The European Bioinformatics Institute · Cambridge

Annual Scientific Report 2015

Digital Edition

EMBL-EBI

On the cover:

The artwork in this publication, created by Spencer Phillips in EMBL-EBI's External Relations team, is inspired by the results of the EMBL-led *Tara* Oceans expeditions. The European Nucleotide Archive and the EBI Metagenomics data service provide different ways to access and analyse marine datasets, including the outputs of *Tara* Oceans and MicroB3's annual Ocean Sampling Day.

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

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

Contents

Foreword

We are pleased to present EMBL-EBI's 2015 Annual Scientific Report, which showcases the scope of our activities and highlights progress made in our core mission areas.

The most important change at EMBL-EBI in 2015 was undoubtedly the transition of leadership from Janet Thornton, who stewarded the institute through a time of incredible change and growth, to the two of us, each of whom brings something different to the role of Director. Janet has been the best of leaders, and a mentor to many of today's leading lights in molecular biology – she will be a very tough act to follow. We welcome input from the community as we settle in to our new roles.

A large-scale economic analysis of EMBL-EBI's impact, carried out in 2015, put some impressive numbers to the value of open data in the life sciences. The many communities that share their data, through us and our partners, are making contributions that have scientific and economic impacts over the long term, and it is thanks to their collective efforts that we can do our part to facilitate innovation and keep science moving forward.

The impact of EMBL-EBI is made possible by our funders, including EMBL member states, the US National Institutes of Health, the Wellcome Trust, the European Commission, our Industry Programme members, UK Research Councils and many others. Their continued support enabled the 2014 launch of a public–private initiative, now called Open Targets, to speed up the discovery of new medicines. In 2015 we were pleased to welcome Biogen to the collaboration and proud to launch the Target Validation platform, a data resource that helps wet- and dry-lab researchers identify evidence of an association between a target and disease.

Taken together, the major achievements of our colleagues over the past year reflect our consistent, long-standing commitment to delivering services, research and training required by an increasingly data-driven life-science community. From deciphering samples of unknown life forms in the world's oceans to unpicking the subtle shifts in expression that lead to human disease, we are at the middle of it all. It is a time of breathtakingly rapid technological change, and we are committed to doing whatever it takes to support scientists pushing new frontiers, delving into the workings of life to discover solutions that will benefit us all.

Sincerely,

Rolf Apweiler, Joint Director

Ewan Birney, Joint Director

Major Achievements in 2015

In 2015 Janet Thornton passed the torch of leadership of EMBL-EBI to Rolf Apweiler and Ewan Birney after 14 years of service to the institute. The handover was celebrated in a major, all-staff event held in June, which brought our community together for a bit of friendly competition and an address by EMBL Director General Iain Mattaj about the major impact of one leading scientist on the world of molecular biology. Janet steered the organisation through times of rapid, exciting change in the world of molecular and computational biology, expanding EMBL-EBI's programmes to serve and explore all aspects of the life sciences and stewarding the ELIXIR data infrastructure to its launch.

As Rolf Apweiler and Ewan Birney took up their new posts, the leadership landscape changed in a number of ways. We bid a sad farewell to ChEMBL team leader John Overington, who moved to the London-based start-up company Stratified Medical, and have been very grateful to Anne Hersey for stepping in to lead the ChEMBL team after his departure. We also said goodbye to Research Group Leader Julio Saez-Rodriguez, who moved to RTWH Aachen in Germany, and to Variation team leader Justin Paschall, who joined the University of California Berkeley. We were pleased to welcome several new team leaders: Bronwen Aken, Andy Yates and Daniel Zerbino, who began to lead different areas of work in vertebrate genomics; Sandra Orchard, who started a new team in molecular interactions; and Robert Petryszak, whose new team focuses on gene expression. We were pleased to welcome Moritz Gerstung, who joined us from the Wellcome Trust Sanger Institute as a new Research Group leader in cancer genomics.

The new Ashburner Library was officially inaugurated by Michael Ashburner, a founder of EMBL-EBI and a leading light in genetics research. The colourful facility is located in the newly refurbished Shared Facilities, which also houses an exploration space for the Campus Public Engagement team.

Rolf Apweiler, Janet Thornton and Ewan Birney, June 2015, in front of a cake of historic significance.

Our impact

In 2015 our newly formed Strategic Project Management Office worked with a management consultancy, Charles Beagrie Ltd., to facilitate a large-scale, economic analysis of the institute's impact on research practice and the global economy. This work, encouraged by the BBSRC, fed into a new framework for impact assessment. The report included a survey of over 4000 data-service users, 45% of whom indicated that they could neither have created nor collected themselves the last data they used, nor obtained it elsewhere. The findings demonstrate the vital role of public databases in life-science research, and indicate that for every million pounds invested in EMBL-EBI, roughly 20 million pounds is returned to the global economy.

The impact of the institute is made possible by our funders, including EMBL member states and the Research Councils UK, and is generally framed in the context of our relevance to biomedical research. In 2015 we welcomed a delegation of EMBL-EBI funders on board the schooner Tara in London, on its way to the Paris meeting on climate change. The event raised

awareness amongst some of our strongest supporters that our research, services and training have an impact on all the life sciences, including explorations of unknown life in the world's oceans and soil.

Industry, innovation and translation

Originally formed by GSK, the Wellcome Trust Sanger Institute and EMBL-EBI, Open Targets (formerly the Centre for Therapeutic Target Validation, CTTV) fosters deep, on-going interactions between academic and industry members. Its purpose is to develop open, transformative approaches to selecting and validating novel targets in drug development. In 2015 the partnership launched the Target Validation platform, helping scientists discover and prioritise evidence-based relationships between targets and diseases. At its launch the service provided evidence for over 21 800 potential therapeutic targets spanning more than 8800 diseases and phenotypes. The design of this comprehensive data service was based on user-experience research, ensuring it is fit for purpose for wet- and dry-lab scientists alike, and in its first six weeks had over 9000 visits.

Following the interim directorship of Ewan Birney, Jeffrey Barrett of the Sanger Institute was appointed Director of the partnership in 2015. In all, 60 EMBL-EBI staff now work on some 30 Open Targets projects, which range from computational pipelines to oncology, induced-pluripotent stem cells and single-cell genomics.

We were very pleased to welcome Astex Pharmaceuticals, part of Otsuka Pharmaceuticals, as a new member of the Industry Programme in December 2015. Our Programme provides neutral ground for bioinformatics specialists in 22 large companies to meet and address shared challenges.

CERN openlab, a unique public–private partnership between CERN and leading ICT companies, welcomed EMBL-EBI as one of its first public-sector members. CERN openlab's mission is to accelerate the development of innovative ICT solutions that help the increasingly data-driven scientific community.

Services

New technologies are enabling the generation of vast quantities of data in publicly funded research, and most funders require these outputs to be made available in public data repositories such as those at EMBL-EBI. Our storage capacity has grown steadily, but the growth in nucleotide and proteomics data generation has been dramatic. Compression techniques such as CRAM, which was rolled out for all depositors in 2015, resolve the issue of handling nucleotide data on a very large scale, and we continue to work on novel methods of compression.

The use of our resources has grown rapidly as well: to take just one example, the Ensembl REST API alone served over 70 million requests in 2015. Every month, the EMBL-EBI domain handled over 560 million requests and delivered over 12.6 million 'jobs' for our users.

Genes, genomes and variation

The rapid, early sharing of pathogen surveillance data and related information is crucial during disease outbreaks, and in 2015 the European Nucleotide Archive (ENA) facilitated such sharing in the EU-funded COMPARE project by developing new 'Data Hubs' and establishing a cloud-compute environment. These and future developments will make it easier for researchers to track and rapidly share information about viruses such as Ebola and Zika.

EBI Metagenomics, which contains sequencing data from environmental samples, grew by leaps and bounds in 2015, perhaps most notably with the addition of datasets from the EMBL-led Tara Oceans expedition.

Invitation to an event for policymakers and funders on board the research schooner Tara in London, 2015.

The 1000 Genomes Project, the most comprehensive, fully open survey of human genetic variation ever performed, wrapped up in 2015. The findings continue to provide fascinating insights into the genetic differences between individuals, and a baseline for studies into how genetic changes can cause disease. Perhaps the most important legacy of the project is the methods and technological innovations that made the work possible, and transformed how genetic and genomic research is done around the world. The project's ~2500 datasets were updated to reflect the latest human reference assembly, published and made available through Ensembl and the European Variation Archive (EVA).

In addition to handling a 10-fold increase in fungal genomes and thousands of new bacterial genomes, Ensembl Genomes issued the most complete assembly of the bread-wheat genome ever released publicly. This dataset provides an invaluable resource for research into this essential staple food crop.

In 2015 the Genome-Wide Association Study (GWAS) Catalog was relaunched at EMBL-EBI, based on a completely new infrastructure. The resource, a collaboration with the National Human Genome Research Institute (NHGRI) at the US National Institutes of Health, now provides enriched, ontology-driven search capabilities and improved visualisation of complex information.

Expression

In the RNA- and protein-expression space, we are working towards a comprehensive, integrated, scalable atlas of expression, drawing on all available high-quality information. RNAcentral welcomed 12 new expert databases, broadening the range of RNA information offered through this integrating portal. Studies based on RNA sequencing began to appear in the literature en masse in 2015, and a substantial number of these datasets were shared through EMBL-EBI resources.

Our Baseline Expression Atlas released its first large-scale proteomics dataset for protein expression in human tissues. PRIDE handled significant growth, made its proteomics datasets available through a new website and Web Service, and offered a new dataset-discovery tool that improves data integration for ProteomeXchange.

Proteins

The incredibly fast, sensitive HMMER search tool, developed at the Janelia Research Campus in the US, was made available through an open-source website at EMBL-EBI in 2015. HMMER helps users infer the function of a protein and its evolutionary

history, and has been incorporated into several of our protein services, including Pfam and InterPro.

Data visualisation and integration are extremely useful for researchers trying to get to grips with large amounts of biological information, and in 2015 the UniProt collaboration developed a feature viewer that makes it easier to explore sequence feature annotations from the Knowledgebase. In addition, the re-launched Enzyme portal, built on user-experience research, integrates all information about enzymes from EMBL-EBI's resources, presenting the information in an intuitive interface.

Bringing structure to biology

Three-dimensional Electron Microscopy (3D EM), named "Technology of the Year" by *Nature Methods* in 2015, produces massive datasets that require processing to optimise their future utility. The Protein Data Bank in Europe (PDBe) team supported the 3D EM user community during this phase of very rapid development, when high-resolution structures began to enter the public archives.

PDBe's new, responsive website raises the bar for the delivery of 3D macromolecular structure information. The mobile-friendly site offers a BioSolr-based search system, as well as new categories of value-added information and images that provide rich insights into the sequence and structure of molecules and cells.
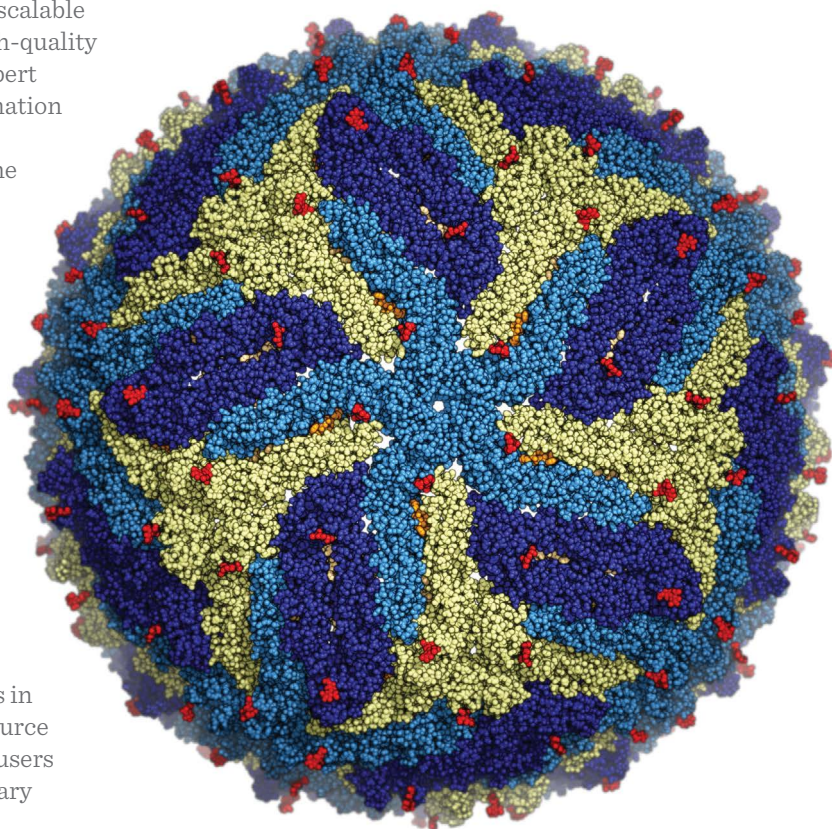
On 31 March 2016, the 3D structure of Zika virus, determined in near-atomic detail by cryo-electron microscopy (EM), was announced in the journal Science. *The structure is freely and publicly available through the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB) with the identifier 5ire.*

Chemical biology

The 17 million novel chemical entities annotated in SureChEMBL can now be searched using an API developed by the IMI-funded Open PHACTS project, and users can maintain a regular stream of this patent data behind a firewall to integrate with their own datasets.

Metabolomics continues to be one of the fastest areas of data growth at EMBL-EBI. The redesigned MetaboLights resource now offers improved search, more flexibility for data submissions and features that accommodate a huge range of taxonomy sources.

Molecular systems

Resolving an important challenge in sharing pharmacometric models of disease, the Innovative Medicines Initiative (IMI)-funded Drug Disease Model Resources (DDMoRe) project, in which our BioModels team plays a key role, launched a flexible format for exchanging computational models: the Pharmacometrics Markup Language, PharmML. PharmML-encoded models can be deposited in the DDMoRe model repository, facilitating collaboration on models to improve the design of cost–effective, reliable clinical trials of new and repurposed drugs.

Reactome updated its pathway diagram viewer, vastly improving the user experience and making it easier to incorporate the tool into third-party applications. The Complex Portal also features a new, innovative graphical tool for visualising complex topology and stoichiometry.

Cross-domain tools and resources

In 2015 Europe PMC reached 3.5 million full-text research articles, text-mined to integrate 19 accession number types with the literature. The Target Validation platform launched with new phenotype–disease annotation representation and gene–disease associations text mined from the literature to enable the integration of rare and common diseases according to shared phenotype.

In 2015 the International Mouse Phenotyping Consortium (IMPC) informatics platform, named as one of five case studies in a G7 report on global research infrastructures, played an important role in standardising high-throughput phenotyping data, which is pivotal to bridging mouse biology and precision medicine. EMBL-EBI also participated in the BioSolr project, which among other achievements provided ontology-enabled search for the European Bank for induced pluripotent Stem Cells (EBiSC) project.

The BioStudies database was launched in 2015, providing a new home for studies and associated files that do not quite fit with traditionally structured archives. We also launched a prototype BioSamples system that links information from public resources such as the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) hub. BioSamples now offers data on over 4 million samples.

Enabling research

Our service teams provide technical solutions that make it possible for translational and biomedical research initiatives to meet their objectives. For example, in 2015 we delivered data-management solutions for the EU-AIMS project on autism spectrum disorder, and built a cloud environment for molecular data in a project for the reuse of patient health records in clinical research, the European Medical Information Framework (EMIF).

We also work to strengthen agricultural and environmental science communities. During 2015 we helped establish community standards for marine data reporting and service provision, which makes it possible for different communities to pool their findings and re-examine them in a wider context. We also led on data and metadata standards for the launch of the Functional Annotation of Animal Genomes (FAANG) project, which aims to identify all functional elements in animal genomes.

We also helped to set minimum standards for genome annotation, which makes it easier for researchers to identify relevant information in public genome and proteome datasets. In the context of the COSMOS project, we helped develop the NMR mark-up language open standard for Nuclear Magnetic Resonance data and launched MetabolomeXchange, a platform for sharing metabolomics data. These efforts are essential to progress in 'omics studies.

Training

EMBL-EBI delivers a comprehensive range of bioinformatics training to help the global research community keep pace with rapid technological development. In 2015, over 100 people at EMBL-EBI were involved in training and scientific outreach, orchestrating and contributing to over 250 events throughout the world. These included face-to-face courses and workshops, demonstrations at conferences and web-based instruction. In addition, we delivered 10 member-selected workshops for our Industry Programme members.

In 2015 we participated in a number of UK National Health Service initiatives, for example informing a Master's syllabus for Genomic Medicine and co-authoring a report with Health Education England on how best to support the introduction of genomic technologies within the UK healthcare system through advanced training.

In 2015 our research leaders mentored and trained 30 EMBL International PhD Programme students, welcoming five newcomers. Three students successfully defended their theses: Ewan Johnstone, Nils Koelling and Michael Menden.

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, multi-dimensional statistical analysis and data-driven biological discovery, from plant biology to mammalian development and disease. Our groups are highly collaborative, and publish high-impact works on sequence and structural alignment, genome analysis, basic biological breakthroughs, algorithms and methods of widespread importance.

In 2015 the Teichmann group, with colleagues at the Cavendish Laboratory, created a new 'periodic table' of protein complexes that provides a unified way to classify and visualise protein complexes. The Table, published in *Science*, offers a new way of looking at the enormous variety of structures that proteins can build in nature, which ones might be discovered next, and predicting how entirely novel structures could be engineered.

New research by the Flicek group and their colleagues at the Cancer Research UK–Cambridge Institute shows how evolution has given rise to a rich diversity of species by repurposing functional elements shared by all mammals. Their study of gene regulation in 20 mammals, published in *Cell*, provides insights into how species diverged millions of years ago, and adapts methods and tools for genetic analysis of humans and mice to the study non-model species, such as whales.

Analysing associations between many different genetic variants and multiple traits amongst hundreds of thousands of individuals is extremely challenging, but a new algorithm developed by the Stegle group in 2015 makes it much easier. Published in *Nature Methods*, the method makes it possible to perform genetic analysis of up to 500,000 individuals – and many traits – at the same time.

A new method developed by the Marioni and Stegle groups for analysing RNA sequence data allows researchers to identify new subtypes of cells, creating order out of seeming chaos. Published in *Nature Biotechnology*, the protocol clarifies the true differences and similarities between cells, helping scientists use known molecular pathways to better understand cancer cells, differentiation processes and the pathogenesis of diseases.

Understanding the biological signalling pathways that regulate metabolism and gene expression is challenging, because so many things are happening at once. The

Saez-Rodriguez group, in collaboration with colleagues at Barts Cancer Institute, developed a new method for studying the targets and effects of cancer drugs using data from discovery mass spectrometry (MS) experiments. Published in *Nature Communications*, the new method reconstructs pathways robustly, allowing researchers to ask more precise questions about how drugs affect proteins and pathways.
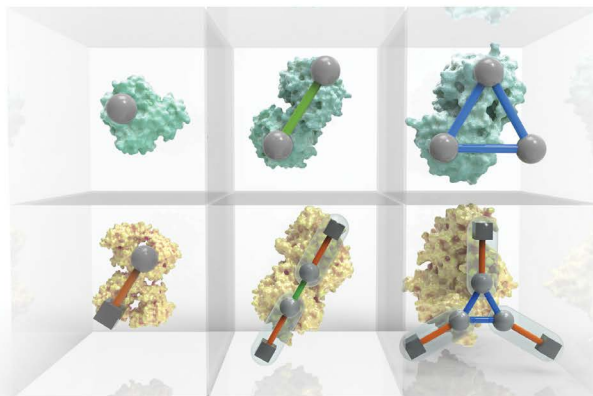
Created by an interdisciplinary team led by researchers at EMBL-EBI, the Sanger Institute and the University of Cambridge, the Periodic Table of Protein Complexes (www.periodicproteincomplexes.org) provides a valuable tool for research into evolution and protein engineering.

The Bertone group, in collaboration with colleagues at the Wellcome Trust–MRC Stem Cell Institute, published a map of gene expression in mouse and marmoset embryos that defines the common origins of pluripotency in mammalian development. Published in *Developmental Cell*, their study identifies specific pathways that are critical for early lineage segregation in the primate, and paves the way for optimising methods to isolate pluripotent stem cells, reprogram cells to pluripotency, or improve human embryo culture.

The Marioni and Teichmann groups, working with their collaborators in the Single Cell Sequencing Centre, revealed new genes involved in the stem-cell regulatory network in a study of gene expression in mouse embryonic stem cells. Published in *Cell Stem Cell*, their research reveals new subpopulations of cells and provides new methods to find meaning in the data.

Molecular similarities between food and environmental proteins that cause allergy (such as pollen) and multicellular parasites were identified systematically for the first time by the Thornton group, as part of an interdisciplinary collaboration with the University of Cambridge, the University of Edinburgh and the Ugandan Ministry of Health. The findings, published in *PLoS Computational Biology*, help demonstrate the evolutionary basis for allergy and support the hypothesis that allergic reactions are a flawed antibody response towards harmless environmental allergens.

The most extensive catalogue of structural variations, the 1000 Genomes Project, provides a reference for large-scale genetic differences in populations across the globe. The final outcomes of the project were published in *Nature* alongside research showing that structural variations are often likely to have functional consequences. Led by the Korbel group in Heidelberg with contributions from the Stegle group, this research clarifies what researchers should be looking for when trying to understand the genetic causes of a certain condition.

The MinION™, a handheld DNA-sequencing device developed by Oxford Nanopore, was tested and evaluated by an international consortium coordinated by Ewan Birney. By the time the analysis of data generated in five different laboratories was published, development of the sequencer had already moved on considerably. In this rapid innovation environment, the group decided to make the data freely available for re-analysis and input on a dedicated channel on *F1000Research*. The device opens up new possibilities for using sequencing technology in the field, for example in tracking disease outbreaks, testing packaged food or the trafficking of protected species.

European coordination

EMBL-EBI is a founder of ELIXIR, the research infrastructure providing access to a global portfolio of life-science data resources maintained by its members, including EMBL-EBI. ELIXIR aims to provide sustainable life-science data resources and related services, and in 2014 was invited to apply to a dedicated call within the EU's Horizon 2020 funding programme. The result of this call is the ELIXIR-EXCELERATE project, which started in September 2015.

ELIXIR EXCELERATE is dedicated to identifying and supporting the development and deployment of essential data services over the next four years, and an important part of this work is integrating services amongst the ELIXIR Nodes. EMBL-EBI teams played an important role in beginning to coordinate these endeavours.

EMBL-EBI is also co-leading the ELIXIR-EXCELERATE data platform in collaboration with the SIB Swiss Institute of Bioinformatics, and leads use cases in marine metagenomics, plant science and human data. We are also very involved in the project's compute platform, training and capacity-building areas.

One ELIXR milestone in 2015 was the automation of data transfer from the controlled-access European Genome-phenome Archive (EGA) to the Centre for Genomic Regulation (CRG) in Barcelona, which co-develops the resource. This was made possible by a new protocol for secure data transfer, optimised

EMBL-EBI hardware, network access supporting large-scale re-encryption processes, and improved monitoring tools for validating data integrity at the CRG.

We contributed to ELIXIR's human data activities in the context of the 'Beacon' collaboration with Global Alliance for Genomics and Health (GA4GH), which aims to facilitate discovery of genomic data stored in EGA and national resources.

Looking ahead

Molecular biology is highly relevant to clinical research and, increasingly, clinical practice. We are called upon to provide support for clinical researchers and healthcare professionals, and have an articulated strategy for medical data. Using this and EMBL's membership in the Global Alliance for Genomics in Health (GA4GH) as a basis, we will continue to foster relationships with the clinical research community, providing international coordination for open, human-disease data initiatives worldwide, providing reference datasets for clinical research and encouraging Europe's nations to develop biomedical informatics capacity.

