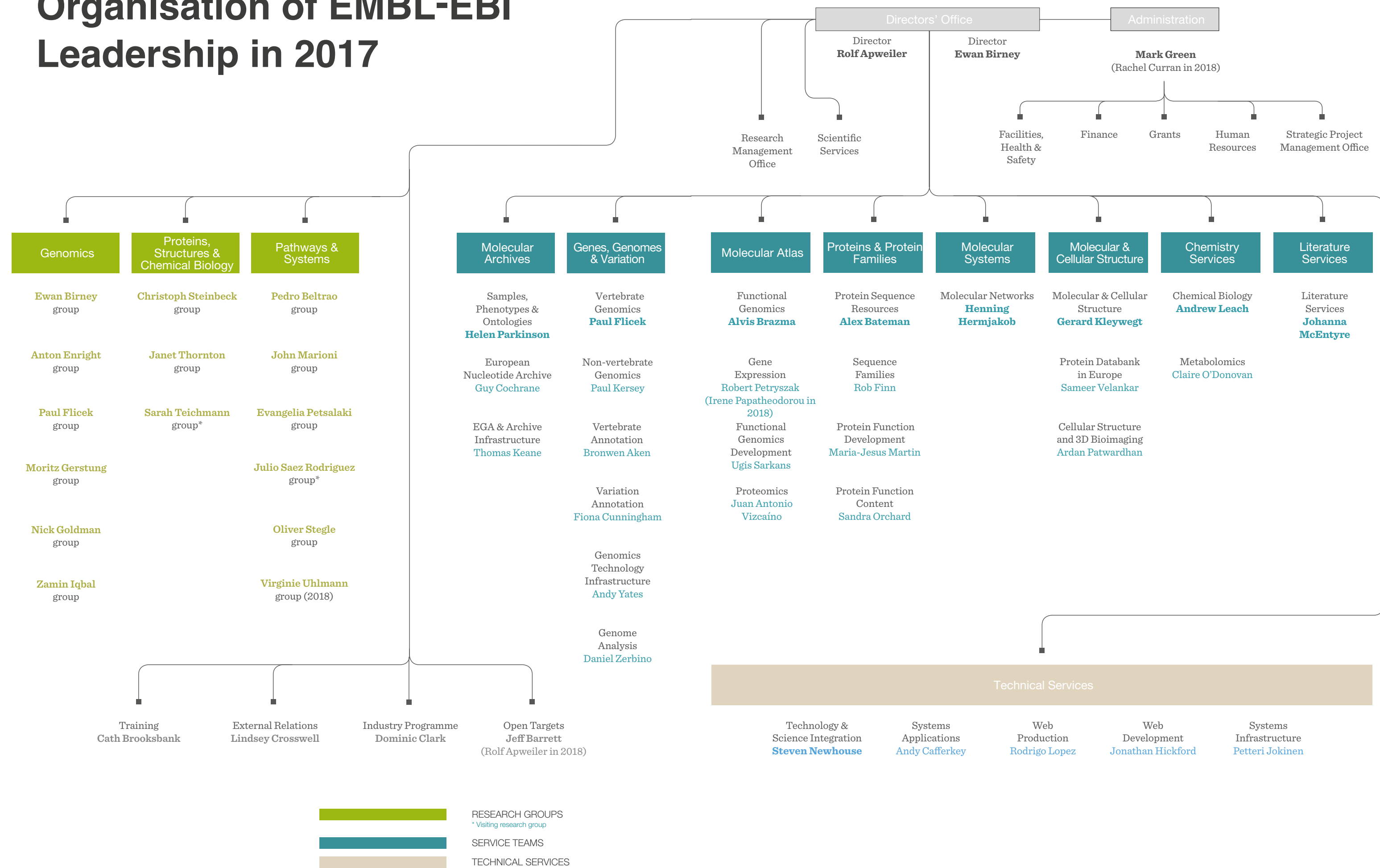
- ⦿ *New York University Langone Health, United States*
- ⦿ *Ontario Institute for Cancer Research, Canada*
- ⦿ *Oregon Health and Science University, United States*

UniProt: The Unified Protein Resource

UniProt at EMBL-EBI is part of the UniProt Consortium. Other partners include:

- ⦿ *UniProt, Protein Information Resource, Georgetown University Medical Centre, United States*
- ⦿ *UniProt, Protein Information Resource, University of Delaware, United States*
- ⦿ *UniProt, Swiss Institute of Bioinformatics, Switzerland*
- ⦿ *Gene Ontology Consortium*
- ⦿ *IMEx Consortium*

Organisation of EMBL-EBI Leadership in 2017








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

 @emblem
 /EMBLEBI
 /EMBLEBI

EMBL-EBI is a part of the European Molecular Biology Laboratory.
A digital version of this publication is available on
www.ebi.ac.uk/about/our-impact

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