Over the past few years we have seen astounding growth in all areas of data generation, perhaps most notably in metabolomics experiments. But with super-high-resolution electron microscopy and other imaging techniques coming online, we are starting to lay the groundwork for hosting reference image data alongside our public molecular data offerings. This undertaking is not trivial, and to be sustainable will require considerable commitment by public funding bodies over the very long term.
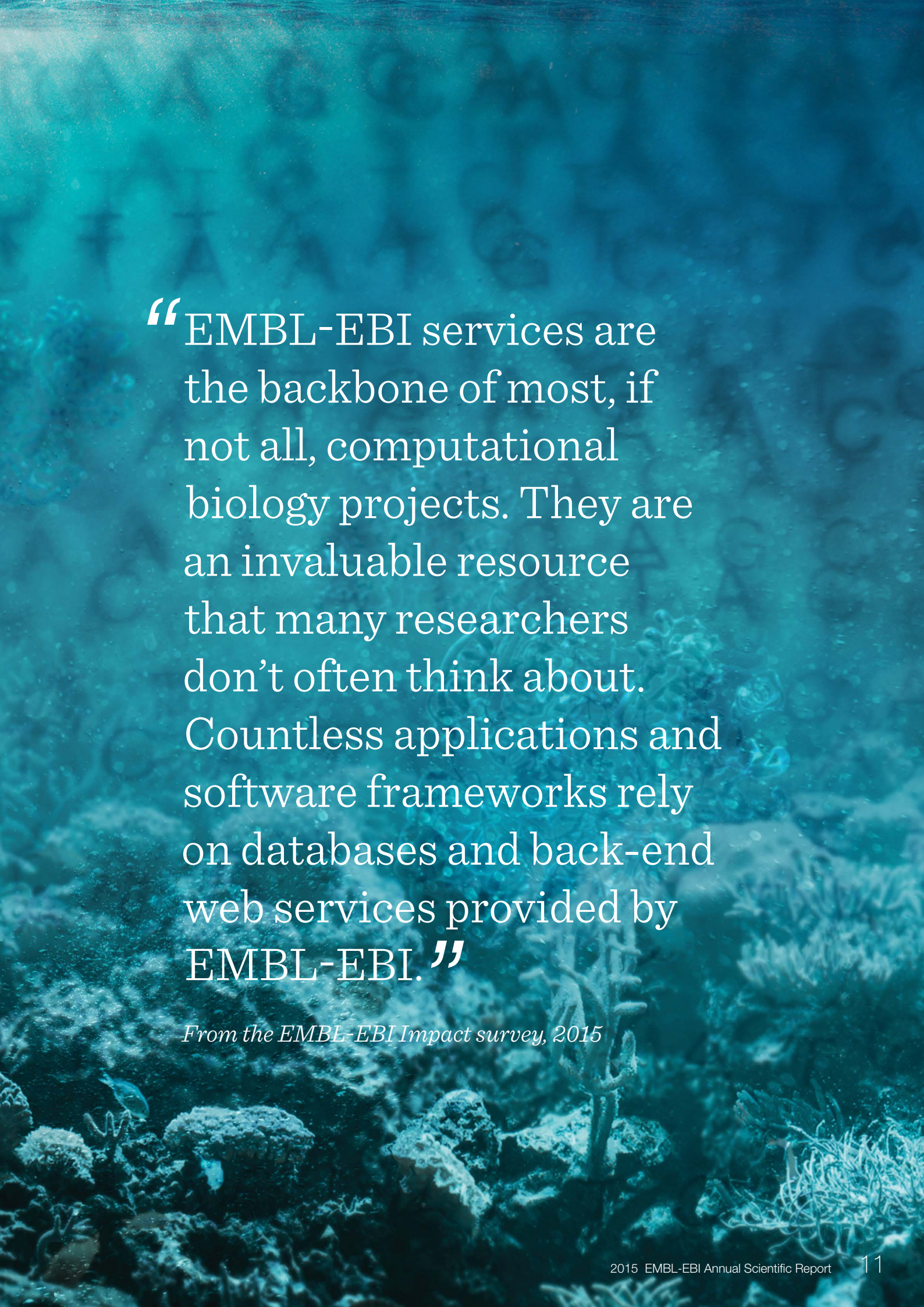
Expanded infrastructure for pairing imaging and molecular data could have a profound impact on biological and biomedical research, enabling a whole new way of exploring connections between genotype and phenotype. It would also be an incredibly positive development for the imaging community, which at present lacks a unified, standardised, central, public repository to serve as an archive of record and to be re-examined in new ways well into the future.

The performance of the MiniION miniature sequencing device developed by Oxford Nanopore was tested by a global consortium led by EMBL-EBI.
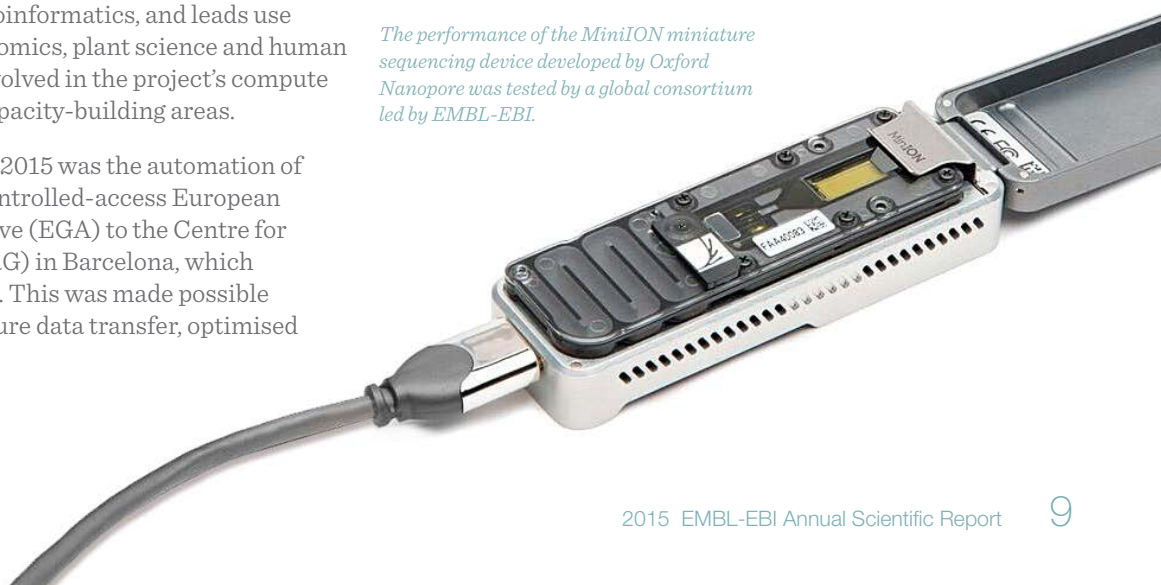
Services

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

From the EMBL-EBI Impact survey, 2015

Genes, Genomes and Variation

Genes, genomes and variation data resources represent the largest service cluster at EMBL-EBI, comprising the European Nucleotide Archive (ENA), Ensembl, Ensembl Genomes, the GWAS Catalog, EBI Metagenomics, RNAcentral, Rfam, the European Variation Archive (EVA) and the European Genome-phenome Archive (EGA), among others. Teams working in this area play a vital role in data coordination for large-scale projects such as BLUEPRINT, HipSci and the 1000 Genomes Project, to name a few.

In 2015 we were pleased to welcome new team leaders Bronwen Akwen, Andy Yates and Daniel Zerbino, who will steer different areas of work in vertebrate genomics. We were grateful to Helen Parkinson for taking up stewardship of our variation resources following the departure of Team Leader Justin Paschall to the University of California Berkeley in the US.

The ENA, a foundational resource for molecular biology, facilitated the rapid, early sharing of pathogen surveillance data and related information for the EU-funded COMPARE project in 2015. New 'Data Hubs' and cloud-compute environment will make it easier for researchers to track and rapidly share information about viruses such as Ebola and Zika.

The ENA also strengthened the marine research community, helping establish the 'M2M3' standard for marine data reporting and service provision.

EBI Metagenomics handled staggering growth during the year, notably with the addition of datasets from the EMBL-led Tara Oceans expedition.

In 2015 the Vertebrate Genomics teams contributed to the launch of the Functional Annotation of Animal Genomes (FAANG) project, leading on data and metadata standards. These efforts are essential to progress in genomic and metagenomic studies.

RNAcentral, an integrating resource, welcomed 12 new expert databases, offering a broader palette of data and deeper integration with other information sources.

As the 1000 Genomes Project wrapped up in 2015, the datasets were updated to reflect the latest human reference assembly, published and made available through Ensembl and the EVA. EMBL-EBI teams were deeply involved in data coordination for the project, and contributed to research findings. Ensembl released a fleet of new features that improve speed, integration, usability, data discoverability and accessibility, all amidst rapidly growing usage: the Ensembl REST API served over 70 million requests during the year.

Plant

Estuary

Sediment

Animal

Ocean
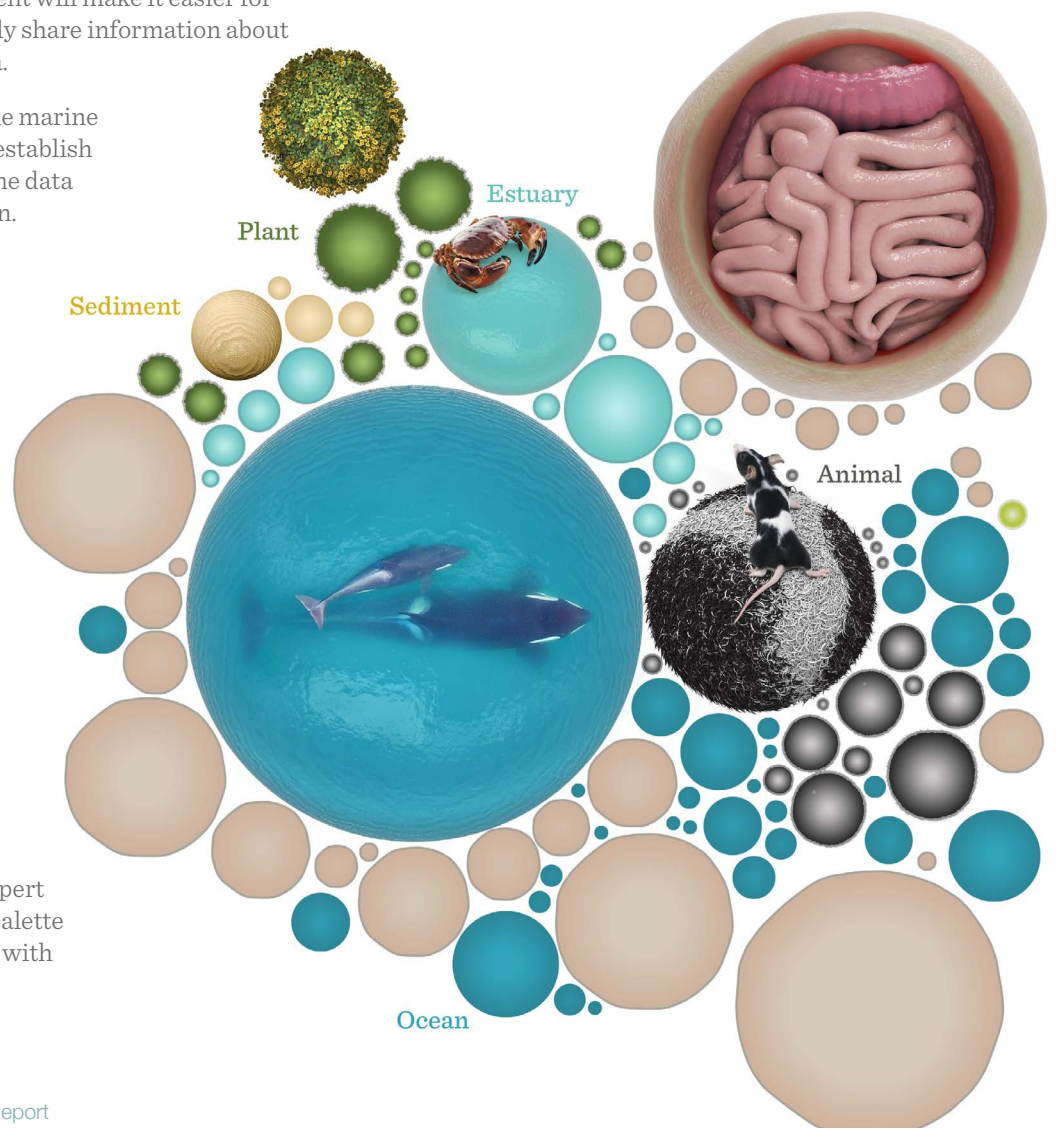
European Nucleotide Archive

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's palette of services is provided over the web and through a powerful programmatic interface. ENA data and services form a core foundation upon which scientific understanding of biological systems has been assembled. With ongoing focus on data presentation, integration within ENA, integration with resources external to ENA, tools provision and services development, our commitment is to the utility of ENA content and achieving the broadest reach of sequencing applications.

www.ebi.ac.uk/ena

Ensembl

The Ensembl project, founded in 1999 to support the results of the Human Genome Project, supports over 80 vertebrate species and provides resources such as reference gene sets, whole genome alignments, gene homology annotation, gene sequence alignments, variant annotation and regulatory regions. Many of these datasets have been adopted as authoritative references within the scientific community.

www.ensembl.org

EMBL-EBI's Metagenomics data service offers tens of thousands of analysed genomes and tools for studying data from environmental samples. This illustration by Spencer Phillips is based on a proportional representation of datasets in the Metagenomics portal.

Engineered

Insect

Ensembl Genomes

Ensembl Genomes is an integrating portal providing access to genome-scale data from across the taxonomic space. Using the infrastructure developed in the context of the Ensembl project, it offers consistent interactive and programmatic user interfaces to data from important invertebrate metazoan, plant, fungal, protist and bacterial species. Ensembl Genomes supports key data types including genome sequence (structural and functional), annotation of genes, regulatory elements and polymorphisms. It also supports comparative and evolutionary analyses.

www.ensemblgenomes.org

GWAS Catalog

The NHGRI-EBI GWAS Catalog is a quality-controlled, manually curated, literature-derived collection of all published genome-wide association studies. Co-developed with the NHGRI, it provides a karyotype visualisation of GWAS Catalog data. The GWAS Catalog is integrated with Ensembl and Europe PubMedCentral.

www.ebi.ac.uk/gwas/

European Genome-phenome Archive

The EGA, co-developed with the Centre for Genomic Regulation (CRG) in Spain, contains human data collected from research participants whose consent agreements authorise data release only to bona fide researchers and possibly for specific uses. Strict protocols govern how information is managed, stored and distributed by the EGA project. The EGA help desk provides service for both data submitters and those seeking access to the available datasets.

www.ebi.ac.uk/ega

European Variation Archive

The European Variation Archive is an open-access database of all types of genetic variation data, from all species. The EVA provides access to highly detailed, granular, raw variant data from many species, including human, sheep and tomato. All users can download data from any study, or submit their own data to the archive. You can also query all variants in the EVA by study, gene, chromosomal location or dbSNP identifier using our VCF Browser.

www.ebi.ac.uk/eva

EBI Metagenomics

Our Metagenomics service is a free-to-use, large-scale platform for analysis and archiving of metagenomic, metatranscriptome and amplicon data. It provides a standardised analysis workflow, capable of producing rich taxonomic diversity and functional annotations, allowing analysis results to be compared both within and across projects at a broad level, and across different data types (e.g., metagenomic, metatranscriptomic).

www.ebi.ac.uk/metagenomics

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

Rfam

Rfam is a curated database of non-coding RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models. Our families may be divided into non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Rfam families are created from ENA sequence data and experimental evidence in the literature. Rfam provides ontology terms and external references, as well as search tools to enable users to query their sequence against Rfam data. Rfam is used for automatic annotation of genome sequences as well as a test dataset for many RNA bioinformatics methods.

http://rfam.xfam.org

Data Coordination

Resequencing Informatics provides data coordination activities for multiple projects and consortia, including EBiSC, BLUEPRINT and HipSci. This part of the Vertebrate Genomics team also leads the International Genome Sample Resource, which is a continuation of the 1000 Genomes Project.

www.blueprint-epigenome.eu
www.hipsci.org
www.internationalgenome.org/

European Nucleotide Archive

Guy Cochrane

- *Established 'Data Hubs' for the COMPARE project to facilitate the rapid, early sharing of structured pathogen surveillance data and related information;*

- *Established a cloud-compute environment for the development and operation of systematic pathogen surveillance-related computational workflows;*

- *Supported a heavy load of data submissions (1 every 6 minutes) and data consumers throughout the world;*

- *Accommodated extensive growth and increasing complexity for a range of data types;*

- *Enhanced data submission and presentation services;*

- *Provided data coordination support for global programmes, including Tara Oceans and the early-access MinION Analysis and Reference Consortium initiative;*

- *Supported EBI Metagenomics, EVA, EGA, Ensembl Genomes, ArrayExpress, UniProt, RNAcentral in their data-submission services and data management;*

- *Worked extensively with communities to establish and implement data standards, e.g. 'M2M3' for marine data;*

- *Improved and applied curation processes using standards-driven checklists;*

- *Released CRAM v. 3, with greater performance and flexibility, and configured systematic indexing ('CRAI') for all raw reads submitted in CRAM format.*

Sequence Families

Rob Finn

- *Managed the five-fold growth of EBI Metagenomics, with over 100 billion reads now analysed, including the marine project Ocean Sampling Day and Tara Oceans;*

- *Developed and implemented a mechanism for updating the EBI Metagenomics pipeline (version 2.1), which resulted in an increase of throughput by over 300%;*

- *Introduced a new layout for results that highlights source biome annotations, improving data discoverability;*

- *Issued three releases of RNAcentral, which included 12 new expert databases, 1.2 million new sequences and 9.3 million new database cross-references;*

- *Developed species-specific identifiers in RNAcentral to enhance curation;*

- *Expanded functionality of the RNAcentral website, particularly with data export and sequence-similarity searches;*

- *Consolidated Rfam and RNAcentral into a single 'RNA resources' project, with production and development activities structured to ensure coordinated data releases.*

Vertebrate Genomics

Paul Flicek
Bronwen Aken
Andrew Yates
Daniel Zerbino

- *Welcomed three new Team Leaders in September: Bronwen Aken, Andrew Yates and Daniel Zerbino;*

- *Updated the data from several major projects (e.g. 1000 Genomes Project, BLUEPRINT) to reflect the new GRCh38 human reference assembly;*

- *Issued five major releases of Ensembl, and provided updates to other highly used resources, e.g. human (now assembly v. GRCh37.p8) and mouse (now GRC38m.p4) genomes;*

- *Published the Ensembl regulatory build and the genome of the vervet monkey;*

- *Released important annotation updates to the rat (Rnor_6.0) and zebrafish (GRCz10) genome assemblies, and introduced a dynamic gene gain/loss view of these datasets;*

- *Released BLUEPRINT data via GenomeStats, a web-based tool for carrying out analyses of epigenomic data;*

- *Released the beta version of our TrackHub registry;*

- *Developed new views and tools, enhanced performance and usability of existing views, extended support for track hubs, and improved our mirror sites;*

- *Developed a new BioMart system to provide fast access to all regulatory data from the new Ensembl Regulatory Build and added new bindings in Bioconductor;*

- *Helped launch the Functional Annotation of Animal Genomes (FAANG) project, in which we lead efforts to define data and metadata standards;*

- *Upgraded the HipSci project website to improve the discoverability of individual cell lines and related data;*

- *Improved display of variation data tables and introduced Manhattan plots for linkage disequilibrium data;*

- *Managed the growth in usage of the Ensembl REST API, which had over 70 million requests in 2015;*

- *Introduced a new visualisation tool for long-range connections between genomic regions;*

- *Promoted Ensembl resources through social media, conferences, webinars and 97 workshops;*

- *Helped complete the relocation of the GWAS Catalog software infrastructure from the NHGRI in the US to EMBL-EBI;*

- *Improved the GWAS Catalog website by updating the search interface with SOLR technology and supporting ontology expansion queries.*

Non-vertebrate Genomics

Paul Kersey

- *Issued six public releases of Ensembl Genomes;*

- *Contributed to the regular data releases of Vector Base, Wormbase and PomBase;*

- *Increased the number of bacterial genomes available through the Ensembl public interface to nearly 30 000;*

- *Increased the number of fungal genomes 10-fold and protist genomes 5-fold;*

- *Made major contributions to the paper describing the genome of* Anopheles stephensi, *the primary mosquito vector of malaria in urban India;*

- *Extended community curation to plant pathogens, and released new data-mining tools for PhytoPath and WormBase ParaSite;*

- *Issued a new, more contiguous and complete assembly of the bread wheat genome to the research community.*

Variation

Justin Paschall and Helen Parkinson

- *Handled a 50% increase in the volume of data archived in the European Genome-phenome Archive (EGA) and a 65% increase in the number of files submitted;*

- *Deployed a new EGA downloader service, which distributed over 1.7 Petabytes of data;*

- *Implemented a Global Alliance for Genomics and Health 'Beacon' for the EGA, enabling users to access a limited collection of variation data through a single, three-tiered entry point;*

- *Re-built the EGA pipeline, reducing the quarterly average processing time from three weeks to one and a half days;*

- *In collaboration with colleagues at CRG Barcelona, increased the EGA's capacity to distribute data via FTP, Aspera and a customised downloader;*

- *Managed the growth of the European Variation Archive to 22 datasets on various organisms, including crop species and domesticated animals;*

- *Made available datasets from Phase 3 of the 1000 Genomes Project and from the Exome Aggregation Consortium (ExAC);*

- *Improved the EVA browser by integrating variant annotations generated by the Ensembl Variant Effect Predictor tool and applying advanced search filters;*

- *Improved the representation of clinical information with a new display for data from ClinVar;*

- *Helped standardise ClinVar data in the context of Open Targets (formerly CTTV) and developed global standards for variation data as part of the GA4GH.*

Expression

Our RNA and protein expression service teams are working to create a comprehensive, integrated, scalable atlas of expression. Our goal is to make it easier for researchers to achieve a systems-based understanding of the human body and the many species with which we cohabitate and interact.

RNA-sequencing-based datasets increased substantially in 2015, and the Baseline Expression Atlas released its first large-scale proteomics dataset for protein expression in human tissues. Our Expression Atlases now offer data from 2620 studies and close to 100 000 assays, and contribute transcriptomic data and visualisations to many public resources, including the Target Validation platform, Ensembl, Reactome, Plant Reactome and the International Mouse Phenotyping Consortium.

A new pipeline launched in 2015 makes it easier to analyse public RNA-seq data for major species in the ENA, and resulted in the inclusion of a large amount of data on 85 species being fed into both the Expression Atlas and Ensembl.

ArrayExpress developments in 2015 include a redesigned experiment-management tool, automated management of sequencing data and, thanks to collaboration with the ENA, new methods for storing raw sequence data and information. We also updated the Annotare submission tool to respond to the needs of data curators. The result is a more robust, user-friendly, quality-controlled data resource.

In the context of the BioMedBridges project and in collaboration with the Samples, Phenotypes and Ontologies team, our teams worked to establish the basic infrastructure components of the BioSamples database. The prototype system launched in 2015 links information from biobanks including the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) Hub and the Resource Entitlement Management System (REMS).

The PRIDE database, a key ProteomeXchange partner processed over 1500 submitted datasets, archived in a relaunched website and available through a new, REST-based Web Service. ProteomeXchange central benefitted from the new Omics Discovery Index (OmicsDI), a dataset-discovery tool providing access to nine data repositories for proteomics, metabolomics and access-restricted studies in the EGA.

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) project for the reuse of patient health records in clinical research.

ArrayExpress

The ArrayExpress Archive is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. ArrayExpress is one of three international repositories recommended by many journals for holding microarray or RNAseq functional genomics data supporting publications.

www.ebi.ac.uk/arrayexpress

Expression Atlas

The Expression Atlas is an added-value database based on re-analysis of (RNA-seq and microarray) gene and protein expression data from EMBL-EBI's ArrayExpress and PRIDE archives respectively, as well as from external sources. It shows you which genes/proteins are expressed under different conditions (e.g. tissues, cell types, cell lines, strains or developmental stages), and how expression differs between conditions (e.g. comparing disease to healthy). Future development will include metabolite expression data.

www.ebi.ac.uk/gxa

PRIDE

The PRIDE PRoteomics IDEntifications 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

Functional Genomics

Alvis Brazma

- *Released the BioStudies Database for storing descriptions of biological studies and unstructured data at EMBL-EBI;*
- *Led the effort to integrate data from large genomics and proteomics studies in the Expression Atlas;*
- *In the context of the ICGC Pan-cancer analysis project, analysed data from over 1400 cancer genomes and transcriptomes.*

Proteomics Services

Henning Hermjakob and Juan Vizcaino

- *As part of the ProteomeXchange consortium, processed over 1500 submitted proteomics datasets in 2015, with downloads reaching 200 Terabytes;*
- *Redeveloped the PRIDE Archive website, developed a REST-based web service and released a new version of the popular, stand-alone PRIDE Inspector tool suite;*
- *Developed the Omics Discovery Index (OmicsDI), a dataset-discovery tool providing access to nine different data repositories for proteomics, metabolomics and access-restricted studies in the EGA.*

Functional Genomics Development

Ugis Sarkans

- *Released the BioStudies Database;*
- *In collaboration with the Literature Services team, populated the BioStudies database with supplementary materials from articles in Europe PMC;*
- *In the context of the BioMedBridges project, developed a secure information access pilot, demonstrating the linkage of biobank data across different infrastructures and levels of access control.*

Gene Expression

Robert Petryszak

- *Led the development of the new, high-quality Expression Atlas and its Baseline Atlas, which comprises large genomics and proteomics studies;*
- *Co-developed a pipeline for large-scale automatic analysis of public RNA-seq data;*
- *Led the submissions handling effort for ArrayExpress;*
- *Brokered sequencing data for the ENA.*

Protein Sequence Resources

EMBL-EBI provides foundational resources for researchers who work with protein sequences and protein families, including the UniProt, InterPro and Pfam data services, and the HMMER homology search tool, among others.

We develop and curate the UniProt, the universal protein resource, in collaboration with the SIB Swiss Institute of Bioinformatics and Protein Information Resource (PIR). Building on the UniProt Knowledgebase, we provide further resources for exploring and comparing protein families, domains and motifs.

HMMER, a fast, sensitive search tool, helps biologists find sequence relationships deep in evolutionary time. In 2015 we made HMMER algorithms available through a dedicated, open-source website at EMBL-EBI, providing an advanced tool to help researchers infer the function of a protein and its evolutionary history.

In 2015 we re-launched the Enzyme portal, which integrates all information about enzymes from EMBL-EBI resources. Built following user-centred design methodology, the service makes it easier to navigate comprehensive summaries, enzyme comparison, sequence search and search entry points to enzymes by disease, pathway, taxonomy and EC.

Reducing redundancy is important to ensure efficiency and quality are maximised, but is a major challenge in data management. In 2015 we implemented a new method for identifying highly redundant proteome datasets, and removed them from UniProtKB. The result is a more streamlined, efficient resource.

We also began distributing new data types: variants with consequences at the protein level. We incorporated variants in the protein context from the Exome Aggregation Consortium (ExAC, hosted by the Broad Institute) and the Exome Sequencing Project (ESP, hosted by the University of Washington). Working with the PeptideAtlas at the Institute for Systems Biology in the US and MaxQB in Gemany, we released peptide data from MS experiments, mapped to UniProt proteins.

Visualisation was a focus area in 2015. Users can now map and visualise UniProtKB sequence feature annotations including domains, sites and post-translational modifications. This feature viewer was released in beta in 2015, and made public in early 2016. We also implemented a PSIQUIC server for visualising protein–protein interaction annotations using the open-source Cytoscape software.

In 2015 we helped establish minimum standards for genome annotation, which will make it easier for diverse communities to work with public genome and proteome datasets.

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. Our Protein Function teams contribute to several resources, each optimised for a different purpose:

UniProt Knowledgebase (UniProtKB): the central database of protein sequences, providing accurate, consistent, rich annotations about sequence and function;

UniProt Archive (UniParc): a stable, comprehensive, non-redundant collection representing the complete body of publicly available protein sequence data;

UniProt Reference Clusters (UniRef): non-redundant data collections that draw on UniProtKB and UniParc to provide complete coverage of the 'sequence space' at multiple resolutions.

www.uniprot.org

InterPro

InterPro is used to classify proteins into families and predict the presence of domains and functionally important sites. The project integrates signatures from 11 major protein signature databases. InterPro rationalises instances where more than one protein signature describes the same protein family or domain, uniting these into a single InterPro entry and noting relationships between them. It adds biological annotation and links to external databases such as GO, PDB, SCOP and CATH. InterPro pre-computes all matches of its signatures to UniParc proteins using the InterProScan software, making the data available in a variety of machine-readable formats and via web-based interfaces. This data is updated and incorporated into each UniProtKB release. InterPro applications include the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects, and large-scale mapping of proteins to GO terms for Ensembl and the GOA project. It forms a core component of the EBI Metagenomics analysis pipeline.

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. As homologous proteins are more likely to share structural and functional features, Pfam families can aid in the annotation of uncharacterised sequences and guide experimental work.

http://pfam.xfam.org

HMMER

HMMER is a sequence-analysis package that can be used with both protein and nucleotide sequences. At the core of the software is an algorithm that enables the searching of one or more probabilistic models (profile hidden Markov models, HMMs) against either a single sequence or a database of sequences. The HMMER website has implemented this software as a set of fast web services, with both a programmatic interface and graphical user interfaces. Profile HMMs are incredibly powerful, allowing users to detect distant evolutionary relationships.

www.ebi.ac.uk/Tools/Hmmer

Protein Function Development
Maria Martin

- *Re-launched the Enzyme portal and developed new interfaces and tools for UniProt and QuickGO, with a focus on optimising user interaction with these websites;*

- *Implemented a method for identification of highly redundant proteomes and removal from UniProtKB;*

- *Extended the provision of variants with consequences at the protein level, incorporated variation data from ExAC and the Exome Sequencing Project (ESP);*

- *Released experimental peptides mapped to UniProt proteins from mass-spectrometry studies in collaboration with PeptideAtlas and MaxQB;*

- *Extended the scope of the annotation tool Protein2GO and the GO browser QuickGO, and implemented a PSIQUIC server for protein-protein interaction annotations visualisation in Cytoscape;*

- *Implemented the automatic annotation of domains, signal peptides, transmembrane and coil-coil regions for millions of protein sequences in UniProtKB/TrEMBL.*

Protein Function Content
Claire O'Donovan

- *In the context of the Consensus Coding Sequence (CCDS) project, ensured the curated, complete synchronisation with the HGNC, which has assigned unique gene symbols and names to 39 000 human loci (19 001 of which are listed as coding for proteins);*

- *Helped establish minimum standards for genome annotation to enable scientists to exploit complete genome and proteome datasets to their full potential;*

- *Improved UniProt Automatic Annotation by significantly increasing the number of UniRules, with an emphasis on enzymes across the taxonomic space;*

- *Secured funding to continue our contribution to the validation of the computational approaches submitted to the Critical Assessment of Function Annotation experiment.*

Sequence Families
Rob Finn

- *Refactored Pfam to utilise UniProt reference proteomes as the underlying sequence database, streamlining curation and production processes while minimising impact on sensitivity;*

- *Optimised Pfam quality control to allow minor overlaps between Pfam entries to allow better modeling of protein families;*

- *Streamlined production and delivered monthly updates of InterPro data to UniProt for their automatic annotation procedures;*

- *Integrated a net gain of over 2000 new member database signatures within InterPro, resulting in over 1800 new entries;*

- *Provided GO terms to UniProt, with the latest release assigning ~110 million terms to approximately 35 million proteins in UniProt release 2016_01;*

- *Migrated the HMMER web services from Janelia Research Campus;*

- *Expanded HMMER services to include PIRSF HMM searches and support for UniProt reference proteomes, now the default sequence database;*

- *Issued two releases of Pfam and six releases of InterPro.*

Molecular and Cellular Structure

Understanding the structure of a molecule is key to understanding how it may function. PDBe, the Protein Data Bank in Europe, aims to 'bring structure to biology' by making this complex field more accessible to non-specialists. PDBe is involved in managing three of the major archives in structural biology: the Protein Data Bank (PDB), the Electron Microscopy Data Bank (EMDB) and the Electron Microscopy Pilot Image Archive (EMPIAR).

In 2015 we launched a completely redesigned, responsive PDBe website that raises the bar for the delivery of 3D structural information. Based on years of usability research and an iterative design and testing process, the mobile-friendly interface features intuitive layout and organisation, new categories of value-added information and a set of images providing rich insights into quaternary structure, ligand binding, and sequence and structure domain annotations. A powerful, BioSolr-based search system, developed in collaboration with the Samples, Phenotypes and Ontologies team and Flax, a Cambridge-based search technology company, offers extended functionality. In addition, a new RESTful API provides easy programmatic access to PDB and EMDB data, supports our new entry pages and can be used freely by external software developers.

As part of our on-going work to improve and enrich PDBe content, we established new quality-control measures focused on accessibility and discoverability, and integrated SIFTS annotations and the PISA data analysis tool. The wwPDB validation pipeline now provides quality information, so users can identify "best quality" structures for given macromolecules. An integrated wwPDB software system for deposition and annotation of structural data, launched in January 2016, features community-driven validation reports.

3D EM is more accessible than ever, generating large datasets that require processing to optimise their utility. High-resolution structures (e.g. 2.2Å) started to enter PDBe, which is well positioned to support the 3D EM user communities during this phase of very rapid development.

In early 2015, our annotation experts began to handle all depositions to the PDB from European and African laboratories and companies, as well as all depositions that come through previous deposition systems (4007 of 10 886 total worldwide depositions to PDB and 452 entries in EMDB in 2015). These team members also reached out to users in promotional events throughout Europe, presentations at international conferences, sustained social media engagement, newsletters, courses, invited lectures and new e-learning materials in Train Online.

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 repositories of molecular and cellular structure data, the Protein Data Bank (PDB) and the Electron Microscopy Data Bank. PDBe also develops advanced services that provide access to the structural data in a variety of ways. PDBe's goal is to 'bring structure to biology'.

www.ebi.ac.uk/pdbe

Electron Microscopy Data Bank

The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography, and electron (2D) crystallography.

www.ebi.ac.uk/pdbe/emdb

EMPIAR

Electron Microscopy Pilot Image Archive is a public resource for raw, 2D electron microscopy images. The purpose of EMPIAR is to facilitate methods development and validation, which will lead to better 3D structures. It is a PDBe project based on input from the EM community

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

Protein Data Bank in Europe

Gerard Kleywegt and Sameer Velankar

- *Curated a record 4007 PDB entries;*

- *Curated 452 EMDB entries;*

- *Released 22 new EMPIAR entries;*

- *In collaboration with partners in the US and Japan, extended the wwPDB deposition and annotation software, as well as the wwPDB structure-validation pipelines, to fully support 3DEM and NMR data;*

- *Further developed EMPIAR, a rapidly growing public archive for raw 3DEM and tomography image data;*

- *Released the fully redesigned PDBe website (including the PDB and EMDB entry pages), a powerful new search system and an API;*

- *Improved and extended the internal production process and data and database infrastructure to support the new website, search system and API;*

- *Intensified outreach and training efforts, ranging from social media to expert workshops.*

PDB entry 1TN2: A transfer RNA (tRNA) molecule structure determined by Aaron Klug. Aaron Klug developed crystallographic electron microscopy methods and elucidated the structure of tRNA, the nucleosome and the zinc-finger DNA-binding domain. For his work on these biologically important nucleic acid-protein complexes he received the 1982 Nobel Prize in Chemistry.

Chemical Biology

EMBL-EBI's chemical biology resources help researchers design and study small molecules and their effects on biological systems. These resources are well integrated with many of our core molecular resources, enabling scientists in industry and academia to explore life-science data in new ways.

In 2015 John Overington left EMBL-EBI to join biotech company Stratified Medical, and Anne Hersey stepped in as Acting Team Leader.

ChEMBL grew to 1.7 million compounds and nearly 14 million bioactivities, serving approximately 15 000 unique visitors per month through its web interface and many more through downloads, Web Services and the RDF platform. New RDkit Web Services allow users to perform more complex queries and combine data and chemistry-aware queries.

At the end of 2015 the number of novel chemical entities annotated in SureChEMBL stood at approximately 17 million. A new data client feed enables users to maintain a regular stream of the patent data behind a firewall and integrate the data with their in-house data.

The fully open-source ChEBI database grew to over 47 500 fully curated chemical entity entries, about a third of which are from direct data submissions, and includes citation information for 11 000 natural product entries. ChEBI is well integrated with MetaboLights, Reactome and the Rhea enzyme database.

MetaboLights was redesigned with authenticated Web Services, improved search, more flexibility for data submissions and a browsable 'tree of life' featuring automatic classification compatible with identifiers from 89 different taxonomy sources. Over 18 000 compounds are now linked with ChEBI, and over 150 datasets were exported to MetabolomeXchange, the EBI Search and the Omics Discovery Index. MetaboLights is now the recommended metabolomics repository for Nature Scientific Data, Metabolomics, PLOS and EMBO journals.

The COSMOS project delivered the NMR mark-up language open standard for Nuclear Magnetic Resonance data, agreed procedures for the management and dissemination of data in metabolomics and launched the MetabolomeXchange platform. This paved the way for the 2015 launch of PhenoMeNal, the incipient e-infrastructure for the processing, analysis, and information-mining of medical molecular phenotyping and genotyping data for the European biomedical community.

ChEMBL

ChEMBL, a database of bioactive compounds, provides curated bioactivity data that quantitatively links compounds to molecular targets, phenotypic effects, exposure and toxicity end-points. ChEMBL focuses on interactions relevant to pharmaceutical and agro-chemical development. Data is organised and can be viewed by pharmaceutically important gene families in ChEMBL.

www.ebi.ac.uk/chembl

SureChEMBL

SureChEMBL extracts chemical structure data from the full text and images of patents, making it easier to check whether a newly developed drug or other product is actually novel.

www.surechembl.org

UniChem: Chemical structure cross-referencing

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 dictionary of molecular entities focused on small chemical compounds. It is a manually annotated database that provides a wide range of related chemical information such as formulae, 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, as well as experimental data from metabolomics experiments. You can submit your own data, and future developments include search services around spectral similarities and chemical structures.

*www.ebi.ac.uk/metabolights*

Chemogenomics

Anne Hersey and John Overington

- *Added biological annotation to the SureChEMBL pipeline, such that disease and protein targets are referenced in the patent database;*

- *Grew the ChEMBL database by adding ~400,000 bioactivity values on ~140,000 compounds;*

- *Released enhanced ChEMBL Web Services with increased functionality;*

- *Expanded the content of the UniChem structure cross-referencing resource to provide database cross references to more than 100 million unique chemical structures from 27 source databases;*

- *Built a pipeline that automates the update of the UniChem database, such that it is updated weekly;*

- *Contributed bioactivity and patent data to the RDF-based OpenPHACTS, an Innovative Medicines Initiative project;*

- *Contributed to Open Targets (formerly CTTV) by providing annotations of target and disease information on FDA-approved drugs;*

- *Contributed to the IDG project by annotating clinical candidates for the four main target classes with their therapeutic targets.*

Cheminformatics and Metabolism

Christoph Steinbeck

- *Handled the addition of over 5400 entries to ChEBI;*

- *Automated ChEBI release procedures and replaced monthly updates of the ChEBI website by a 'live' service;*

- *Introduced a 'species table' that can be curated and searched, which facilitates the handling of species information for entities in ChEBI that are natural products;*

- *Introduced an automatic classifier that immediately classifies bulk-added data within the ChEBI ontology;*

- *Maintained and developed the ChEBI application suite, including BiNChE software for ontology enrichment analysis and OntoQuery for online ontology-based logical querying;*

- *Enhanced the ChEBI web application by adding interactive statistics graphing, links to supplier information websites and a JavaScript-based chemical structure editor;*

- *Led workpackage 7 in the EU-funded METASPACE project to develop algorithms for the analysis of spatial metabolomics data, led by Theodore Alexandrov at EMBL Heidelberg;*

- *Provided biomedical use case for the EU-funded OpenMinTed project to develop a European e-infrastructure for text mining, led by Natalia Manola in Athens, Greece;*

- *Led the PhenoMeNal project, an e-infrastructure for the analysis of clinical metabolic phenotype data.*

The MetaboLights taxonomy browser
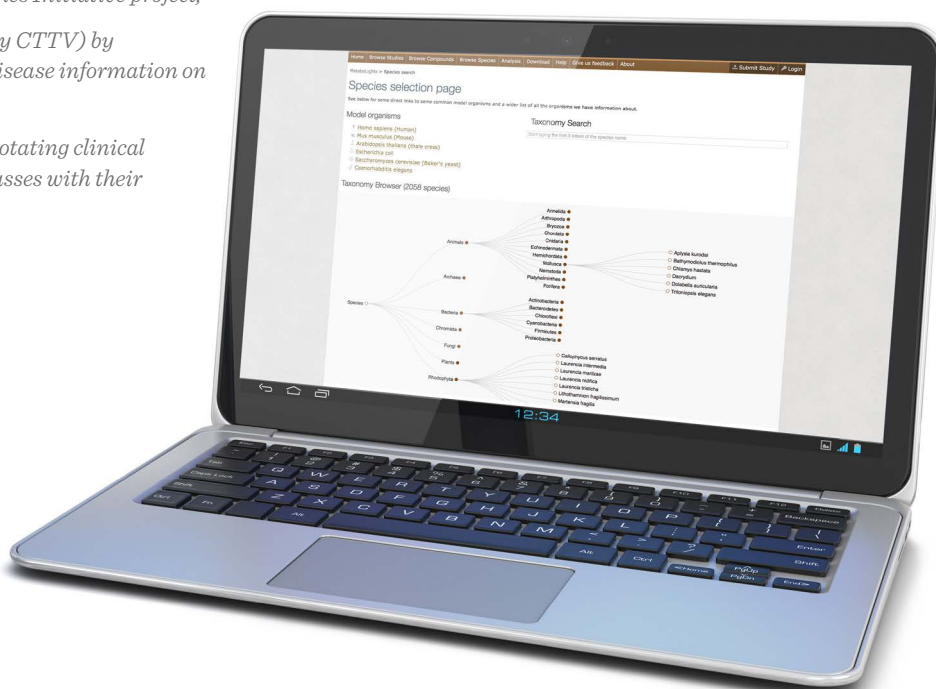
Molecular Systems

Our Molecular Systems teams provide public databases as reference implementations for community standards, for example the IntAct molecular interaction database, the Reactome pathway database, and the BioModels repository of computational models of biological systems. IntAct became an independent project under the leadership of Sandra Orchard in the new Molecular Interactions team in April 2015.

We are major contributors to the IMEx Consortium of interaction databases, distributing almost 600 000 binary interaction evidences in 2015. We published a detailed visual analysis of the interactome of LRRK2, a complex, multidomain protein strongly implied in Parkinson's disease; the dataset represents integrated data from several sources, including IMEx partners and the Reactome database.

We redeveloped the Reactome pathway diagram viewer to provide a faster, clearer interface and smooth zooming from the entire reaction network to view the details of individual reactions. All Reactome major components are now available as Web Services or JavaScript-based widgets suitable for integration into third party applications.

The Complex Portal, which contained over 1400 manually curated complexes at the end of 2015, now features an innovative graphical tool for visualising complex topology and stoichiometry. Originally developed by the Rappsilber group at the University of Edinburgh, the tool now works with a new Java library.

The BioModels database celebrated its tenth anniversary in 2015, and by the end of the year contained more than 1000 literature-based models and featured a new disease summary page. As part of DDMoRe, an IMI project well aligned with the development goals of BioModels, we provided a public platform for sharing models: the Pharmacometrics Markup Language, PharmML.

We reached hundreds of scientists in around 20 outreach and training events in 2015, including courses and workshops locally, internationally and virtually through Train online and webinars.

BioModels

BioModels Database is a repository of peer-reviewed, published, 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. In addition, models in the database can be used to generate sub-models, can be simulated online, and can be converted between different representational formats. This resource also features programmatic access via Web Services.

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

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.

Opposite: Detail from a dataset in IntAct: Phospho-tyrosine dependent protein-protein interaction network, published in 2015 by Grossmann et al., Otto-Warburg Laboratory, Max-Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Mol. Syst. Biol. 11:794.
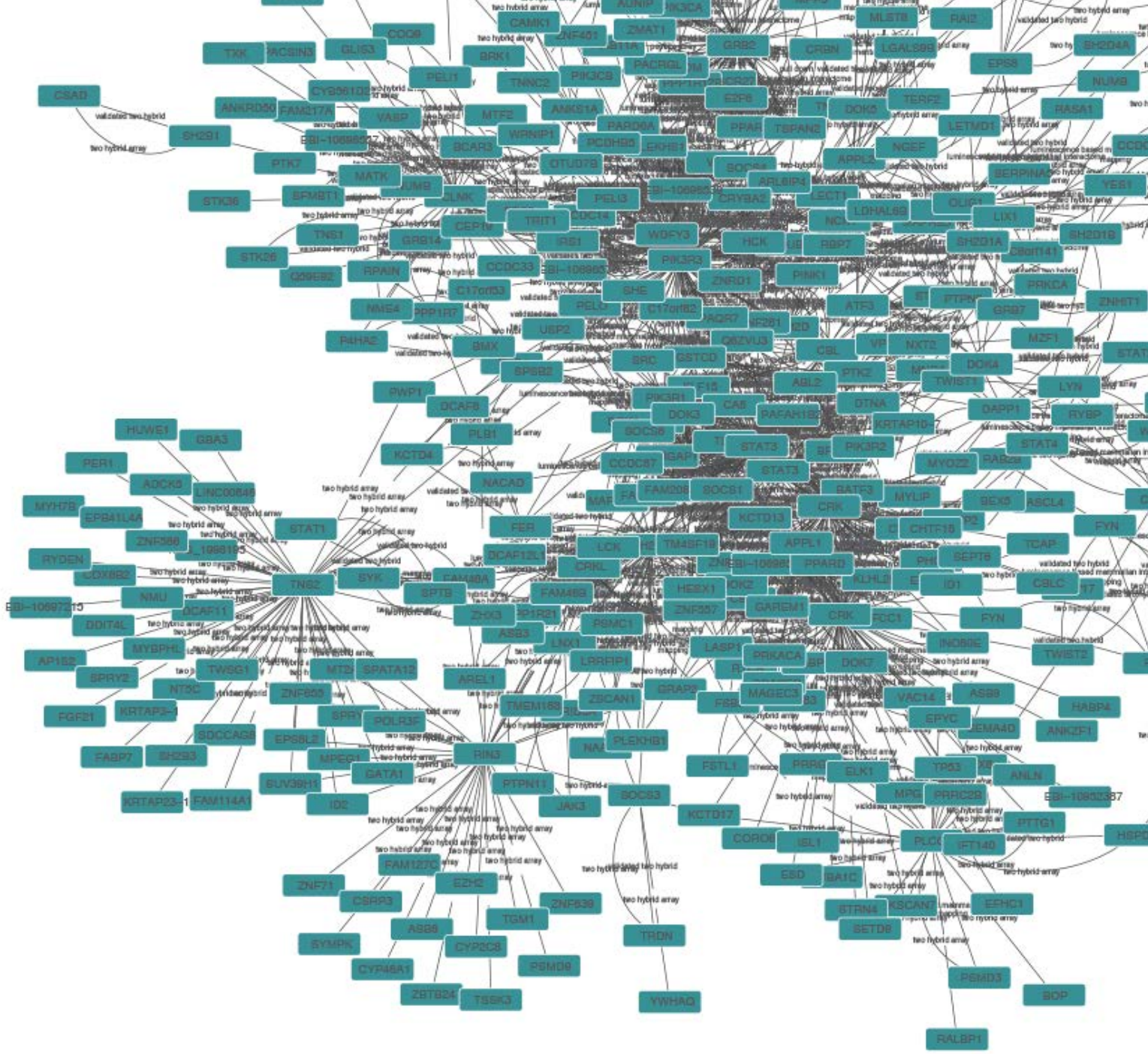
Proteomics Services and Molecular Interactions

Henning Hermjakob and Sandra Orchard

- IntAct team members and their collaborators added 100,000 binary Interaction evidences to the database;

- IntAct published a detailed visual analysis of the interactome of LRRK2, a complex, multidomain protein strongly implied in Parkinson's disease;

- Complex Portal released a novel graphical display enabling the visualisation of the topology of a protein complex;

- Reactome released a completely redeveloped pathway diagram viewer to provide a faster, clearer interface and smooth zooming from the entire reaction network to the details of individual reactions;

- BioModels team members released the Pharmacometrics Markup Language (PharmML), in the context of the Drug Disease Model Resources (DDMoRe) project.

Cross-Domain Tools and Resources

EMBL-EBI offers comprehensive, high-quality data resources covering the full spectrum of molecular biology. Our cross-domain tools and resources include literature, ontologies, samples and studies, and are pivotal in connecting diverse data types to serve researchers working in all life-science domains.

In 2015 the International Mouse Phenotyping Consortium (IMPC) informatics platform was named as one of five case studies in a G7 report on global research infrastructures, which promoted the further development of a framework for transitioning national research infrastructures to international ones. It was also cited in the literature as an example of how infrastructure can facilitate reproducibility and replicability of results by incorporating ARRIVE guidelines into data management. The project played an important role in standardising high-throughput phenotyping data, which is pivotal to bridging mouse biology and precision medicine.

The Experimental Factor Ontology played an important role in the first public release of the new Target Validation platform, with new mapping tools and content to support the platform's source disease annotations in EMBL-EBI databases. A new phenotype–disease annotation representation and gene–disease associations text mined from the literature allows this public–private partnership to integrate rare and common diseases according to shared phenotype.

As of December 2015, Europe PMC offered more than 30.5 million abstracts and around 3.5 million full-text research articles from PubMed and PubMed Central. It includes metadata for Agricola, biological patents and a new collection of books and documents, extending the scope of full-text clinical guidelines. Users can also search over 50 000 biomedical research grants that have been awarded by Europe PMC's 27 funders.

Users at over 10 million unique IP addresses visited the Europe PMC website in 2015, and the resource handled on average 24 million requests every month through its Web Services. The site was relaunched with a responsive design based on user-experience research, and features author profile and citation pages based on ORCIDs – unique identifiers for researchers. To support literature–data integration, 19 accession number types are text mined. In addition, there are now over 500 000 links to Wikipedia pages from articles in Europe PMC.

We launched the BioStudies database, which provides a new home for over half a million studies and associated files that do not quite fit with traditionally structured archives, for example toxicogenomics data from the diXa project. Authors can now link supplementary and manuscript files within the broader context of molecular data services, all through an intuitive user interface.

Solr and ElasticSearch users at EMBL-EBI, the NCBI and several other organisations came together in the BioSolr project, which brought ontology-enabled search to a wide range of life-science users, including the European Bank for induced pluripotent Stem Cells (EBiSC) project.

In 2015 EMBL-EBI took on leadership of informatics activities in two new, large-scale collaborative projects: ELIXIR Excelerate and ELIXIR CORBEL, both of which focus on interoperability and cross-domain infrastructure to support life sciences in Europe.

Europe PMC

Europe PMC contains over 30 million abstracts (including PubMed, patents, and Agricola records) and over 3.5 million full-text life-science research articles. Of these full text articles, over 1 million are available for reuse and download. As well as a sophisticated search and retrieval across all content, Europe PMC provides information on how many times the articles have been cited and by whom, links to related data resources, and text-mined terms. 'Europe PMC Labs' showcases integrated text-mining tools. PIs on grants awarded by the 27 Europe PMC Funders can use 'Europe PMC Plus' to self-deposit full-text articles and link those articles to the grant that supported the work.

https://europepmc.org

BioStudies

The BioStudies database holds descriptions of biological studies, links to data from these studies in other databases at EMBL-EBI or outside, as well as data that do not fit in the structured archives at EMBL-EBI. The database can accept a wide range of types of studies described via a simple format. It also enables manuscript authors to submit supplementary information and link to it from the publication.

www.ebi.ac.uk/biostudies

BioSamples

The BioSamples database aggregates sample information for reference samples (e.g. Coriell Cell lines) and samples for which data exist in one of EMBL-EBI's assay databases such as ArrayExpress, the European Nucleotide Archive or PRIDE, the proteomics identifications database. It provides links to assays on specific samples, and accepts direct submissions of sample information. Samples in this database can be referenced by accession numbers from data submissions to other EMBL-EBI resources.

www.ebi.ac.uk/biosamples

International Mouse Phenotyping Consortium

The International Mouse Phenotyping Consortium (IMPC) has initiated a global effort to generate and characterise a knockout mouse strain for every protein-coding gene in the mouse genome. EMBL-EBI, the Sanger Institute and MRC Harwell deliver informatics as part of the MPI2 consortium activities. The IMPC Informatics resource coordinates production of mouse strains, harmonises protocols and data export across centers, associates genes to phenotypes using an automated statistical pipeline, and performs data integration to gain new insights into human disease. Data is annotated to widely used ontologies, archived at EMBL-EBI using robust relational database structures and is freely available via an intuitive browser, an API and code downloads.

www.mousephenotype.org

Literature Services
Johanna McEntyre

- *Launched the BioStudies database, in collaboration with the Functional Genomics Development team;*

- *Was funded by the Wellcome Trust, on behalf of all Europe PMC funders, to take on the full development of the resource;*

- *Managed the growth of Europe PMC to over 3.5 million full-text articles, and created links to several outside sources including Wikipedia;*

- *Launched a re-designed Europe PMC website to make it easier to navigate and view content on mobile devices;*

- *Developed a user accounts system to enable users to save their most frequent searches;*

- *Launched ORCID-based author profiles that illustrate the impact of a person's published research over time;*

Samples, Phenotypes and Ontologies
Helen Parkinson

- *Enhanced ontology services and integrated them with industry efforts such as the Target Validation platform and Roche drug discovery;*

- *Developed mapping tools and new ontology content to support disease annotations in the European Variation Archive, UniProt and Reactome;*

- *In collaboration with colleagues in Germany and Australia, associated diseases with phenotypes based on text mining of the scientific literature;*

- *Designed a phenotype–disease annotation representation that allows integration of rare and common diseases according to shared phenotype;*

- *Participated in the BioSolr project to develop Solr and ElasticSearch expansion plugins for ontology-enabled search, for example for the EBiSC project;*

- *Updated the Gene Ontology with content for apoptosis, cilia and viruses, and improved its representation of human intestinal parasites;*

- *Collaborated with the Molecular Interactions team to standardised protein-complex annotations;*

- *Incorporated the BioSamples database into several projects that access biological sample data at the point of acquisition;*

- *In the context of the EBiSC consortium, ensured that BioSamples can model induced pluripotent stem cells accurately, and that cell lines can be registered as soon as they are generated;*

- *Launched a new BioSamples RESTful API to enable programmatic submission;*

- *Applied a versionable statistical analysis software package, Phenstat, to the IMPC infrastructure;*

- *Developed and maintained the IMPC and its informatics platform, named as one of five case studies in a G7 report on global research infrastructures, which promoted the further development of a framework for transitioning national research infrastructures to international ones;*

- *Contributed to INFRAFRONTIER, which coordinates the global distribution of mouse models produced in Europe, and PhenoImageshare, an online, cross-species, cross-repository tool enabling semantic discovery, browsing and complex annotations of phenotype images.*

Training

"EMBL-EBI is a critical mass of bioinformatics expertise and as such one of the few ways to plan for the unplannable and remain competitive. It is also a 'phone number' for any biocomputational problem at hand."

From the EMBL-EBI Impact survey, 2015

Training

Rapid changes in life-science technologies demand sustained training on all fronts, and EMBL-EBI delivers a comprehensive range of bioinformatics training to help the global research community keep pace. Our extensive Training Programme delivers multi-platform courses with contributions from experts across the institute. In addition to Ensembl's dedicated outreach operation, our service teams have representatives who work with the Training programme to deliver courses and workshops. In 2015 over 100 people at EMBL-EBI were involved in training and scientific outreach. In addition, our research programme trains the next generation of computational biologists through the EMBL International PhD Programme.

Training Programme

EMBL-EBI's diversifying user community is reflected in our Training Programme's broad range of courses, workshops and online training, which cover the full spectrum of the institute's activities. It serves the needs of researchers working in academia and companies of all sizes, and supports the Industry Programme in delivering workshops tailored to the needs of its members.

EMBL-EBI staff orchestrated and contributed to over 250 events throughout the world in 2015, including face-to-face courses at EMBL-EBI, off-site training and workshops at host institutes, demonstrations at conferences and web-based presentations and courses.

We provided a growing number of opportunities for those with limited time get up to speed with data resources. Building on the success of webinars hosted by our service teams and members of our Industry Programme, the Training Programme launched a webinar series for users of EMBL-EBI services. Train online offered 19 new or updated online courses and had over 210,000 visitors in 2015, underscoring the utility of this flexible, web-based approach.

Bioinformatics training is a highly specialised field, with few organisations throughout the world offering the concentration of expertise on offer at EMBL-EBI. Our 'train the trainer' activities supported 19 new trainers in 2015, including staff on other EMBL sites and institutes on the Wellcome Genome Campus. We continued to collaborate on trainer support with the H3Africa initiative, and with the Australian Bioinformatics Network.

Emily Perry of the Vertebrate Genomics team presenting in our custom-built IT training suite in the EMBL-EBI South building.

EMBL International PhD Programme at EMBL-EBI

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

In 2015 we benefited from the presence of 30 PhD students, welcoming five newcomers. Three students successfully defended their theses: Ewan Johnstone on cancer subtypes and glioma stem cell characterisation, Nils Koelling on quantitative genetics of gene expression, and Michael Menden on models of drug response in cancer cell lines.

EMBL International PhD Programme students at EMBL-EBI. Back row, left to right: Ananth Prakash, Emanuel Gonçalves, Damien Arnol, Valentine Svensson, Claudia Hernandez. Middle row: Anna Gawedzka, Nadezda Volkova, Lara Urban, Michael Schubert, Rachel Spicer, Sergio Miguel Santos. Front row: Greg Slodkowicz, Matthew Jeffryes, Jack Monahan, Marta Strumillo, Daniel Elias Martin Herranz, William Coleman-Smith, and David Bradley.

Healthcare professionals: a growing user community

Clinical practitioners have emerged as a growing user base, and in 2015 we participated in a number of UK National Health Service initiatives, for example informing a Master's syllabus for Genomic Medicine and co-authoring a report with Health Education England on how best to support the introduction of genomic technologies within the UK healthcare system through advanced training.

Continuing professional development

We are working to shape the continuing professional development of Europe's research community, in both bioinformatics and transferrable skills necessary for all successful research professionals. In 2015 we made substantial contributions to the IMI-funded EMTRAIN project and its comprehensive online catalogue of biomedical and life-science training opportunities throughout Europe. We also laid the groundwork for training contributions to large-scale efforts such as BioExcel for high-end computing, and the EU-funded ELIXIR Excelerate, RItrain and Corbel projects.

Industry workshops
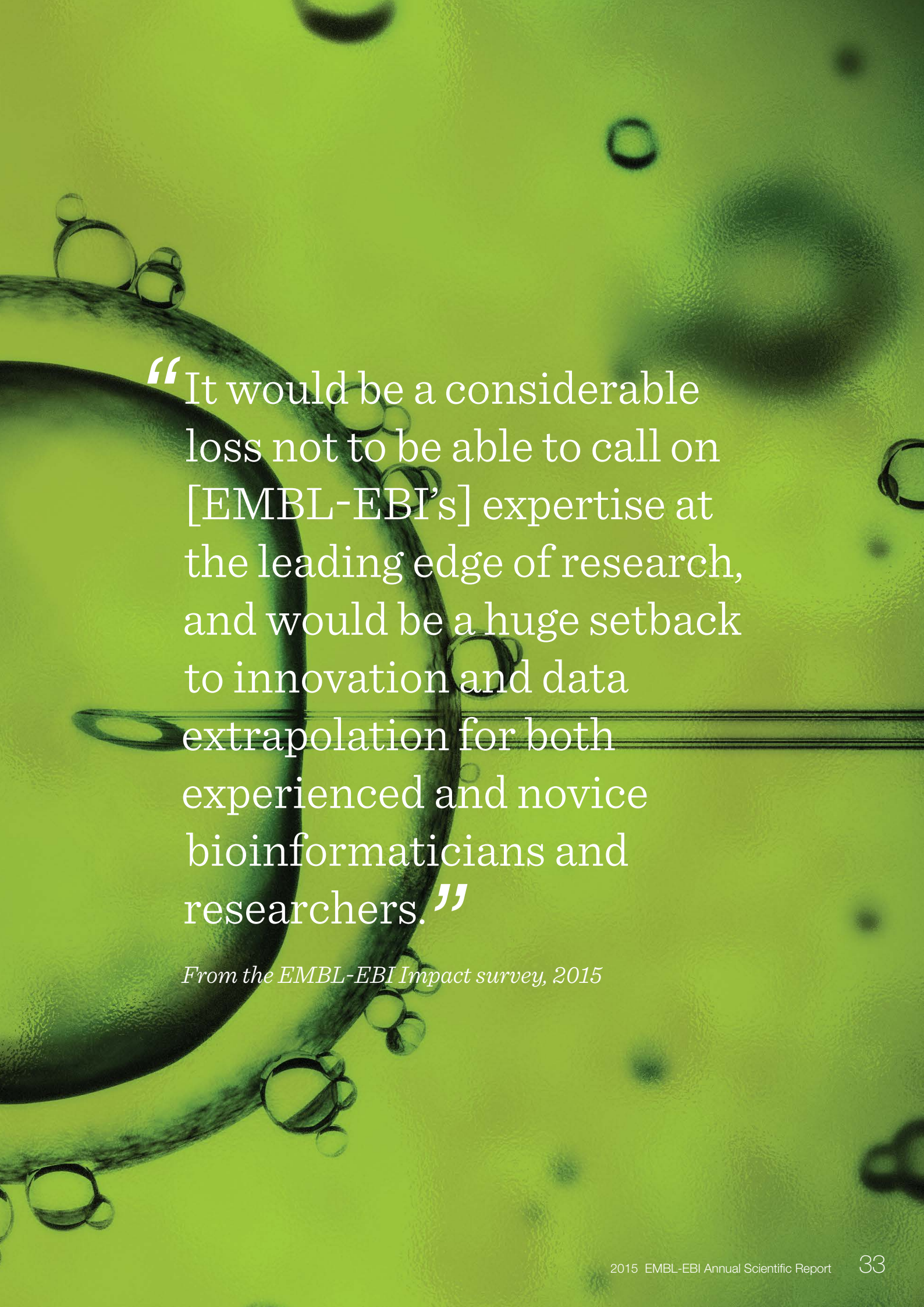
The 22 members of our Industry Programme request workshops on topics of shared relevance, and in 2015 we delivered 10 events on topics ranging from in silico ADMET prediction and immunogenomics to Semantic Web and electronic medical records for drug discovery. Two of these workshops were delivered in the US, and eight at EMBL-EBI.

EMBL-EBI Industry Programme members at a 2015 workshop.

Research

"It would be a considerable loss not to be able to call on [EMBL-EBI's] expertise at the leading edge of research, and would be a huge setback to innovation and data extrapolation for both experienced and novice bioinformaticians and researchers."

From the EMBL-EBI Impact survey, 2015

Research Achievements in 2015

In 2015 we welcomed new group leader Moritz Gerstung, whose work focuses on cancer genomics. John Marioni was appointed Senior Group Leader at the Cancer Research UK–Cambridge Institute at the University of Cambridge, a position held in tandem with his role at EMBL-EBI. Sarah Teichmann was awarded the prestigious EMBO Gold Award for her contributions to science. Sarah and Ewan Birney were both elected to the Fellowship of the Academy of Medical Sciences.

Evolution

Research from the Flicek group (Villar et al., *Cell* 2015) provided deep insights into the 'mammalian radiation', a time of rapid morphological evolution. Leveraging findings from a study comparing the genome sequences of 29 mammals, and with the help of conservation organisations such as the UK Cetacean Strandings Investigation Programme and the Copenhagen Zoo, the team studied and compared gene regulation in liver cells from 20 key species including human, Tasmanian devil, dolphin and Sei whale. Their study showed how evolution has given rise to a rich diversity of species by repurposing functional elements shared by all mammals. With their colleagues at the University of Cambridge Cancer Research UK–Cambridge Institute (CRUK CI), the group demonstrated how existing methods for understanding human biology can be used to understand a broad range of species.

The Bateman group demonstrated that protein domains, previously thought to be indivisible units that make up proteins, can exist and function despite having lost large parts of their structure (Prakash and Bateman, *Genome Biology* 2015). Their research into 'domain atrophy' offers new insights into the evolution, stability and function of protein domains.

Biotechnology

Enzymes form the basis of all forms of life, but how these catalysts have evolved their functions remains a fundamental question in biology. The Thornton group used a range of computational tools and resources to compile information about changes in enzyme function within hundreds of protein-domain superfamilies, linking changes in functions during evolution to changes in reaction chemistry (Furnham et al, *Journal of Molecular Biology* 2015). Their analyses provide insights that are useful in predicting the function of uncharacterised sequences and in the design of new synthetic enzymes.

The Teichmann group used mass spectrometry (MS) data and a large-scale analysis of structures of protein complexes to examine the fundamental steps of protein assembly. In collaboration with colleagues at the Cavendish Laboratory, they analysed tens of thousands of protein complexes and identified repeating patterns in the assembly transitions that occur (Ahnert et al., *Science*, 2015). They used these patterns to create a new 'Periodic Table' of protein complexes, which provides a predictive framework for anticipating new, unobserved topologies of protein complexes.

In 2015 the Beltrao group published a study that brings biotechnology one step closer to the regulation of protein modification by design. Their study of the structural properties of phosphosites in *Xenopus laevis* and their conservation across 13 other species (Johnson et al., *PLoS Computational Biology* 2015) showed that the degree of conservation is predictive of functionally relevant sites and interactions, and suggests that some of these sites might exert their function by controlling protein conformation.

The Thornton group tackled the tricky problem of comparing enzyme function in their study of isomerases (Martinez-Cuesta et al, *PNAS* 2016), which catalyse inter-conversion of molecules that share the same atomic composition but have different arrangements of chemical groups. Using a combination of manual and computational approaches, they catalogued known isomerisation, clustering the reactions into classes and comparing the results with the Enzyme Commission (EC) classification. Their work provides an overview of which isomerases occur in nature, how we should describe and classify them, and their diversity.

Enabling technologies

Gaining insights into disease by finding links between many different genetic traits—in hundreds of thousands of individuals at once—has been a major challenge in computational biology. In 2015 the Stegle group developed and applied methods for linking genetic variation and phenotype data (Casale et al., *Nature Methods* 2015). Their new statistical model lets researchers study associations between sets of genetic variants and multiple phenotypes, using up to 500,000 samples at once.

The Birney group used methods developed by the Stegle group to explore links between molecular events in a variety of normal and diseased human samples, and collaborated with MRI researchers and cardiologists at Imperial College, UK to detail the molecular structure and physiology of the human heart.

In 2015 the Stegle, Marioni and Teichmann groups devised new ways to tease out the heterogeneity of gene expression between single cells (Buettner et al., *Nature Biotechnology* 2015). The new protocol clarifies the true differences and similarities between cells, helping scientists better understand cancer cells, differentiation processes and the pathogenesis of diseases.

The Marioni group developed and validated a high-throughput method to identify the precise spatial origin of cells assayed using scRNA-seq (Achim, Pettit et al., *Nature Biotechnology* 2015). This approach compares complete mRNA 'fingerprints' of a cell with gene-expression profiles derived from a gene-expression atlas.

Gene expression regulation and culture conditions are both critical for maintaining the pluripotency of mouse embryonic stem cells (mESCs) in vitro. Using the single-cell RNA-seq (scRNA-seq) approach, the Marioni and Teichmann groups investigated the transcriptome profiles of mESCs in different culture conditions (Kolodziejczyk et al., *Cell Stem Cell* 2015) and showed that globally, expression in specific sets of genes varies systematically. The study brought to light new pluripotency network genes, demonstrating the value of scRNA-seq for future discovery.

A new approach to cell signalling developed by the Beltrao group (Wagih et al., *Nature Methods* 2016) combines protein phosphorylation and interaction-network data to predict sequence determinants for kinase recognition. The group used the method to predict the specificity of hundreds of human kinases, which helps clarify how protein kinases identify their target substrates.

The Enright group collaborated with the Furlong group at EMBL Heidelberg and the O'Carroll group at the University of Edinburgh to compile a catalogue of long non-coding RNAs (lncRNAs) expressed through the murine germline. They used this incredibly detailed atlas of transcription to identify a set of lncRNAs whose functions may be extremely important to the maintenance of the germline and genomic integrity.

Ewan Birney coordinated an independent, international consortium that tested and evaluated the MinION™, a handheld DNA-sequencing device developed by Oxford Nanopore. Reflecting a rapid innovation environment, the data from these evaluations was made freely available for re-analysis on a dedicated *F1000Research* channel. The device opens up new possibilities for using sequencing technology in the field, for example in tracking disease outbreaks, testing packaged food or the trafficking of protected species.

Development

The Stegle group developed new statistical methods to study the consequence of structural variants in the human genome on gene-expression levels. With the Korbel team at EMBL Heidelberg, they surveyed these effects at a genome-wide scale using the data from the final release of the most extensive catalogue of structural variations, the 1000 Genomes Project (Sudmant et al., *Science* 2015). The group found that structural variations – more than SNPs – are often likely to have functional consequences. This knowledge helps focus research into the genetic causes of a given condition.

In 2015 the Bertone group, with colleagues at the Wellcome Trust–MRC Stem Cell Institute, published a map of gene expression in mouse and primate embryos that defines the common origins of pluripotency in mammalian development. (Marmoset Genome Sequencing and Analysis Consortium, *Developmental Cell* 2015). The group analysed the complex network of gene regulation that supports pluripotency, examining how this network comes together and later collapses as cells exit the pluripotent state to become specialised cell types. The findings have implications for optimising methods to reprogram cells to pluripotency, or to improve human embryo culture.

Understanding ageing and disease

The Saez-Rodriguez group developed a novel method to build logic signalling networks from phospho-proteomic data generated in MS shotgun data and, with colleagues at Barts and the London School of Medicine and Dentistry, used it to study the effect of drugs on breast cancer cells (Terfve et al., *Nature Communications* 2015). The method reconstructs pathways robustly, allowing researchers to ask more precise questions about how drugs affect proteins and pathways.

In their work on ageing, the Thornton group, building on the results of their GWAS study for *Drosophila* lifespan (Ivanov et al., *Journals of Gerontology* 2015), collaborated with the Institute of Healthy Ageing at UCL to test the effect of manipulating 10 of the *Drosophila* genes most significantly associated with lifespan. Preliminary results suggest that a proportion of the genes significantly reduced lifespan, and over-expression of one significantly increased lifespan as compared to control.

In a series of computational studies (Tyagi et al., *Frontiers in Immunology* 2015), the Thornton group established molecular similarities between parasite proteins and allergens that affect the nature of immune response, and predicted regions of parasite proteins that potentially share similarity with the IgE-binding regions of the allergens. Their findings, part of an interdisciplinary collaboration with the University of Cambridge, the University of Edinburgh and the Ugandan Ministry of Health, have implications for the prediction of likely allergens and the design of molecules to treat allergy.

Research Summaries

Beltrao group

- *Studied the conservation and structural properties of phosphosites in proteins from the African clawed frog (X. laevis);*

- *Developed a method to predict the specificity of protein kinases using protein phosphorylation data and functional interaction information;*

- *Measured 250 000 conditional gene–gene interactions in yeast to study condition-dependent genetic interactions.*

Birney group

- *Published an overview of Oxford Nanopore Reference data as part of the MARC consortium;*

- *Published the first transcription factor QTL study, the CTCF transcription factor in human LCL lines;*

- *Published an analysis of the Kiyosu population of Medaka, and showed that this population has good properties for establishing an isogenic panel appropriate for quantitative trait mapping.*

Enright group

- *Worked on two important papers illustrating the CaptureSeq methodology for deep characterisation of transcript isoforms;*

- *Published a major new computational system for the assessment of microRNA uridylation and modifcations;*

- *Developed new computational techniques for the prediction and characterisation of long non-coding RNAs in mouse and D. melanogaster.*

Flicek group

- *In collaboration the Odom group at the University of Cambridge, mapped and analysed evolution of genomic promoter and enhancer elements across 20 mammalian species to produce the most comprehensive view of shared and lineage-specific promoter and enhancer elements in the mammalian lineage;*

- *In collaboration with the Merkenschlager group at Imperial College London, gained new insights into the role of cohesin in genome regulation;*

- *In collaboration with the Spector group at Cold Spring Harbor, investigated chromatin changes in response to a large genomic deletion.*

Goldman group

- *With support from the BBSRC, continued our work to re-purpose DNA as a medium for archiving digital information, developing computational and laboratory DNA-handling technologies needed to bring DNA-storage closer to market.*

- *Investigated how the presence of gaps in a multiple sequence alignment (MSA) affects the accuracy of inferred phylogenies, and derived a new proof of statistical consistency of maximum likelihood phylogenetic reconstruction for un-gapped alignments;*

- *To improve 'cell lineage trees', extending existing phylogenetic methods to cope with error-prone data,, developed a scalable algorithm that gives more accurate trees than those produced using standard methods;*

- *Investigated the impact of popular MSA programs on reconstruction of ancestral sequences, and discovered that different aligners introduce biases;*

- *Published the results of a long-running comparative study of MSA filtering methods, and showed that alignment filtering generally worsens the resulting trees;*

- *Developed treeCl, a clustering method that groups loci that share a common evolutionary history and distinguishes sets that do not, offering insights into the underlying causes of incongruence.*

Marioni group

- *Improved and extended models for finding heterogeneously expressed genes using scRNA-sequencing;*

- *Used scRNA-seq to characterise the first cell fate decisions in the early embryo;*

- *Applied scRNA-seq to study heterogeneity in embryonic stem cells.*

Saez-Rodriguez group

- *Identified biomarkers of drug efficacy in cancer cell lines from multiple types of data: genomic, epigenomic and transcriptomic;*

- *Completed analysis of collaborative competitions (DREAM challenges), assessing the state of the art of prediction of drug toxicity and of inference of signalling networks in cancer cells;*

- *Developed a method to reconstruct signalling pathways from mass-spectrometry phospho-proteomic data, and applied it to the study of breast cancer.*

Stegle group

- *Developed advanced statistical fast models to test genetic effects across multiple phenotypes in populations;*

- *Surveyed the regulatory effect of structural variants in the context of the 1000 Genomes Project;*

- *Devised integrative analytical strategies to investigate interaction effects between multiple genetic variants and environmental factors;*

- *Developed one of the first statistical approaches to account for confounding factors in single-cell transcriptome sequencing studies;*

- *In applications to T-cell differentiation studies, we find that this method reveals otherwise masked subpopulations of cells.*

Teichmann group

- *Investigated the transcriptome profiles of mESCs in three different culture conditions representing different pluripotent states, and revealed additional pluripotency network genes, including Ptma and Zfp640;*

- *Developed a novel computational method, TraCeR, that links TCR sequence with transcriptional profiles in individual cells with high accuracy and sensitivity;*

- *Using mass spectrometry data and a large-scale analysis of structures of protein complexes, identified repeating patterns in the assembly transitions and created a new*

'Periodic Table' of protein complexes, providing a predictive framework for anticipating new, unobserved topologies of protein complexes;
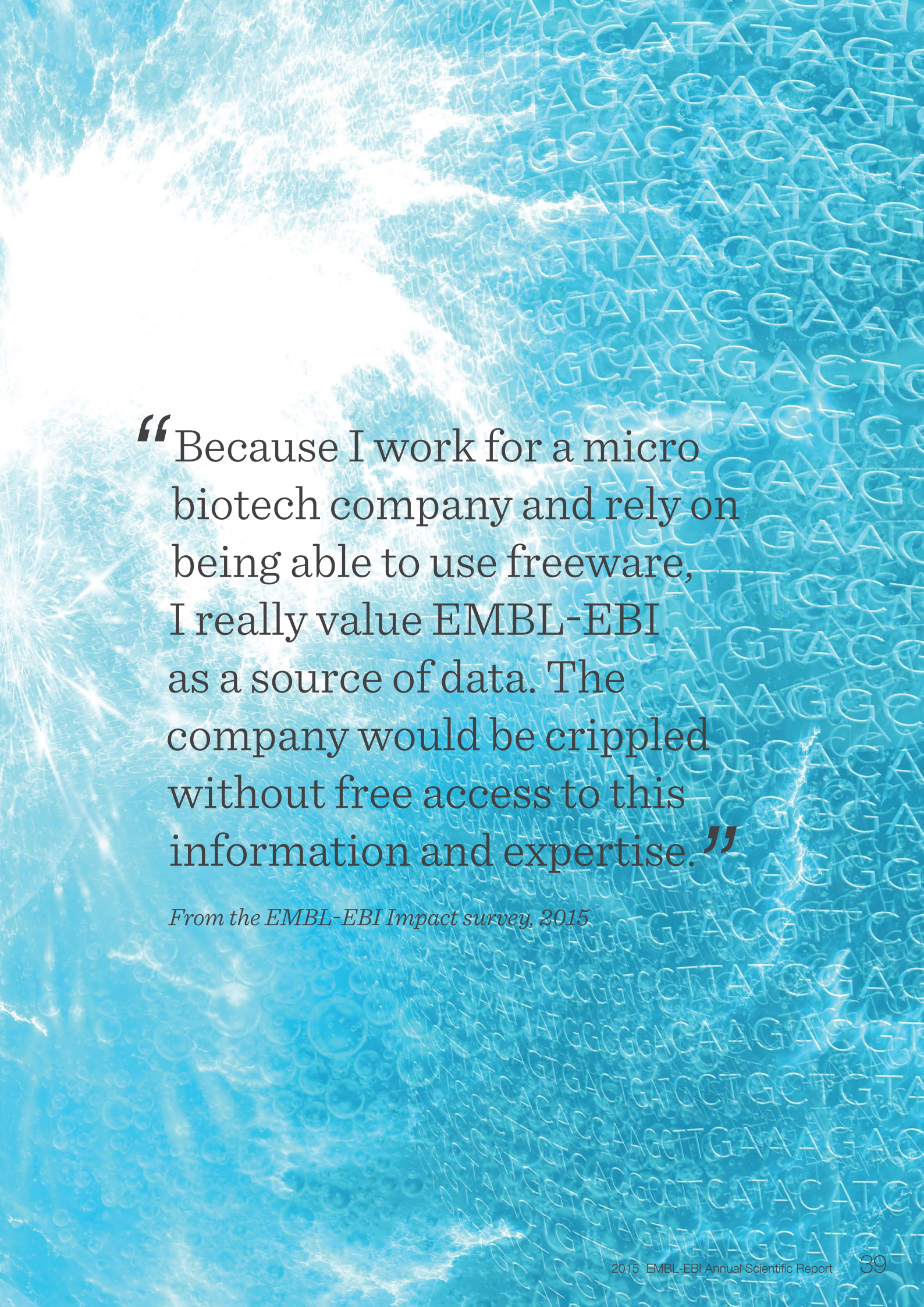
- *Developed a computational pipeline that provides a generic approach for processing scRNA-seq data and removing low-quality cells, ensuring that only correctly annotated cells are included in downstream analyses.*

Thornton group

- *Expanded understanding of the biochemistry and diversity of isomerisation by creating a robust method for comparison and clustering of isomerase reactions into classes;*

- *Studied enzyme function within 379 protein domain superfamilies, and provided insights useful for predicting the function of uncharacterised sequences and in the design of new synthetic enzymes;*

- *Conducted a genome-wide association study of Drosophila lifespan, and found that the top associated genes provide good candidates for further investigation into their relationship with lifespan and ageing;*

- *In collaboration with the Institute of Healthy Ageing at UCL, tested the effect of manipulating 10 of the genes most significantly associated with lifespan;*

- *In a combined computational–experimental study, presented the first confirmed example of a plant-pollen-like protein in a worm that is targeted by IgE, addressing an important question in allergy and parasitology.*

Industry and Innovation

"Because I work for a micro biotech company and rely on being able to use freeware, I really value EMBL-EBI as a source of data. The company would be crippled without free access to this information and expertise."

From the EMBL-EBI Impact survey, 2015

Industry, Innovation and Translation

EMBL-EBI supports researchers in biobusiness in a number of ways: directly through the provision of services and infrastructure, and indirectly by facilitating pre-competitive collaboration. Indeed, industry users of our services account for up to a quarter of data and tools usage.

Our collaborations with industry partners have increased significantly in recent years, most markedly with pre-competitive initiatives supporting drug development in large pharmaceutical companies. This has been made possible by the availability of collaborative working space within the EMBL-EBI South Building, funded through a large grant from the UK Government's Large Facilities Capital Fund in 2011 under the auspices of its Office of Business, Innovation and Skills.

Innovation and translation

The first of our Innovation & Translation activities, Open Targets (formerly the CTTV), launched its Target Validation platform in 2015, helping scientists discover and prioritise evidence-based relationships between targets and diseases. At its launch the service provided evidence for over 21 800 therapeutic targets spanning more than 8800 diseases and phenotypes. This comprehensive data service was built based on user-experience research, ensuring it is fit for purpose for wet- and dry-lab scientists alike, and in its first six weeks had over 9000 visits.

Originally formed by GSK, the Wellcome Trust Sanger Institute and EMBL-EBI, Open Targets fosters deep, on-going interactions between academic and industry members for the purpose of developing open, transformative approaches to selecting and validating novel targets in drug development. Following the interim directorship of Ewan Birney, Jeffrey Barrett of the Wellcome Trust Sanger Institute was appointed Director of the partnership in 2015.

EMBL-EBI's role in the design, development and implementation of the Target Validation platform cannot be understated. For example, the Samples, Phenotypes and Ontologies team enhanced and integrated ontology services and content supporting disease annotations, making it possible for this public–private partnership to integrate rare and common diseases according to shared phenotype. Our Web Development team led user-experience research as well as website design, development and deployment.

In all, 60 EMBL-EBI staff work on some 30 Open Targets projects, which range from computational pipelines to oncology, induced-pluripotent stem cells and single-cell genomics. Results of the experimental systems set up in 2015 are expected to be published during 2016.

Industry programme

We were very pleased to welcome Astex Pharmaceuticals, part of Otsuka Pharmaceuticals, as a new member of the Industry Programme in December 2015. Our Programme provided neutral ground for bioinformatics specialists in large companies to meet and address shared challenges, at quarterly meetings and 10 member-driven workshops with topics ranging from immunogenomics to Semantic Web applications.

In 2015 our Industry Programme organised a two-day workshop in Argentina, EMBL's newest associate member state. In collaboration with the Argentine Ministry of Science, Technology and Productive Innovation and the Argentine Chamber of Biotechnology, we delivered an event focussed on applications of bioinformatics and genomics in healthcare, agriculture and livestock breeding.
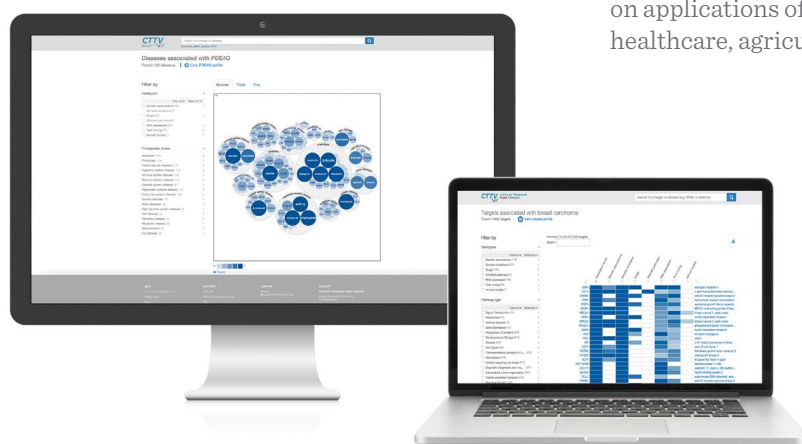
The Target Validation Platform, launched in 2015, is designed based on user-experience research.
www.targetvalidation.org

Industry workshops

- *In silico ADMET prediction*
- *Immunogenomics*
- *Translational NGS*
- *Enabling the translational bioinformatician*
- *Quantitative systems pharmacology*
- *Data enhancement through scientific literature: integrating the literature with data to enable discovery*
- *The EMBL-EBI RDF Platform*
- *Electronic Medical Records for Drug Discovery:*
- *Connectivity Map and LINCS (organised in association with The Broad Institute)*

Supporting companies of all sizes

Our collaborations are international, interdisciplinary and cross-sector in scope. Ten of our scholarly publications in 2015 were in collaboration with companies, and we worked with industry on a range of projects, for example expanding and improving mapping to the Gene Ontology with F. Hoffmann-La Roche.

Our service teams worked on several innovative projects with SMEs. The PDBe Content and Integration and Samples, Phenotypes and Ontologies teams collaborated with Flax on the BioSolr project, which aims to advance technologies to explore biomedical data in open-source software. Our Literature Services team collaborated with Publons, making it easier for authors to take credit for their work, and with Kudos, a company offering lay descriptions of research articles.

45% of users
said they could not have neither have created/ collected the last data they used themselves, nor obtained it elsewhere

£920m/year
EMBL-EBI data and services contributed to the wider realisation of future research impacts worth £920m annually or £6.9b over 30 years (NPV)

£335m/year
EMBL-EBI data and services directly underpinned an estimated £335m of research last year or £2.5b over 30 years(NPV)

£1bn to £5bn
Direct efficiency impact of EMBL-EBI data, representing direct worth of between £5,382 to £26,000 per user per annum

Chief Medical Officer for England, Professor Dame Sally Davies, delivering the keynote address at BioBeat15, Translating genomics into biobusiness: Defining the "Why Now?"

Making connections

In 2015 we co-organised the BioBeat conference: "Translating Genomics Into Biomedicine" - a major event featuring Innovate UK Chief Executive Ruth McKernan, Chief Medical Officer for England Sally Davies and over a dozen leaders in biobusiness. Enabled by BioBeat and the Wellcome Genome Campus Sex in Science Programme, the event attracted 300 attendees. We co-organised the Cambridge New Therapeutics Forum, a bimonthly networking event that attracts over 100 attendees, and co-organised a workshop for SMEs with InnovateUK on integrated 'omics. Our Industry Programme and OneNucleus also jointly organised the annual SME Bioinformatics Forum, which showcases our services in the context of innovations by local companies.

Impact

In 2015 our newly formed Strategic Project Management Office worked with an external management consultancy, Charles Beagrie Ltd., to facilitate a large-scale, economic analysis of the institute's impact on research practice and the global economy. This work, encouraged by the BBSRC, fed into a new framework for impact assessment. The report included a survey of over 4000 data service users, 45% of whom indicated that they could neither have created nor collected the last data they used themselves, nor obtained it elsewhere. The findings demonstrate the vital role of public databases in life-science and biomedical research, and indicate that for every million invested in EMBL-EBI, roughly 20 million is returned to the global economy.

NPV, net present value. Figures are from an external assessment of EMBL-EBI services undertaken in 2015.
See http://bit.ly/embl-ebi-impact

SERVICE TEAMS

TRAINING PROGRAMME

RESEARCH GROUPS

INDUSTRY PROGRAMME

EXTERNAL-FACING ACTIVITIES

CONTENTS

Technical Services, External Relations and Administration

The work of our institute relies on the infrastructure maintained by our technical service teams, depends on administrative support and benefits from engagement with diverse communities.

Technical Services, External Relations & Administration Highlights

Our Technical Services Cluster comprises four teams, delivering a broad portfolio of IT services that support the service provision, research and administrative activities of EMBL-EBI. External Relations handles public relations and communications for the institute as a whole, engaging with diverse stakeholders on multiple platforms and in person. Administration facilitates the work of the institute by contributing to the EMBL-wide implementation of efficient administrative processes, enabling the effective deployment and development of resources within a complex regulatory environment.

External Relations
Lindsey Crosswell

We support the work of EMBL-EBI's many ambassadors, in particular the institute's Directors and team leaders, in fostering good relations with policymakers, funders, potential collaborators and service users throughout the world. We work with leadership to refine and deliver key messages, and host visiting delegations of scientists, politicians and industry representatives. We endeavour to raise the profile of the EMBL-EBI brand by generating high-quality content and disseminating it through the press, our top-level website, social media channels, newsletters and printed publications. We also provide editorial and graphic design support to individuals throughout the organisation, helping them raise awareness of EMBL-EBI data services.

- *Organised BioBeat15: Translating Genomics Into BioBusiness, a one-day conference featuring key business leaders and entrepreneurs, under the auspices of the campus Sex in Science programme;*
- *Organised an event for life-science funders and policymakers aboard the Tara research vessel in London, on its way to the COP21 climate-change talks in Paris;*
- *Wrote and distributed 33 press releases and other news stories, managed social media channels, supported news writers in other teams and colleagues in Heidelberg, hosted 15 journalist and film crew visits;*
- *Created and distributed print and digital publications including the EMBL-EBI Annual Scientific Report, overview brochure, Innovation and Translation brochure and others, and contributed to EMBL-wide publications including Research at a Glance, the EMBL Annual Report and the Indicative Scheme;*
- *Served as a point of reference for branding, photography, logo design and graphics for services, and published cover artwork for the journal Genome Research.*

Technical Services
Steven Newhouse

We provide the institute's physical infrastructure, with over 60 Petabyes of disc storage, a 30 000-core cluster to support advanced data analytics, a 36-node Hadoop cluster and a high-performance networking infrastructure linking three data centres. Our teams also operate a new OpenStack cloud infrastructure that complements existing cloud and virtual infrastructure. Together, these support the smooth running of over 450 Terabytes of SQL and NoSQL databases.

Our technical teams deliver over 2000 Web Services, and manage 700 virtual machines dedicated to service teams and research groups that provide a variety of search functionality and access points to a diverse range of life-science data resources. We also design and develop the main EMBL-EBI website, provide User Experience Research consultation to internal teams and develop websites for strategic projects such as Open Targets.

In 2015 we launched an integrated information and support service for EMBL-EBI staff, greatly improving communication and response time. We trained over 30 members of staff on IT service management processes, and initiated staff inductions for desktop services.

Our teams played an important role in the ELIXIR Excelerate project, developing and delivering bioinformatics services and infrastructure. We also investigated hybrid cloud computing models as a supplement to our internal infrastructure. Our Technology Science Integration group participated in several influential EU-funded projects, including EUDAT2020, BioExcel, PhenoMeNal, EGI-Engage and Helix Nebula Science Cloud. This work is supported in part by funding from from Cancer Research UK.

Web production

Rodrigo Lopez

- *Indexed over 1.5 billion data records from EMBL-EBI data resources;*
- *Handled more than 150 million job requests via Job Dispatcher (36% increase over 2014);*
- *Hosted approximately 13 000 distinct sequence libraries, primarily in the ENA, UniProt and Ensembl Genomes;*
- *Managed FTP and Aspera traffic, which totalled 148.1 Petebytes;*
- *Improved the EBI Search, which is used in 20 EMBL-EBI data resources, by reducing its memory footprint through virtualisation and gaining efficiencies in its core Lucene libraries;*
- *Participated in 14 outreach and training events.*

Web development

Brendan Vaughan

- *Launched an integrated events portal;*
- *Redesigned the Training Programme web pages, helping users find events of interest easily;*
- *Re-engineered the global EMBL-EBI website to be responsive on mobile devices;*
- *Designed and launched a fully functional, UX-driven Target Validation Platform;*
- *Moved the platform, Open Targets website and Intranet to an Embassy Cloud workspace;*
- *Created a dedicated Technical Services information portal;*
- *Upgraded or moved 14 Drupal6 websites and portals;*
- *Provided User Experience Research consultation to teams and groups across the institute.*

Systems applications

Andy Cafferkey

- *Deployed a new Delphix based infratructure for our databases across our distributed data centres. This work was recognized with a 'Best Data Centre Project of the Year' award at the Computer Weekly European User Awards 2015.*
- *Maintained 2000 virtual machines, with 100% uptime;*
- *Commissioned and maintained a new OpenStack infrastructure for Embassy Cloud;*
- *Contributed to the operation of eMedLab project;*
- *Supported the 950 client devices in use by EMBL-EBI internal users across OSX, Windows and Linux platforms;*
- *Began to join the Embassy Cloud with the European Grid Infrastructure as part of the ELIXIR-Excelerate;*
- *Consolidated all database instances to virtual machines on a shared infrastructure, reducing the data-centre capacity required for databases.*

Systems infrastructure

Petteri Jokinen

- *Upgraded the core network links in Hinxton from 10 to 40 Gigabits per second;*
- *Re-implemented the vault firewall for the EGA using a software approach, substantially improving system performance;*
- *Rebuilt the EMBL-EBI compute clusters' General Parallel File System, to improve the high-speed file access.*
- *Enabled internal users to run Docker containers, resulting in improved efficiency and more consistent performance;*
- *Deployed a Lustre (Linux Cluster) for the Functional Annotation of Animal Genomes project;*
- *Maintained EMBL-EBI's growing compute infrastructure, now based on ~30 000 Central Processing Unit (CPU) cores*

Administration

Mark Green

In 2015 all of the Administration sub-teams examined processes and procedures in their immediate areas and in activities that contribute to the overall development of EMBL-EBI and EMBL. This led to improvements that are deeply appreciated by those whom administration serves.

- *Established the Strategic Project Management Office, which supports on-going infrastructure projects;*
- *Contracted and contributed to a large-scale, independent analysis of the economic impact of the institute, and finalised the report for distribution;*
- *Initiated a Project Management Network to support project coordinators in teams throughout the institute;*
- *Introduced flexible working arrangements for EMBL-EBI staff;*
- *Managed a nearly 20% increase in recruitment;*
- *Welcomed a member of the EMBL Budget Office to EMBL-EBI, improving the coordination of Finance activities with the EMBL Grants Office.*

Facts and Figures

Funding and Resource Allocation

EMBL-EBI funding remained stable in 2015. This continued support of our member states and other funding bodies helped us retain staff, maintain our core public resources and, thanks to additional support from the UK Government, absorb the doubling of the data we store in our archives.

Here we show our sources of funding, and how we spent these funds in 2015. The 'external funds' shown here represent funds that were available for our use in 2015, but not those earmarked for subcontractors, as part of our grant-funded activities.

Sources of funding

Funding for EMBL-EBI in 2015, excluding sums earmarked for project subcontractors, was €73 million and comes primarily from EMBL member states. Our major sources of external funding include the European Commission (€5.5 million), the Wellcome Trust (€5.4 million), the US National Institutes of Health (€7.1 million), the UK Research Councils (€5.6 million) and the EMBL-EBI Industry Programme (€0.8 million). We also benefit from a large number of grants from various other sources (total, €3.8 million). These major sources of funding are shown in Figure 1.
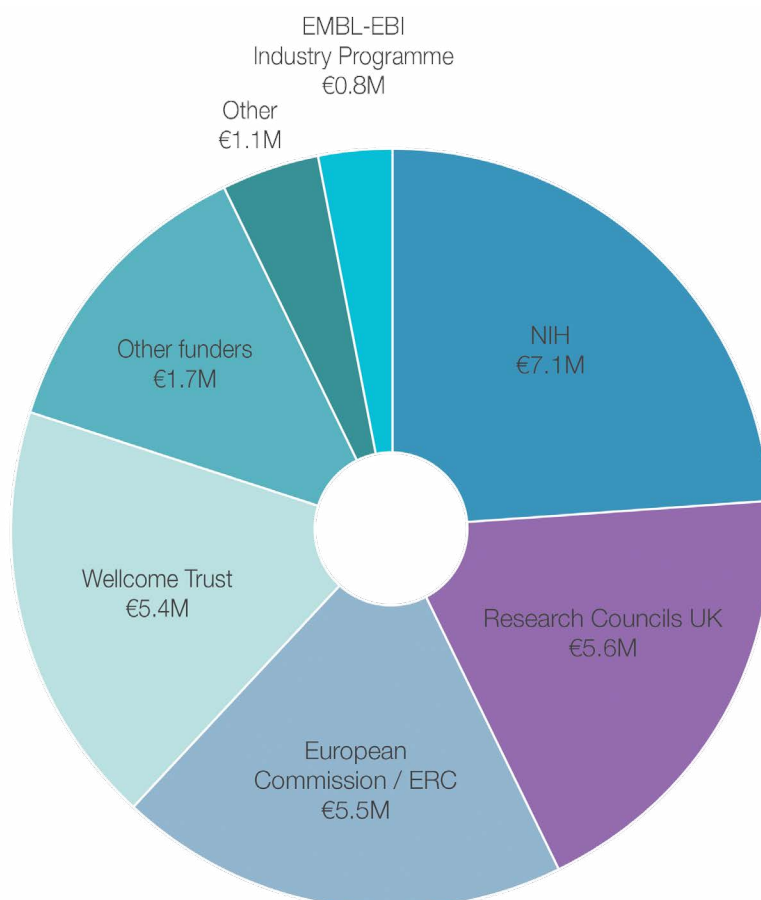
* Figures exclude external funds passed straight through to grant subcontractors. In addition to these sums, EMBL-EBI's funding in 2015 included funds earmarked for project subcontractors of €4.5 million (NIH), €0.6 million (Wellcome Trust) and € 0.1 million (others).

Capital investment

## Support from the United Kingdom Government's Large Facilities Capital Fund

The UK Government's Large Facilities Capital Fund has provided for the EMBL-EBI South Building, which houses ELIXIR and an Innovation and Translation Suite, and for the on-going use of Tier III+ data centres and the equipment to enable data service provision.

Year	Data centre capacity	Funding for Technical Hub	Total funding received
2013	€ 10.363 M	€ 15.4 M	€ 22.45 M
2014	€ 7.4 M	€ 1.7 M	€ 9.1 M
2015	€ 8.1M	€ 0.3 M	€ 8.4 M

Table. Support from the UK Government's Large Facilities Capital Fund.

Spending

Figure 2 shows the breakdown of EMBL-EBI's total spend for 2015 (€53.48 million, excluding sums expended by grant subcontractors). Our largest expenditure was staff (73%), followed by running costs (23%).
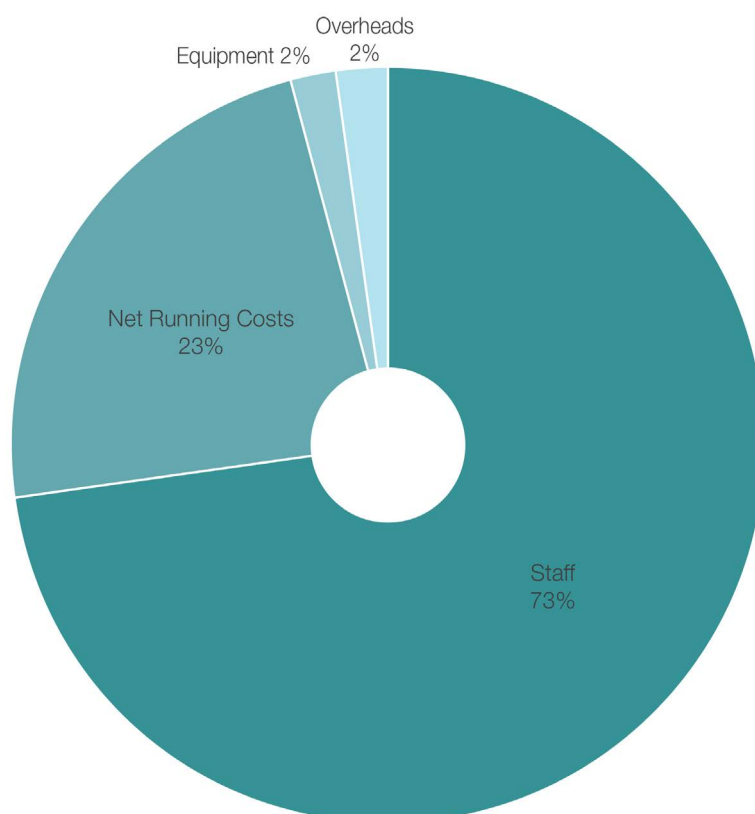
Growth of Core Resources

On an average weekday at the end of 2015 we saw in excess of 16 million requests to our websites. The absolute number of requests is influenced heavily by the way the website is implemented, so these figures should be viewed as an indicator of trends, and the trend clearly shows continued growth in use of our websites over time. It is even more difficult to extract information on the number of users from web logs (often a whole organisation appears as a single user); however, from our records we see that there were 562 million requests to our sites each month, and in 2015 over nine million unique hosts accessed our pages.

In 2015 we ran more than 151 million jobs on the Dispatcher framework, which helps users establish analytical pipelines to automatically query data and combine it with their own. Job Dispatcher also helps users who are exploring the data in a more organic way. The average number of jobs per month was 12.6 million (comare to 9.2 million in 2014). The most heavily used tools were InterProScan, NCBI Blast, ClustalW, and Needle.
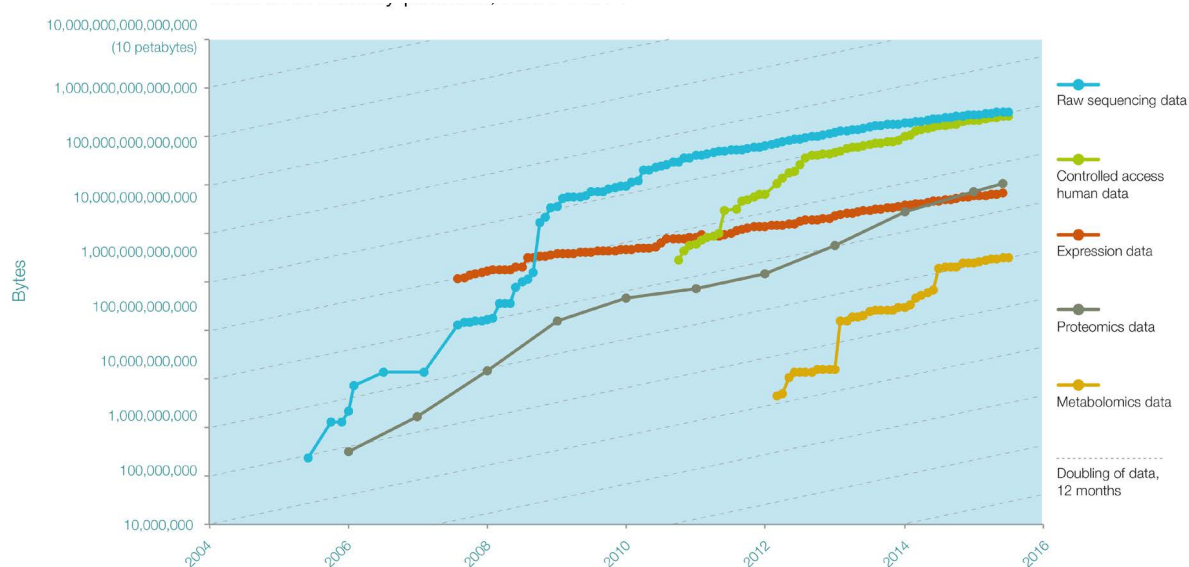
Growth of data platforms at EMBL-EBI, 2004 through 2015
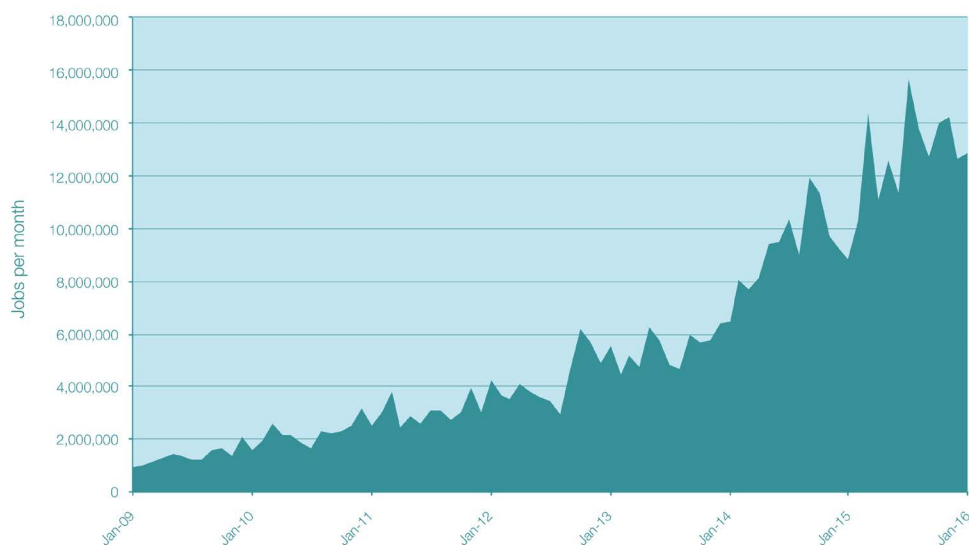
Requests per day on EMBL-EBI services, 2009 through 2015

Nucleotide sequence data (compressed)

Genomes (all species)

Gene expression data

Protein sequence data

Macromolecular structures

Protein families motifs and domains
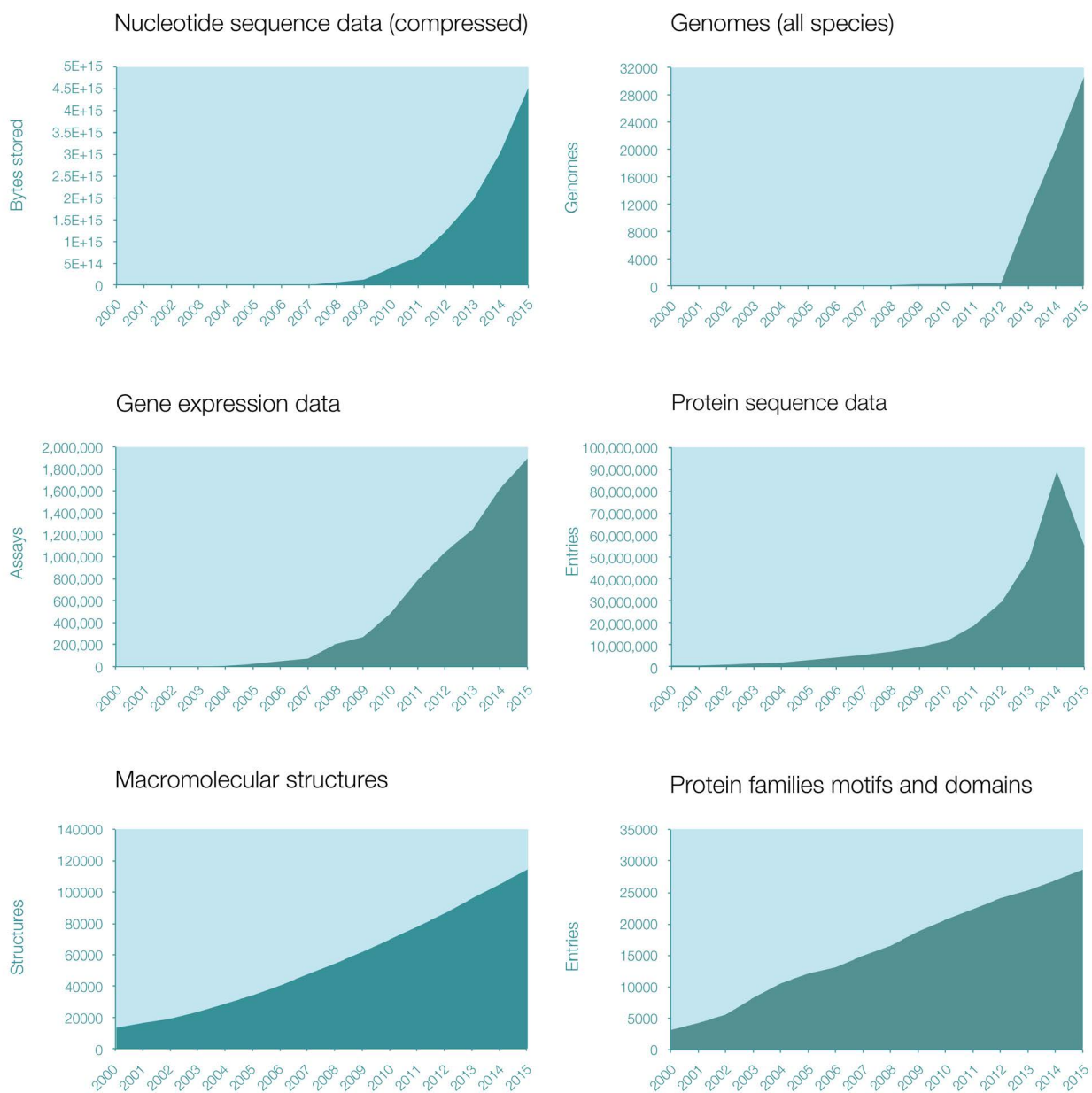
In 2015 our core data resources continued their steady growth. The cost of generating data continued to fall, which has a dramatic impact on EMBL-EBI databases as it enables researchers to generate more data. EMBL-EBI continues to develop and implement innovative data-storage methods.

- *Nucleotide sequence data: 4.52 Petabytes stored (compare to 3.06 petabytes in 2014). A Petabyte is 1 x 10^{15} bytes;*

- *Genomes, all species and strains: 30 674 (compare to 20 343 in 2014);*

- *Gene expression assays: 1.90 million (compare to 1.62 million in 2014);*

- *Protein sequences: 55.3 million (compare to 89.1 million in 2014). The large reduction is due to removal of millions of identical sequences from closely related organisms to reduce response times for user queries;*

- *Macromolecular structures: 114 691 (compare to 105 444 in 2014);*

- *Protein families, motifs and domains—entries in InterPro: 28 678 (compare to 27 002 in 2014).*

Scientific Collaborations

We work with communities throughout the world to establish standards, exchange information, improve methods for analysis and share 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 2015, EMBL-EBI had joint grant funding with researchers and institutes in 52 countries throughout the world, most notably in the United Kingdom, Germany, France and the Netherlands but also with colleagues in countries with modest research communities such as Senegal. Of the 162 grants received, 51 were exclusively for EMBL. These figures are potentially underestimated, as all partners are not always listed on grants.
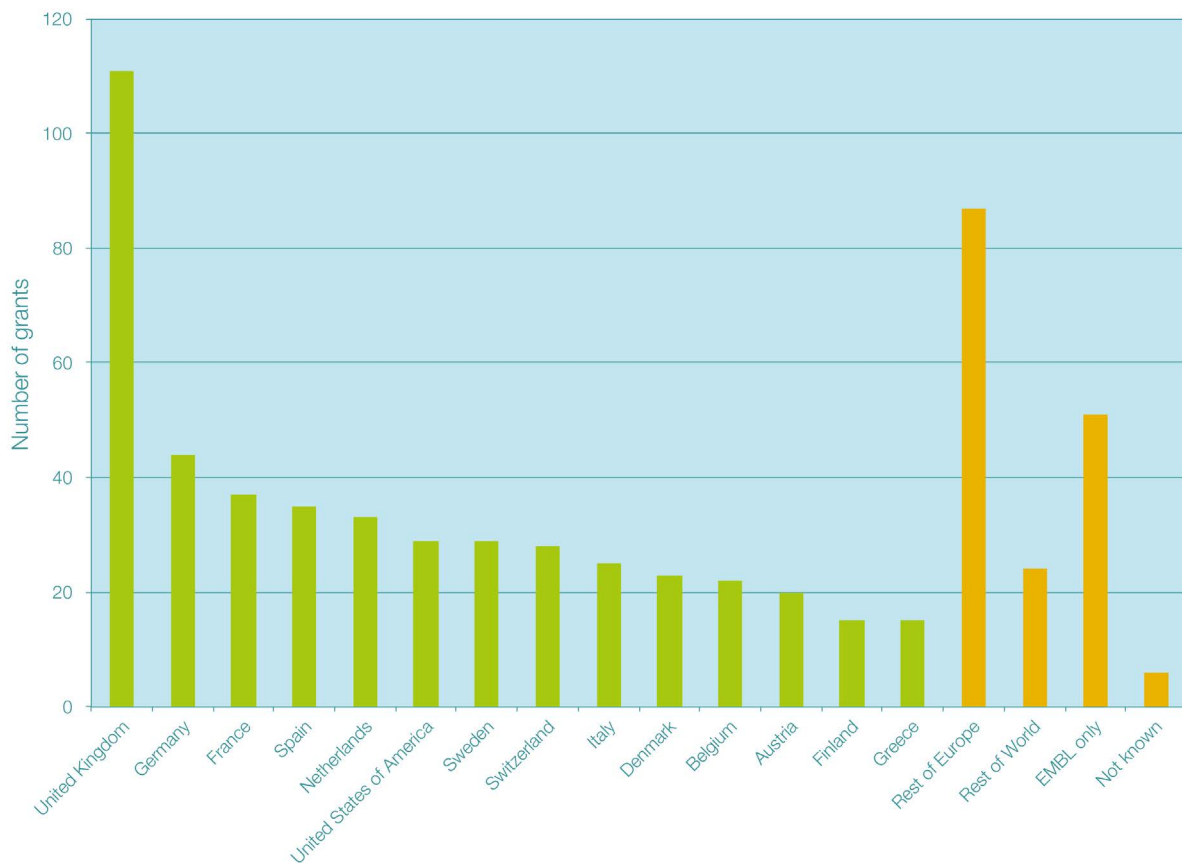
Joint publications

Most of our 299 articles published online in 2015 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, the Netherlands, Spain and France, and our collaborations extended well beyond Europe to Argentina, Brazil and South Africa.
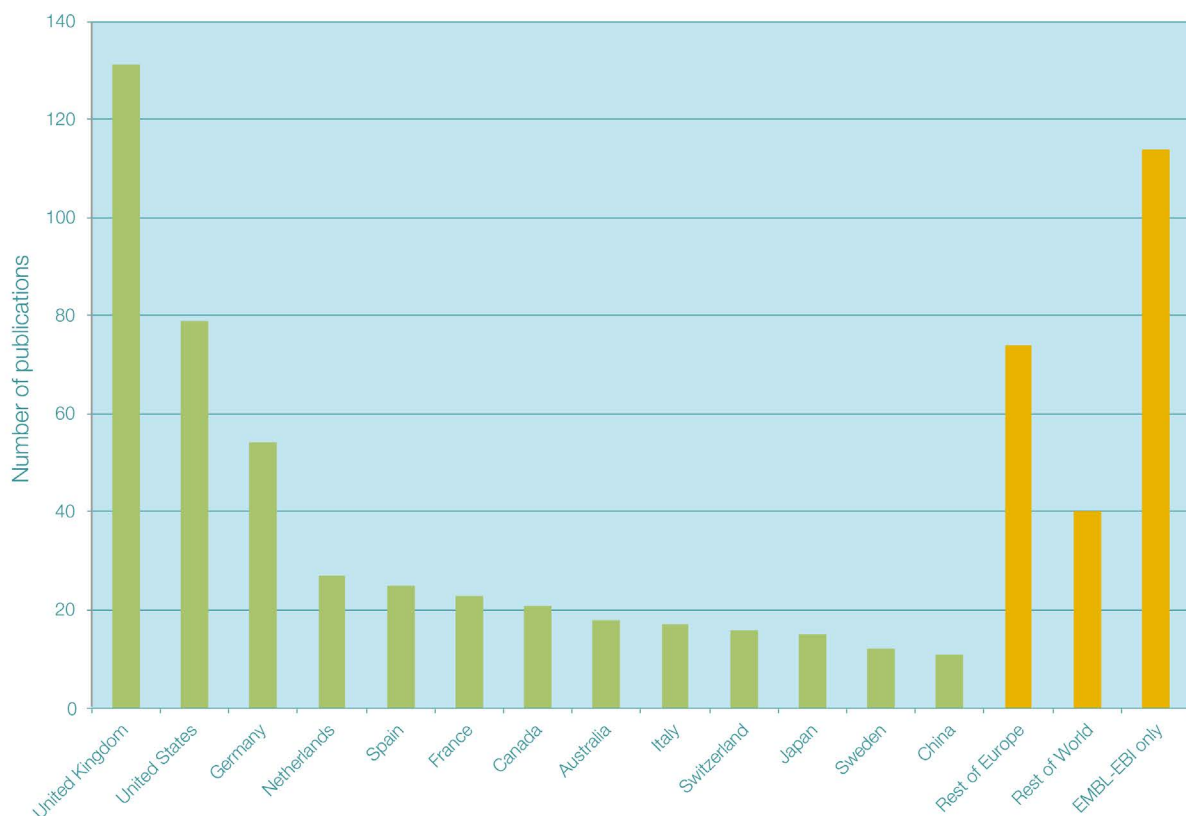
Our Staff in 2015

We are proud to report that in 2015, our staff represented 62 nationalities (compare to 57 in 2014). We had 513 members of staff in 2015, and hosted 100 visitors and 51 trainees. These visitor figures include those who joined us for longer than one month (compare to 91 in 2014).

Nationalities represented at EMBL-EBI in 2015
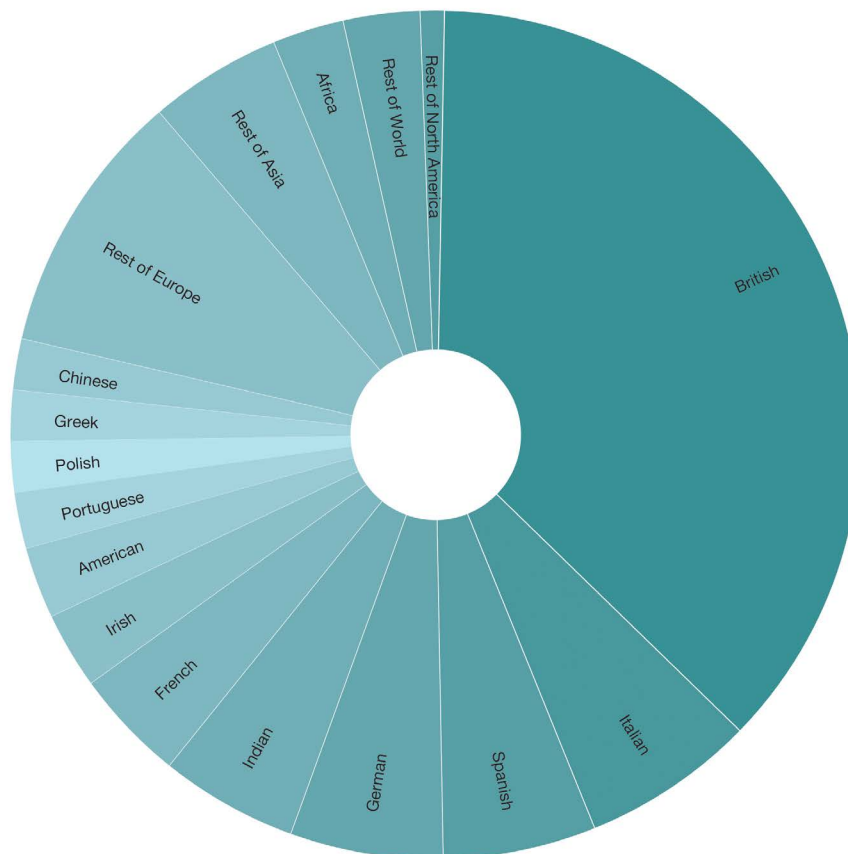
Staff growth at EMBL-EBI, 1998 to 2015
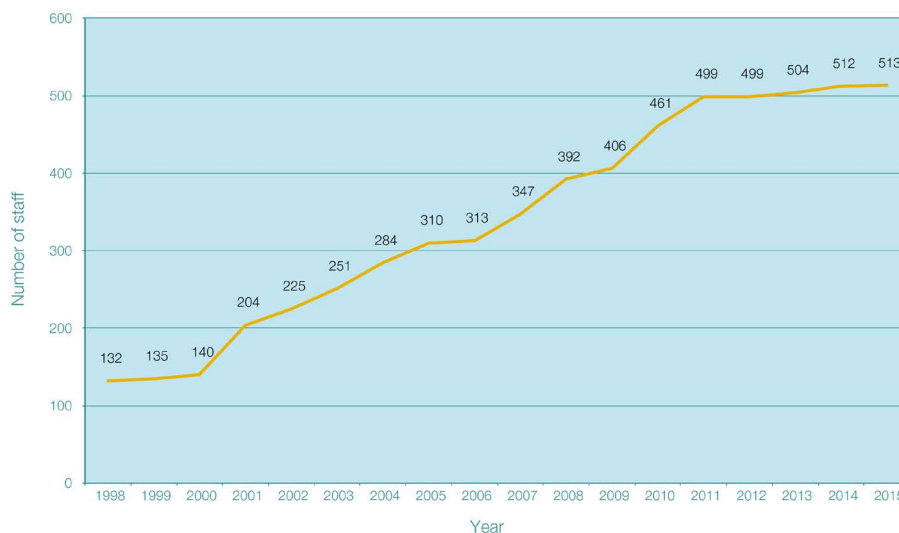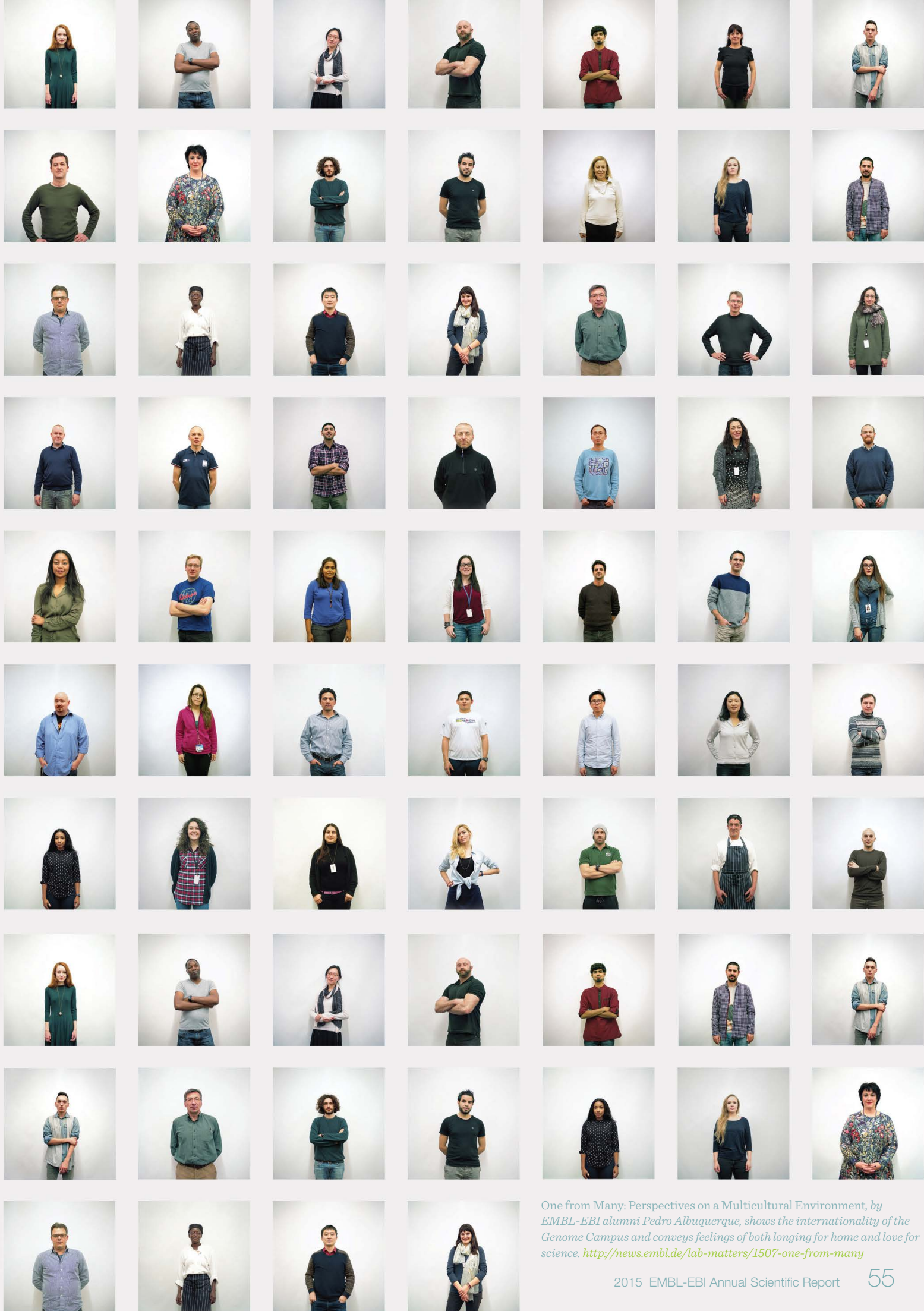
One from Many: Perspectives on a Multicultural Environment, *by EMBL-EBI alumni Pedro Albuquerque, shows the internationality of the Genome Campus and conveys feelings of both longing for home and love for science.* http://news.embl.de/lab-matters/1507-one-from-many

Scientific Advisory Committees

Scientific Advisory Committees

ArrayExpress and Expression Atlas

- Roderic Guigo Serra, Centre de Regualció Genòmica, Barcelona, Spain
- Frank Holstege, University Medical Center Utrecht, the Netherlands
- Wolfgang Huber, EMBL, Germany
- Jill Mesirov, Broad Institute of MIT and Harvard, Cambridge, United States
- Chris Ponting, University of Oxford, United Kingdom

BioModels

- Carole Goble, University of Manchester, United Kingdom
- Thomas Lemberger, Nature Publishing Group/EMBO
- Pedro Mendes, University of Manchester, United Kingdom
- Wolfgang Mueller, HITS, Germany
- Philippe Sanseau, GSK, United Kingdom

Cheminformatics: ChEMBL and ChEBI

- Steve Bryant, National Institutes of Health, United States
- Edgar Jabcoby, Novartis, Basel, Switzerland
- Andrew Leach, GlaxoSmithKline Plc, United Kingdom (Chair)
- Tudor Oprea, University of New Mexico, Albuquerque, United States
- Alfonso Valencia, CNIO, Madrid, Spain
- Peter Willett, University of Sheffield, United Kingdom

EMDataBank

- Paul Adams, Lawrence Berkeley Laboratory, United States (Chair)
- Richard Henderson, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom
- Bram Koster, Leiden University Medical Center, the Netherlands
- Maryanne Martone, University of California San Diego, United States
- Andrej Sali, University of California San Francisco, United States

Ensembl

- Ian Bird, CERN, Switzerland
- Deanna Church, Personalis, Palo Alto, United States

- Federica Di Palma, Vertebrate and Health Genomics, The Genome Analysis Centre, Norwich, United Kingdom
- Mark Diekhans, Center for Biomolecular Science & Engineering, University of California Santa Cruz, United States
- Anne Ferguson-Smith, Department of Physiology, Development and Neuroscience, Cambridge University, Cambridge, United Kingdom
- Ivo Gut, Centro Nacional de Análisis Genómico, Barcelona, Spain
- Matt Hurles, Genomic Mutation and Genetic Disease, Wellcome Trust Sanger Institute, Hinxton, United Kingdom
- Erich Jarvis, Department of Neurobiology, Duke University Medical Center, Durham, United States
- Felicity Jones, Friedrich Miescher Laboratory, Tuebingen, Germany
- Jim Reecy, Department of Animal Science, Iowa State University, Ames, United States

Ensembl Genomes

- Martin Donnelly, University of Liverpool, United Kingdom
- Klaus Mayer, Helmholtz Institute for Pharmaceutical Research, Saarland, Germany
- Claudine Medigue, Genoscope, France
- Allison Milller, University of St. Louis, United States
- Rolf Mueller, Helmholtz Institute, Bonn, Germany
- Chris Rawlings, Rothamsted Research, United Kingdom
- Jason Staijich, University of Riverside, United States
- Denis Tagu, INRA, Rennes, France

European Genome-phoneme Archive

- Dixie Baker, Martin-Blanck and Associates, Alexandria, Virginia, United States
- Jan-Willem Boiten, Dutch Techcentre for Life Sciences, Eindhoven, the Netherlands
- Anne Cambon-Thomsen, Institut National de la Santé et de la Recherche Medicale (INSERM), Toulouse, France
- Joaquin Dopazo, Centro de Investigación Príncipe Felipe, Valencia, Spain
- Teri Manolio, National Human Genome Research Institute, Bethesda, United States
- Gil McVean, Oxford University, United Kingdom

European Nucleotide Archive

- *Mark Blaxter, University of Edinburgh, United Kingdom (Chair)*

- *Antoine Danchin, CNRS, Institut Pasteur, Paris, France*

- *Frank Oliver Glöckner, Max Planck Institute for Marine Microbiology, Bremen, Germany*

- *Tim Hubbard, King's College London, United Kingdom*

- *Babis Savakis, University of Crete & IMBB-FORTH, Heraklion, Greece*

- *Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin, Germany*

- *Patrick Wincker, Genoscope, Evry, France*

The Gene Ontology

- *Philip Bourne, National Institutes of Health, Bethesda, United States*

- *Richard Scheuermann, University of Texas Southwestern Medical Centre, Dallas, United States*

- *Michael Schroeder, Technische Universität Dresden, Germany*

- *Barry Smith, SUNY Buffalo, United States*

- *Olga Troyanskaya, Department of Computer Science and Molecular Biology, Princeton University, United States*

- *Michael Tyers, Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Canada*

GWAS Catalog

- *Ines Barroso, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom*

- *Alexis Battle, Johns Hopkins University, Baltimore, United States*

- *Nancy Cox, Vanderbilt University, Nashville, United States*

- *Josh Denny, Vanderbilt University, Nashville, United States*

- *Mike Feolo, National Center for Biotechnology Information, Bethesda, United States*

- *Marylyn Ritchie, Pennsylvania State University, University Park, United States*

IntAct and Complex Portal

- *Pascal Braun, Technische Universität München, Germany*

- *Alex Jones, University of Warwick, United Kingdom*

- *Giovanni Cesareni, Tor Vergata University Rome, Italy*

- *Willem Ouwehand, Department of Haematology, University of Cambridge NHS Blood and Transplant Centre, United Kingdom*

- *Peter Woollard, GSK, Stevenage, United Kingdom*

International Genome Sample Resource

- *Piero Carninci, Riken, Japan*

- *Richard Durbin, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom*

- *Jane Kaye, Oxford University, United Kingdom*

- *Eimear Kerry, Icahn School of Medicine at Mount Sinai, New York, United States*

- *Jan Korbel, EMBL, Heidelberg, Germany*

- *Cisca Wijmenga, University of Groningen, the Netherlands*

International Nucleotide Sequence Database Collaboration International Advisory Committee, European chapter

- *Mark Blaxter, University of Edinburgh, United Kingdom*

- *Antoine Danchin, CNRS, Institut Pasteur, Paris, France (Chair)*

- *Babis Savakis, University of Crete and Institute of Molecular Biology and Biotechnology-Foundation for Research and Technology, Heraklion, Greece*

- *Jean Weissenbach, Genoscope, Evry, France*

InterPro/Pfam

- *Patrick Aloy, Institute for Research in Biomedicine, Barcelona, Spain*

- *Michael Galperin, National Center for Biotechnology Information, Bethesda, United States*

- *Nicola Mulder, Institue of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa*

- *Sean Munro, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom*

- *Erik Sonnhammer, Stockholm University, Sweden (Chair)*

- *Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain*

Scientific Advisory Committees

Literature Services Scientific Advisory Board

- *Terry Attwood, University of Manchester, United Kingdom*
- *Theo Bloom, British Medical Journal, London, United Kingdom*
- *Jan Brasse, Göttingen State and University Library, Germany*
- *Martin Fenner, Hannover Medical School Cancer Center, Public Library of Science and Technical and Human Infrastructure for Open Research/DataCite, Germany*
- *Tim Hubbard, King's College London, United Kingdom*
- *Jenny Malloy, Univeristy of Cambridge and ContentMine, Cambridge, United Kingdom*
- *Patrick Ruch, University of Applied Sciences, Geneva, Switzerland*

Metagenomics Scientific Advisory Board

- *Mark Blaxter, University of Edinburgh, United Kingdom (Chair)*
- *Chris Bowler, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France*
- *Mark Forster, Syngenta, Cambridge, United Kingdom*
- *Eric Pelletier, Genescope, Paris, France*
- *Phil Poole, University of Oxford, United Kingdom*

Open Targets (formerly Centre for Therapeutic Target Validation)

- *Bissan Al-Lazikani, The Institute of Cancer Research, London, United Kingdom*
- *Søren Brunak, Technical University of Denmark, Kongens Lyngby, Denmark*
- *Fiona Marshall, Heptares Therapeutics, Welwyn Garden City, United Kingdom*
- *Robert M. Plenge, Merck Research Laboratories, Boston, United States*
- *Berent J. Prakken, Children's Hospital, University Medical Centre Utrecht, the Netherlands*
- *Rob G.J. Vries, Foundation Hubrecht Organoid Technology, Utrecht, the Netherlands*
-

Proteomics Identifications Database

- *Ruedi Aebersold, Swiss Federal Institute of Technology, ETH, Zurich, Switzerland*
- *Roz Banks, University of Leeds, United Kingdom*
- *Angus Lamond, University of Dundee, United Kingdom*
- *Kathryn Lilley, University of Cambridge, United Kingdom*
- *Juri Rappsilber, University of Edinburgh, United Kingdom*
- *Ioannis Xenarios, SIB Swiss Institute of Bioinformatics, Geneva, University of Lausanne, Switzerland*

Reactome Scientific Advisory Committee

- *Russ Altman, Stanford University, Palo Alto, United States*
- *Gary Bader, University of Toronto, Canada*
- *Richard Belew, University of California San Diego, United States*
- *Edda Klipp, Max Planck Institute for Molecular Genetics, Berlin, Germany*
- *Adrian Krainer, Cold Spring Harbor Laboratory, Cold Spring Harbor, United States*
- *Ed Marcotte, University of Texas at Austin, United States*
- *Mark McCarthy, Oxford University, United Kingdom*
- *Jill Mesirov, Broad Institute of MIT and Harvard, Cambridge, United States*
- *John Overington, Stratified Medical, London, United Kingdom*
- *Bill Pearson, University of Virginia, Charlottesville, United States*
- *Brian Shoichet, University of California San Francisco, United States*

RNA group (RNAcentral and Rfam) Scientific Advisory Board

- *Sean Eddy, Harvard University, Cambridge, United States*
- *Michele Meyer, Boston College, Boston, Massachusetts, United States*
- *John Rinn, Harvard Medical School, Boston, Massachusetts, United States*
- *Eric Westhof, University of Strasbourg, France*

Technical Services Cluster Scientific Advisory Board

- Rolf Apweiler, EMBL-EBI
- Ewan Birney, EMBL-EBI
- Alvis Brazma, EMBL-EBI
- Tony Cox, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
- Nick Goldman, EMBL-EBI
- Henning Hermjakob, EMBL-EBI
- Rupert Lueck, EMBL, Heidelberg, Germany
- Johanna McEntyre, EMBL-EBI
- Julio Saez-Rodriguez, EMBL-EBI
- Ugis Sarkans, EMBL-EBI

Training Programme Scientific Advisory Committee

- Alex Bateman, EMBL-EBI
- Bogi Eliasen, FarGen: the Faroe Genome Project, Faroe Islands, Denmark
- Mark Forster, Syngenta, Cambridge, United Kingdom
- Nick Goldman, EMBL-EBI
- Paul Kersey, EMBL-EBI
- Gos Micklem, University of Cambridge, United Kingdom
- Chris Ponting, University of Oxford, United Kingdom (Chair)

UniProt Scientific Advisory Committee

- Patricia Babbitt, University of California San Francisco, United States
- Helen Berman, Rutgers University, New Brunswick, United States
- Takashi Gojobori, National Institute of Genetics, Mishima, Japan
- Minoru Kanehisa, Institute for Chemical Research, Kyoto, Japan
- Maricel Kann, University of Maryland, Baltimore, United States
- Edward Marcotte, University of Texas, Austin, United States
- William Pearson, University of Virginia, Charlottesville, United States
- Lyne Regan, Yale University, New Haven, United States

- Philippe Sansaeau, GlaxoSmithKline, Stevenage, United Kingdom
- Paul Thomas, University of Southern California, Los Angeles, United States
- Mathias Uhlén, Royal Institute of Technology (KTH), Stockholm, Sweden (Chair)
- Timothy Wells, Medicines for Malaria Venture, Geneva, Switzerland

Worldwide Protein Data Bank Advisory Committee

- Paul Adams, Lawrence Berkeley Laboratory, Berkeley, United States
- Edward N. Baker, University of Auckland, New Zealand (Ex Officio)
- Manju Bansal, Indian Institute of Science, Bengaluru, India
- R. Andrew Byrd, National Institutes of Health, United States (Chair)
- Jianping Ding, Shanghai Institutes for Biological Sciences, China
- Wayne Hendrickson, Columbia University, United States
- Genji Kurisu, Institute for Protein Research, Osaka University, Japan
- Gaetano Montelione, Rutgers University, New Brunswick, United States
- Helen Saibil, Birkbeck College London, United Kingdom
- Soichi Wakatsuki, Stanford University, United States

Protein Data Bank in Europe Scientific Advisory Committee

- David Brown, University of Kent, Canterbury, United Kingdom
- Sarah Butcher, University of Helsinki, Finland
- Manuela Helmer Citterich, University of Rome Tor Vergata, Rome, Italy
- Tomas Lundqvist, Max IV Laboratory, Lund University, Sweden
- Michael Nilges, Institut Pasteur, Paris, France
- Randy J. Read, University of Cambridge, United Kingdom (Chair)
- Helen Saibil, Birkbeck College London, United Kingdom
- Michael Sattler, TUM, Munich, Germany
- Titia Sixma, Netherlands Cancer Institute, Amsterdam, the Netherlands

Major Database Collaborations

ArrayExpress

- *Gene Expression Omnibus, National Center for Biotechnology Information, Bethesda, United States*

BioModels Database

- *Database of Quantitative Cellular Signalling, National Center for Biological Sciences, Bengaluru, India*
- *JWS Online, Stellenbosch University, South Africa*
- *Physiome Model Repository, Auckland Bioengineering Institute, New Zealand*
- *The Virtual Cell, University of Connecticut Health Center, Farmington, United States*

ChEBI

- *ChemIdPlus, National Library of Medicine, Bethesda, United States*
- *DrugBank, University of Alberta, Edmonton, Canada*
- *Immune Epitope Database (IEDB) at La Jolla Institute for Allergy and Immunology, United States*
- *KEGG Compound, Kyoto University Bioinformatics Centre, Kyoto, Japan*
- *OBI Ontology Consortium*
- *PubChem, National Institutes of Health, Bethesda, United States*
- *UniPathways, Swiss Institute of Bioinformatics, Geneva, Switzerland*

ChEMBL

- *BindingDB, University of California San Diego, United States*
- *CanSAR, Institute of Cancer Research, London, United Kingdom*
- *PubChem, National Center for Biotechnology Information, Bethesda, United States*

Ensembl

Here we list collaborations with the major genome centres and representative collaborations for the human, mouse, rat and chicken genomes. There are many others.

- *Baylor College of Medicine, Houston, United States*
- *Broad Institute, Cambridge, United States*
- *DOE Joint Genome Institute, Walnut Creek, United States*
- *Ensembl at the Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom*

- *Genome Browser at the University of California, Santa Cruz, United States*
- *Map Viewer at the National Center for Biotechnology Information, Bethesda, United States*
- *Mouse Genome Informatics at the Jackson Laboratory, Bar Harbor, United States*
- *Rat Genome Database at the Medical College of Wisconsin, Milwaukee, United States*
- *The Roslin Institute, Midlothian, Scotland, United Kingdom*

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, Harpenden, 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*
- *WormBase, a collaboration with the California Institute of Technology and Washington University, United States; Ontario Institute for Cancer Research, Canada; Wellcome Trust Sanger Institute and Oxford University, United Kingdom*

European Nucleotide Archive (ENA)

The ENA is part of the International Nucleotide Sequence Database Collaboration. Other partners include:

- *National Center for Biotechnology Information, Bethesda, United States (GenBank, Trace Archive and Sequence Read Archive)*
- *National Institute of Genetics, Mishima, 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, Corvallis, United States

- Cold Spring Harbor Laboratory, Cold Spring Harbor, United States

- Rothamsted Research, Harpenden, United Kingdom

- Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

Europe PubMed Central

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

Gene Ontology Consortium

- Agbase, Mississippi State University, Starkville, United States

- The Arabidopsis Information Resource, Carnegie Institution of Washington, Stanford, United States

- Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, Berkeley, United States

- British Heart Foundation, University College London, London, United Kingdom

- Candida Genome Database, Stanford University, Stanford, United States

- DictyBase at Northwestern University, Chicago, United States

- EcoliWiki

- FlyBase at the University of Cambridge, United Kingdom

- GeneDB S. pombe and GeneDB for protozoa at the Wellcome Trust Sanger Institute, Hinxton, United Kingdom

- Gramene at Cornell University, Ithaca, United States

- Institute for Genome Sciences, University of Maryland, Baltimore, United States

- The J. Craig Venter Institute, Rockville, United States

- Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, United States

- Muscle TRAIT, University of Padua, Padua, Italy

- Plant-Association Microbe Gene Ontology, Virginia Polytechnic Institute and State University, Blacksburg, United States

- Rat Genome Database at the Medical College of Wisconsin, Milwaukee, United States

- Reactome at Cold Spring Harbor Laboratory, United States

- Saccharomyces Genome Database, Stanford University, Stanford, United States

- WormBase at California Institute of Technology, Pasadena, United States

- The Zebrafish Information Network at the University of Oregon, Eugene, United States

IMEx Consortium

- Centro Nacional de Biotecnologia, Madrid, Spain

- DIP at the University of California, Los Angeles, United States

- MINT at University Tor Vergata, Rome, Italy

- MIPS at the National Research Centre for Environment and Health, Munich, Germany

- Neuroproteomics platform of National Neurosciences Facility, Melbourne, Australia

- Shanghai Institutes for Biological Sciences, Shanghai, China

InterPro

- CATH-Gene3D at University College London, London, United Kingdom

- HAMAP at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

- InterPro at EMBL-EBI, Hinxton, United Kingdom

- PANTHER at University of Southern California, Los Angeles, United States

- Pfam at the Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

- PIRSF at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, United States

- PRINTS at the University of Manchester, United Kingdom

- ProDom at INRA and CNRS, Toulouse, France

- PROSITE at the Swiss Institute of Bioinformatics, Geneva, Switzerland

- SCOP at the Laboratory of Molecular Biology, University of Cambridge, United Kingdom

- SMART at EMBL, Heidelberg, Germany

- SUPERFAMILY at the University of Bristol, United Kingdom

- TIGRFAMs at The Institute of Genome Research, Rockville, United States

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

- *Faculty of Life Sciences, The University of Manchester, United Kingdom*
- *Ghent University, Ghent, Belgium*
- *The Yonsei Proteome Research Center, Yonsei University, Seoul, Korea*

RNAcentral

- *dictyBase at Northwestern University, Chicago, United States*
- *the Greengenes Consortium, a collaboration with University of Colorado, Boulder, United States; University of Queensland, Australia; and Second Genome Inc., San Francisco, United States*
- *gtRNAdb at University of California Santa Cruz, Santa Cruz, United States*
- *LNCipedia at Ghent University, Ghent, Belgium*
- *lncRNAdb at Garvan Institute of Medical Research, Darlinghurst, Australia*
- *miRBase at University of Manchester, United Kingdom*
- *Modomics at International Institute of Molecular and Cell Biology, Warsaw, Poland*
- *NONCODE at Institute of Biophysics at Chinese Academy of Sciences, Beijing, China*
- *Pombase at University College London and the University of Cambridge, UK*
- *Ribosomal Database Project at Michigan State University, East Lansing, United States*
- *RefSeq at National Center for Biotechnology Information, Bethesda, United States*
- *Saccharomyces Genome Database, Stanford University, Stanford, United States*
- *SILVA at Max Planck Institute for Marine Microbiology, Bremen, Germany*
- *snOPY at University of Miyazaki, Miyazaki, Japan*
- *SRPDB at University of Texas Health Science Center, San Antonio, United States*

- *The Arabidopsis Information Resource (TAIR) at Phoenix Bioinformatics Corporation, Redwood City, United States*
- *tmRNA Website at Sandia National Laboratories, Livermore, United States*
- *Vega at the Wellcome Trust Sanger Institute, Hinxton, United Kingdom*
- *Wormbase, a collaboration with the California Institute of Technology and Washington University, United States; Ontario Institute for Cancer Research, Canada; the Wellcome Trust Sanger Institute and Oxford University, United Kingdom*
- *RNAcentral databases at EMBL-EBI*

Reactome

- *New York University Medical Center, New York, United States*
- *Ontario Institute for Cancer Research, Toronto, Canada*
- *Reactome at Cold Spring Harbor Laboratory, United States*
- *Reactome at EMBL-EBI*

Rfam

- *Protein Data Bank in Europe (PDBe)*
- *European Nucleotide Archive (ENA)*

UniProt: The Universal Protein Resource

UniProt at EMBL-EBI is part of the UniProt Consortium. Other partners include:

- *UniProt at the Protein Information Resource, Georgetown University Medical Centre, Washington, DC, United States*
- *UniProt at the Protein Information Resource, University of Delaware, United States*
- *UniProt at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*

Publications

Publications in 2015

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 ORCIDs and affiliation data extracted from Web of Science, represents EMBL-EBI articles published online in 2015.

001. 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68-74. doi:10.1038/nature15393

002. Achim K, Pettit JB, Saraiva LR, et al. (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. 33:503-509. doi:10.1038/nbt.3209

003. Agirre X, Castellano G, Pascual M, et al. (2015) Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. Genome Res. 25:478-487. doi:10.1101/gr.180240.114

004. Aguiar B, Vieira J, Cunha AE, et al. (2015) Convergent evolution at the gametophytic self-incompatibility system in malus and prunus. PLoS ONE 10: e0126138. doi:10.1371/journal.pone.0126138

005. Ahnert SE, Marsh JA, Hernández H, et al. (2015) Principles of assembly reveal a periodic table of protein complexes. Science 350: aaa2245. doi:10.1126/science.aaa2245

006. Alvaro N, Conway M, Doan S, et al. (2015) Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. J. Biomed. Inform. 58:280-287. doi:10.1016/j.jbi.2015.11.004

007. Andersson L, Archibald AL, Bottema CD, et al. (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project Genome Biol. 16:57. doi:10.1186/s13059-015-0622-4

008. Anjum S, Morganella S, D'Angelo F, et al. (2015) VEGAWES: variational segmentation on whole exome sequencing for copy number detection. BMC Bioinformatics 16:315-315. doi:10.1186/s12859-015-0748-0

009. Arrowsmith CH, Audia JE, Austin C, et al. (2015) The promise and peril of chemical probes Nat. Chem. Biol. 11:536-541. doi:10.1038/nchembio.1867

010. Audain E, Ramos Y, Hermjakob H, et al. (2015) Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. Bioinformatics (in press); doi:10.1093/bioinformatics/btv674

011. Babbitt PC, Bagos PG, Bairoch A, et al. (2015) Creating a specialist protein resource network: a meeting report for the protein bioinformatics and community resources retreat. Database 2015:bav063. doi:10.1093/database/bav063

012. Bacci G, Ceccherini MT, Bani A, et al. (2015) Exploring the dynamics of bacterial community composition in soil: the pan-bacteriome approach. Antonie Van Leeuwenhoek 107:785-797. doi:10.1007/s10482-014-0372-4

013. Barrett JC, Dunham I, Birney E (2015) Using human genetics to make new medicines. Nat. Rev. Genet. 16:561-562. doi:10.1038/nrg3998

014. Bastian FB, Chibucos MC, Gaudet P, et al. (2015) The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database (Oxford) 2015. doi:10.1093/database/bav043

015. Baxevanis AD, Bateman A (2015) The importance of biological databases in biological discovery. Curr Protoc Bioinformatics 50:1.1.1-8. doi:10.1002/0471250953.bi0101s50

016. Beisken S, Conesa P, Haug K, et al. (2015) SpeckTackle: JavaScript charts for spectroscopy. J Cheminform 7:17. doi:10.1186/s13321-015-0065-7

017. Benjelloun B, Alberto FJ, Streeter I, et al. (2015) Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. Front Genet 6:107. doi:10.3389/fgene.2015.00107

018. Birney E, Soranzo N (2015) Human genomics: The end of the start for population sequencing. Nature 526:52-53. doi:10.1038/526052a

019. Blake JA, Christie KR, Dolan ME, et al. (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res. 43(D1):D1049-D1056. doi:10.1093/nar/gku1179

020. Blumenthal T, Davis P, Garrido-Lecca A (2015) Operon and non-operon gene clusters in the C. elegans genome. WormBook 2015:1-20. doi:10.1895/wormbook.1.175.1

021. Boeckmann B, Marcet-Houben M, Rees JA, et al. (2015) Quest for orthologs entails quest for tree of life: In search of the gene stream. Genome Biol. Evol. 7:1988-1999. doi:10.1093/gbe/evv121

022. Bolser D, Staines DM, Pritchard E, et al. (2015) Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomics data. Methods Mol. Biol. 1374:115-140. doi:10.1007/978-1-4939-3167-5_6

023. Boroviak T, Loos R, Lombard P, et al. (2015) Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. Dev. Cell 35:366-382. doi:10.1016/j.devcel.2015.10.011

024. Bosi E, Donati B, Galardini M, et al. (2015) MeDuSa: A multi-draft based scaffolder. Bioinformatics 31:2443-2451. doi:10.1093/bioinformatics/btv171

025. Brage S, Westgate K, Franks PW, et al. (2015) Estimation of free-living energy expenditure by heart rate and movement sensing: A doubly-labelled water study. PLoS One 10:e0137206. doi:10.1371/journal.pone.0137206

026. Brooksbank C, Johnson C (2015) Europe: Lifelong learning for all in biomedicine. Nature 524:415-415. doi:10.1038/524415c

027. Bruford EA, Lane L, Harrow J (2015) Devising a consensus framework for validation of novel human coding loci. J. Proteome Res. 14:4945-4948. doi:10.1021/acs.jproteome.5b00688

028. Bruford MW, Ginja C, Hoffmann I, et al. (2015) Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Front Genet 6:314. doi:10.3389/fgene.2015.00314

029. Budd A, Dinkel H, Corpas M, et al. (2015) Ten simple rules for organizing an unconference. PLoS Comput. Biol. 11: e1003905. doi:10.1371/journal.pcbi.1003905

030. Buettner F, Natarajan KN, Casale FP, et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. 33:155-160. doi:10.1038/nbt.3102

031. Cannone JJ, Sweeney BA, Petrov AI, et al. (2015) R3D-2-MSA: the RNA 3D structure-to-multiple sequence alignment server. Nucleic Acids Res. 43:W15-W23. doi:10.1093/nar/gkv543

032. Carén H, Stricker SH, Bulstrode H, et al. (2015) Glioblastoma stem cells respond to differentiation cues but fail to undergo commitment and terminal cell-cycle arrest. Stem Cell Reports 5:829-842. doi:10.1016/j.stemcr.2015.09.014

033. Carvalho PC, Padron G, Calvete JJ, et al. (2015) Computational proteomics: Integrating mass spectral data into a biological context. J. Proteomics 129:1-2. doi:10.1016/j.jprot.2015.10.013

034. Casale FP, Rakitsch B, Lippert C, et al. (2015) Efficient set tests for the genetic analysis of correlated traits. Nat. Methods 12:755-U93. doi:10.1038/NMETH.3439

035. Chaouiya C, Keating SM, Berenguier D, et al. (2015) The Systems Biology Markup Language (SBML) Level 3 Package: Qualitative Models, Version 1, Release 1. J Integr. Bioinform. 12:270. doi:10.2390/biecoll-jib-2015-270

036. Chiang Z, Vastermark A, Punta M, et al. (2015) The complexity, challenges and benefits of comparing two transporter classification systems in TCDB and Pfam. Brief. Bioinform. 16:865-872. doi:10.1093/bib/bbu053

037. Chiapparino A, Maeda K, Turei D, et al. (2015) The orchestra of lipid-transfer proteins at the crossroads between metabolism and signaling. Prog. Lipid Res. 61:30-39. doi:10.1016/j.plipres.2015.10.004

038. Church DM, Schneider VA, Steinberg KM, et al. (2015) Extending reference assembly models. Genome Biol. 16:13. doi:10.1186/s13059-015-0587-3

039. Clark MB, Mercer TR, Bussotti G, et al. (2015) Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. Nat. Methods 12:339-342. doi:10.1038/nmeth.3321

040. Collier N, Groza T, Smedley D, et al. (2015) PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. Database (Oxford) 2015: doi:10.1093/database/bav104

041. Collier N, Oellrich A, Groza T (2015) Concept selection for phenotypes and diseases using learn to rank J. Biomed. Semantics 6:24. doi:10.1186/s13326-015-0019-z

042. Cook CE, Bergman MT, Finn RD, et al. (2015) The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 44:D20-D26. doi:10.1093/nar/gkv1352

043. Cortes-Ciriano I, Murrell DS, van Westen GJ, et al. (2015) Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. J Cheminform 7:1. doi:10.1186/s13321-014-0049-z

044. Cortes-Ciriano I, Ul Ain Q, Subramanian V, et al. (2015) Polypharmacology modelling using proteochemometrics (PCM): Recent methodological developments, applications to target families, and future prospects. Med. Chem. Comm. 6:24-50. doi:10.1039/c4md00216d

045. Cortés-Ciriano I, van Westen GJ, Bouvier G, et al. (2015) Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics (in press); doi:10.1093/bioinformatics/btv529

046. Cunningham F, Moore B, Ruiz-Schultz N, et al. (2015) Improving the Sequence Ontology terminology for genomic variant annotation. J. Biomed. Semantics 6:32. doi:10.1186/s13326-015-0030-4

047. Dahlman I, Sinha I, Gao H, et al. (2015) The fat cell epigenetic signature in post-obese women is characterized by global hypomethylation and differential DNA methylation of adipogenesis genes. Int. J. Obes. (Lond) 39:910-919. doi:10.1038/ijo.2015.31

048. Davies M, Dedman N, Hersey A, et al. (2015) ADME SARfari: comparative genomics of drug metabolizing systems. Bioinformatics 31:1695-1697. doi:10.1093/bioinformatics/btv010

049. Davies M, Nowotka M, Papadatos G, et al. (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. 43:w612-w620. doi:10.1093/nar/gkv352

050. de Lange KM, Barrett JC (2015) Understanding inflammatory bowel disease via immunogenetics. J. Autoimmun. 64:91-100. doi:10.1016/j.jaut.2015.07.013

051. de Villavicencio-Díaz TN, Gómez YR, Argüelles BO, et al. (2015) Comparative proteomics analysis of the antitumor effect of CIGB-552 peptide in HT-29 colon adenocarcinoma cells. J. Proteomics 126:163-171. doi:10.1016/j.jprot.2015.05.024

052. Deans AR, Lewis SE, Huala E, et al. (2015) Finding our way through phenotypes. PLoS Biol. 13:e1002033. doi:10.1371/journal.pbio.1002033

053. Denas O, Sandstrom R, Cheng Y, et al. (2015) Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution BMC Genomics 16:87. doi:10.1186/s12864-015-1245-6

054. Deutsch EW, Albar JP, Binz PA, et al. (2015) Development of data representation standards by the human proteome organization proteomics standards initiative. J. Am. Med. Inform. Assoc. 22:495-506. doi:10.1093/jamia/ocv001

055. Diaz-Muñoz MD, Bell SE, Fairfax K, et al. (2015) The RNA-binding protein HuR is essential for the B cell antibody response. Nat. Immunol. 16:415-425. doi:10.1038/ni.3115

056. Ding Z, Ni Y, Timmer SW, et al. (2015) Correction: Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. PLoS Genet. 11: e1005177. doi:10.1371/journal.pgen.1005177

057. Dong H, Nebert DW, Bruford EA, et al. (2015) Update of the human and mouse Fanconi anemia genes. Hum. Genomics 9:32. doi:10.1186/s40246-015-0054-y

058. Dritsou V, Topalis P, Windbichler N, et al. (2015) A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus. Pathog. Glob. Health 109:207-220. doi:10.1179/2047773215Y.0000000031

059. Duarte AM, Psomopoulos FE, Blanchet C, et al. (2015) Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis. Front Genet 6:197. doi:10.3389/fgene.2015.00197

060. Dubin MJ, Zhang P, Meng D, et al. (2015) DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife 4: e05255 doi:10.7554/elife.05255

061. Dutta-Roy R, Rosenmund C, Edelstein SJ, et al. (2015) Ligand-dependent opening of the multiple AMPA receptor conductance States: a concerted model. PLoS ONE 10: e0116616. doi:10.1371/journal.pone.0116616

062. Dyke SOM, Cheung WA, Joly Y, et al. (2015) Epigenome data release: A participant-centered approach to privacy protection. Genome Biol. 16:142. doi:10.1186/s13059-015-0723-0

063. Eduati F, Mangravite LM, Wang T, et al. (2015) Prediction of human population responses to toxic compounds by a collaborative competition. Nat. Biotechnol. 33:933-U172. doi:10.1038/nbt.3299

064. Emwas AH, Luchinat C, Turano P, et al. (2015) Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. Metabolomics 11:872-894. doi:10.1007/s11306-014-0746-7

065. Eory L, Gilbert MTP, Li C, et al. (2015) Avianbase: a community resource for bird genomics Genome Biol. 16:21. doi:10.1186/s13059-015-0588-2

066. Fabregat A, Sidiropoulos K, Garapati P, et al. (2015) The Reactome Pathway Knowledgebase. Nucleic Acids Res. 44:D481-D487. doi:10.1093/nar/gkv1351

067. Faddeeva A, Studer RA, Kraaijeveld K, et al. (2015) Collembolan transcriptomes highlight molecular evolution of hexapods and provide clues on the adaptation to terrestrial life. PLoS ONE 10: e0130600. doi:10.1371/journal.pone.0130600

068. Farnell EJ, Tyagi N, Ryan S, et al. (2015) Known allergen structures predict Schistosoma mansoni IgE-binding antigens in human infection. Front. Immunol. 6:26. doi:10.3389/fimmu.2015.00026

069. Ferreira P, Fonseca NA, Dutra I, et al. (2015) Predicting malignancy from mammography findings and image-guided core biopsies Int. J. Data Min. Bioinform. 11:257-276. doi:10.1504/IJDMB.2015.067319

070. Fiehn O, Putri SP, Saito K, et al. (2015) Metabolomics continues to expand: highlights from the 2015 metabolomics conference. Metabolomics 11:1036-1040. doi:10.1007/s11306-015-0846-z

071. Finn RD, Clements J, Arndt W, et al. (2015) HMMER web server: 2015 update. Nucleic Acids Res. 43:w30-w38. doi:10.1093/nar/gkv397

072. Finn RD, Coggill P, Eberhardt RY, et al. (2015) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279-D285. doi:10.1093/nar/gkv1344

073. Forster SC, Browne HP, Kumar N, et al. (2015) HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res. 44:D604-D609. doi:10.1093/nar/gkv1216

074. Foulger RE, Osumi-Sutherland D, McIntosh BK, et al. (2015) Representing virus-host interactions and other multi-organism processes in the Gene Ontology. BMC Microbiol. 15:146. doi:10.1186/s12866-015-0481-x

075. Frankish A, Uszczynska B, Ritchie GRS, et al. (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics 16: doi:10.1186/1471-2164-16-S8-S2

Publications in 2015

076. Fu G, Batchelor C, Dumontier M, et al. (2015) PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. J. Cheminformatics 7:34. doi:10.1186/s13321-015-0084-4

077. Furnham N, Dawson NL, Rahman SA, et al. (2015) Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. J. Mol. Biol. (in press); doi:10.1016/j.jmb.2015.11.010

078. Fuster-Matanzo A, Gessler F, Leonardi T, et al. (2015) Acellular approaches for regenerative medicine: on the verge of clinical trials with extracellular membrane vesicles? Stem Cell Res. Ther. 6:227. doi:10.1186/s13287-015-0232-9

079. Galardini M, Mengoni A, Bazzicalupo M (2015) Mapping contigs using CONTIGuator. Methods Mol. Biol. 1231:163-176. doi:10.1007/978-1-4939-1720-4_11

080. Galardini M, Mengoni A, Mocali S (2015) From pangenome to panphenome and back. Methods Mol. Biol. 1231:257-270. doi:10.1007/978-1-4939-1720-4_16

081. Gardner PP, Fasold M, Burge SW, et al. (2015) Conservation and losses of non-coding RNAs in avian genomes. PLoS ONE 10:e0121797. doi:10.1371/journal.pone.0121797

082. Gatto L, Hansen KD, Hoopmann MR, et al. (2015) Testing and validation of computational methods for Mass Spectrometry. J. Proteome Res. (in press); doi:10.1021/acs.jproteome.5b00852

083. Gaulton A, Kale N, van Westen GJ, et al. (2015) A large-scale crop protection bioassay data set. Sci. Data 2:150032-150032. doi:10.1038/sdata.2015.32

084. Gaulton KJ, Ferreira T, Lee Y, et al. (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat. Genet. 47:1415-1425. doi:10.1038/ng.3437

085. Geihs M, Yan Y, Walter K, et al. (2015) An interactive genome browser of association results from the UK10K cohorts project. Bioinformatics 31:4029-4031. doi:10.1093/bioinformatics/btv491

086. Gibson R, Alako B, Amid C, et al. (2015) Biocuration of functional annotation at the European Nucleotide Archive. Nucleic Acids Res. 44:D58-D66. doi:10.1093/nar/gkv1311

087. Golanowska M, Galardini M, Bazzicalupo M, et al. (2015) Draft genome sequence of a highly virulent strain of the plant pathogen Dickeya solani, IFB0099. Genome Announc. 3:e00109-15-e00109-15. doi:10.1128/genomeA.00109-15

088. Griss J (2015) Spectral library searching in proteomics. Proteomics 16:729-740. doi:10.1002/pmic.201500296

089. Griss J, Perez-Riverol Y, Hermjakob H, et al. (2015) Identifying novel biomarkers through data mining - a realistic scenario? Proteomics Clin Appl 9:437-443. doi:10.1002/prca.201400107

090. Groza T, Köhler S, Moldenhauer D, et al. (2015) The Human Phenotype Ontology: Semantic unification of common and rare disease. Am. J. Hum. Genet. 97:111-124. doi:10.1016/j.ajhg.2015.05.020

091. Gutmanas A, Adams PD, Bardiaux B, et al. (2015) NMR Exchange Format: A unified and open standard for representation of NMR restraint data. Nat. Struct. Mol. Biol. 22:433-434. doi:10.1038/nsmb.3041

092. Hagmann J, Becker C, Müller J, et al. (2015) Century-scale Methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS Genet. 11: e1004920. doi:10.1371/journal.pgen.1004920

093. Hamann J, Aust G, Arac D, et al. (2015) International Union of Basic and Clinical Pharmacology. XCIV. Adhesion G protein-coupled receptors. Pharmacol. Rev. 67:338-367. doi:10.1124/pr.114.009647

094. Hastings J, Jeliazkova N, Owen G, et al. (2015) eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. J. Biomed. Semantics 6:10. doi:10.1186/s13326-015-0005-5

095. Hastings J, Owen G, Dekker A, et al. (2015) ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res (in press); doi:10.1093/nar/gkv1031

096. Heinzel A, Muhlberger I, Stelzer G, et al. (2015) Molecular disease presentation in diabetic nephropathy Nephrol. Dial. Transplant. 30:17-25. doi:10.1093/ndt/gfv267

097. Henriques D, Rocha M, Saez-Rodriguez J, et al. (2015) Reverse engineering of logic-based differential equation models using a mixed-integer dynamic optimisation approach. Bioinformatics 31:2999-3007. doi:10.1093/bioinformatics/btv314

098. Herrero-Zazo M, Segura-Bedmar I, Hastings J, et al. (2015) DINTO: Using OWL ontologies and SWRL rules to Infer drug-drug interactions and their mechanisms J. Chem Inf. Model. 55:1698-1707. doi:10.1021/acs.jcim.5b00119

099. Hersey A, Chambers J, Bellis L, et al. (2015) Chemical databases: curation or integration by user-defined equivalence? Drug Discov. Today Technol. 14:17-24. doi:10.1016/j.ddtec.2015.01.005

100. Hodson N, Invergo B, Rayner JC, et al. (2015) Palmitoylation and palmitoyl-transferases in Plasmodium parasites. Biochem. Soc. Trans. 43:240-245. doi:10.1042/BST20140289

101. Holliday GL, Bairoch A, Bagos PG, et al. (2015) Key challenges for the creation and maintenance of specialist protein resources. Proteins 83:1005-1013. doi:10.1002/prot.24803

102. Holmes RK, Tuck AC, Zhu C, et al. (2015) Loss of the yeast SR protein Npl3 alters gene expression due to transcription readthrough. PLoS Genet. 11: e1005735. doi:10.1371/journal.pgen.1005735

103. Horikoshi M, M gi R, van de Bunt M, et al. (2015) Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. PLoS Genet. 11: e1005230. doi:10.1371/journal.pgen.1005230

104. Horvatovich P, Lundberg EK, Chen YJ, et al. (2015) Quest for missing proteins: update 2015 on chromosome-centric Human Proteome Project. J. Proteome Res. 14:3415-3431. doi:10.1021/pr5013009

105. Hostas J, Jakubec D, Laskowski RA, et al. (2015) Representative amino acid side-chain interactions in protein-DNA complexes: A comparison of highly accurate correlated ab initio quantum mechanical calculations and efficient approaches for applications to large systems J. Chem. Theory Comput. 11:4086-4092. doi:10.1021/acs.jctc.5b00398

106. Howe KL, Bolt BJ, Cain S, et al. (2015) WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 44:D774-D780. doi:10.1093/nar/gkv1217

107. Huang YH, Xu BS, Zhou XY, et al. (2015) Systematic characterization and prediction of post-translational modification cross-talk. Mol. Cell. Proteomics 14:761-770. doi:10.1074/mcp.M114.037994

108. Hubley R, Finn RD, Clements J, et al. (2015) The Dfam database of repetitive DNA families. Nucleic Acids Res. 44:D81-D89. doi:10.1093/nar/gkv1272

109. Hucka M, Bergmann FT, Dräger A, et al. (2015) Systems Biology Markup Language (SBML) Level 2 Version 5: Structures and Facilities for Model Definitions. J Integr Bioinform 12:271. doi:10.2390/biecoll-jib-2015-271

110. Hucka M, Nickerson DP, Bader GD, et al. (2015) Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. Front Bioeng Biotechnol 3:19. doi:10.3389/fbioe.2015.00019

111. Ing-Simmons E, Seitan V, Faure A, et al. (2015) Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res. 25:504-513. doi:10.1101/gr.184986.114

112. Iorio F, Shrestha RL, Levin N, et al. (2015) A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions. PLoS One 10(10):e0139446-e0139446. doi:10.1371/journal.pone.0139446

113. Ison J, Rapacki K, Ménager H, et al. (2015) Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 44:D38-D47. doi:10.1093/nar/gkv1116

114. Ivanov DK, Escott-Price V, Ziehm M, et al. (2015) Longevity GWAS using the Drosophila genetic reference panel. J. Gerontol. A Biol. Sci. Med. Sci. 70:1470-1478. doi:10.1093/gerona/glv047

115. Jakubec D, Hostas J, Laskowski RA, et al. (2015) Large-scale quantitative assessment of binding preferences in protein-nucleic acid complexes J. Chem. Theory Comput. 11:1939-1948. doi:10.1021/ct501168n

116. Jeliazkova N, Chomenidis C, Doganis P, et al. (2015) The eNanoMapper database for nanomaterial safety information. Beilstein J. Nanotechnol. 6:1609-1634. doi:10.3762/bjnano.6.165

117. Johnson JR, Santos SD, Johnson T, et al. (2015) Prediction of functionally important phospho-regulatory events in Xenopus laevis oocytes. PLoS Comput. Biol. 11: e1004362. doi:10.1371/journal.pcbi.1004362

118. Jupe S, Fabregat A, Hermjakob H (2015) Expression data analysis with Reactome. Curr Protoc Bioinformatics 49:8.20.1-9. doi:10.1002/0471250953.bi0820s49

119. Juty N, Ali R, Glont M, et al. (2015) BioModels: Content, features, functionality, and use. CPT Pharmacometrics Syst. Pharmacol. 4(2). doi:10.1002/psp4.3

120. Kafkas Ş, Kim JH, Pi X, et al. (2015) Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. J. Biomed. Semantics 6:1-1. doi:10.1186/2041-1480-6-1

121. Kafkas Ş, Pi X, Marinos N, et al. (2015) Section level search functionality in Europe PMC. J. Biomed. Semantics 6:7. doi:10.1186/s13326-015-0003-7

122. Karp NA, Meehan TF, Morgan H, et al. (2015) Applying the ARRIVE Guidelines to an in vivo database. PLoS Biol. 13: e1002151. doi:10.1371/journal.pbio.1002151

123. Kersey PJ, Allen JE, Armean I, et al. (2015) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. 44:D574-D580. doi:10.1093/nar/gkv1209

124. Kim JK, Cho Y, Lee M, et al. (2015) BetaCavityWeb: a webserver for molecular voids and channels. Nucleic Acids Res. 43:w413-w418. doi:10.1093/nar/gkv360

125. Kim JK, Kolodziejczyk AA, Illicic T, et al. (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. 6: doi:10.1038/ncomms9687

126. Kirsanova C, Brazma A, Rustici G, et al. (2015) Cellular Phenotype Database: a repository for systems microscopy data. Bioinformatics 31:2736-2740. doi:10.1093/bioinformatics/btv199

127. Kiselev VY, Juvin V, Malek M, et al. (2015) Perturbations of PIP3 signalling trigger a global remodelling of mRNA landscape and reveal a transcriptional feedback loop. Nucleic Acids Res 43:9663-9679. doi:10.1093/nar/gkv1015

128. Kolodziejczyk AA, Kim JK, Svensson V, et al. (2015) The technology and biology of single-cell RNA sequencing. Mol. Cell 58:610-620. doi:10.1016/j.molcel.2015.04.005

129. Kolodziejczyk AA, Kim JK, Tsang JC, et al. (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 17:471-485. doi:10.1016/j.stem.2015.09.011

130. Kopf A, Bicak M, Kottmann R, et al. (2015) The Ocean Sampling Day consortium. Gigascience 4:27. doi:10.1186/s13742-015-0066-5

131. Krajewski P, Chen D, Ćwiek H, et al. (2015) Towards recommendations for metadata and data handling in plant phenotyping. J. Exp. Bot. 66(18):5417-5427. doi:10.1093/jxb/erv271

132. Kretzmer H, Bernhart SH, Wang W, et al. (2015) DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nat Genet 47:1316-1325. doi:10.1038/ng.3413

133. Krupska I, Bruford EA, Chaqour B (2015) Eyeing the Cyr61/CTGF/NOV (CCN) group of genes in development and diseases: highlights of their structural likenesses and functional dissimilarities. Hum Genomics 9: 24. doi:10.1186/s40246-015-0046-y

134. Kubagawa H, Carroll MC, Jacob CO, et al. (2015) Nomenclature of Toso, Fas apoptosis inhibitory molecule 3, and IgM FcR. J. Immunol. 194:4055-4057. doi:10.4049/jimmunol.1500222

135. Kulis M, Merkel A, Heath S, et al. (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat. Genet. 47(7):746-756. doi:10.1038/ng.3291

136. Kurbatova N, Mason JC, Morgan H, et al. (2015) PhenStat: A tool kit for standardized analysis of high throughput phenotypic data. PLoS ONE 10: e0131274. doi:10.1371/journal.pone.0131274

137. Lai M, Brun D, Edelstein SJ, et al. (2015) Modulation of calmodulin lobes by different targets: an allosteric model with hemiconcerted conformational transitions. PLoS Comput. Biol. 11: e1004063. doi:10.1371/journal.pcbi.1004063

138. Lappalainen I, Almeida-King J, Kumanduri V, et al. (2015) The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. 47:692-695. doi:10.1038/ng.3312

139. Laskowski RA, Thornton JM (2015) Proteins: interaction at a distance. IUCrJ 2(Pt 6):609-610. doi:10.1107/S2052252515020217

140. Lawson CL, Patwardhan A, Baker ML, et al. (2015) EMDataBank unified data resource for 3DEM. Nucleic Acids Res. 44:D396-D403. doi:10.1093/nar/gkv1126

141. Le V, Khanh Nhu NT, Cerdeno-Tarraga A, et al. (2015) Genetic characterization of three qnrS1-harbouring multidrug-resistance plasmids and qnrS1-containing transposons circulating in Ho Chi Minh City, Vietnam. J. Med. Microbiol. 64:869-878. doi:10.1099/jmm.0.000100

142. Leha A, Moens N, Meleckyte R, et al. (2015) A high-content platform to characterise human induced pluripotent stem cell lines. Methods (in press); doi:10.1016/j.ymeth.2015.11.012

143. Lehmann KV, Kahles A, Kandoth C, et al. (2015) Integrative genome-wide analysis of the determinants of RNA splicing in kidney renal clear cell carcinoma. Pac Symp Biocomput :44-55. Paper presented at the Pacific Symposium, Kohala Coast, Hawaii, USA, 4–8 January 2015. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, Klein TE (eds.), Proceedings of the Pacific Symposium. pp 44-55. doi:10.1142/9789814644730_0006

144. Leigh-Brown S, Goncalves A, Thybert D, et al. (2015) Regulatory divergence of transcript isoforms in a mammalian model system. PLoS One 10(9): doi:10.1371/journal.pone.0137367

145. Lener T, Gimona M, Aigner L, et al. (2015) Applying extracellular vesicles based therapeutics in clinical trials - an ISEV position paper. J. Extracell. Vesicles 4: 30087. doi:10.3402/jev.v4.30087

146. Li H, Tong P, Gallegos J, et al. (2015) PAND: A distribution to identify functional linkage from networks with preferential attachment property. PLoS ONE 10: e0127968. doi:10.1371/journal.pone.0127968

147. Li HJ, Leung KS, Wong MH, et al. (2015) Improving AutoDock Vina using Random Forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. Mol. Inf. 34:115-126. doi:10.1002/minf.201400132

148. Li P, Seneviratne CJ, Alpi E, et al. (2015) Delicate metabolic control and coordinated stress response critically determine antifungal tolerance of Candida albicans biofilm persisters. Antimicrob. Agents Chemother. 59:6101-6112. doi:10.1128/aac.00543-15

149. Li WZ, Cowley A, Uludag M, et al. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 43:W580-W584. doi:10.1093/nar/gkv279

150. Lizio M, Harshbarger J, Shimoji H, et al. (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 16:22. doi:10.1186/s13059-014-0560-6

151. Lloyd KCK, Meehan T, Beaudet A, et al. (2015) Precision medicine: Look to the mice. Science 349(6246):390

152. Lowe R, Slodkowicz G, Goldman N, et al. (2015) The human blood DNA methylome displays a highly distinctive profile compared with other somatic tissues. Epigenetics 10:274-281. doi:10.1080/15592294.2014.1003744

153. MacNamara A, Stein F, Feng S, et al. (2015) A single-cell model of PIP3 dynamics using chemical dimerization. Bioorg. Med. Chem. 23:2868-2876. doi:10.1016/j.bmc.2015.04.074

Publications in 2015

154. Mak L, Marcus D, Howlett A, et al. (2015) Metrabase: A cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. J. Cheminform 7:31. doi:10.1186/s13321-015-0083-5

155. Maree S, Maree FF, Putterill JF, et al. (2015) Synthesis of empty African horse sickness virus particles. Virus Res. 213:184-194. doi:10.1016/j.virusres.2015.12.006

156. Marsh JA, Rees HA, Ahnert SE, et al. (2015) Structural and evolutionary versatility in protein complexes with uneven stoichiometry. Nat. Commun. 6:6394; doi:10.1038/ncomms7394

157. Marsh JA, Teichmann SA (2015) Structure, dynamics, assembly, and evolution of protein complexes. Annu. Rev. Biochem. 84:551-575. doi:10.1146/annurev-biochem-060614-034142

158. Martin H, Shales M, Fernandez-Piñar P, et al. (2015) Differential genetic interactions of yeast stress response MAPK pathways. Mol. Syst. Biol. 11:800. doi:10.15252/msb.20145606

159. Martínez Cuesta S, Rahman SA, Furnham N, et al. (2015) The classification and evolution of enzyme function. Biophys. J. 109:1082-1086. doi:10.1016/j.bpj.2015.04.020

160. Martínez H, Tárraga J, Medina I, et al. (2015) Concurrent and accurate short read mapping on multicore processors. IEEE/ACM Trans. Comput. Biol. Bioinform. 12:995-1007. doi:10.1109/tcbb.2015.2392077

161. Mattila J, Havula E, Suominen E, et al. (2015) Mondo-Mlx mediates organismal sugar sensing through the gli-similar transcription factor sugarbabe. Cell Rep 13:350-364. doi:10.1016/j.celrep.2015.08.081

162. McEntyre J, Sarkans U, Brazma A (2015) The BioStudies database. Mol. Syst. Biol. 11:847. doi:10.15252/msb.20156658

163. Medina-Rivera A, Defrance M, Sand O, et al. (2015) RSAT 2015: Regulatory sequence analysis tools. Nucleic Acids Res. 43:w50-w56. doi:10.1093/nar/gkv362

164. Melas IN, Sakellaropoulos T, Iorio F, et al. (2015) Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. Integr. Biol. 7:904-920. doi:10.1039/c4ib00294f

165. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, et al. (2015) Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc. Natl. Acad. Sci. U.S.A. 112:14936-14941. doi:10.1073/pnas.1506226112

166. Mitchell A, Bucchini F, Cochrane G, et al. (2015) EBI Metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. 44:D595-D603. doi:10.1093/nar/gkv1195

167. Moreno P, Beisken S, Harsha B, et al. (2015) BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. BMC Bioinformatics 16: doi:10.1186/s12859-015-0486-3

168. Moretto M, Sonego P, Dierckxsens N, et al. (2015) COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res. 44:D620-D623. doi:10.1093/nar/gkv1251

169. Mugumbate G, Abrahams KA, Cox JA, et al. (2015) Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. PLoS ONE 10:e0121492. doi:10.1371/journal.pone.0121492

170. Mugumbate G, Overington JP (2015) The relationship between target-class and the physicochemical properties of antibacterial drugs. Bioorg. Med. Chem. 23:5218-5224. doi:10.1016/j.bmc.2015.04.063

171. Murn J, Zarnack K, Yang YJ, et al. (2015) Control of a neuronal morphology program by an RNA-binding zinc finger protein, Unkempt. Genes Dev. 29:501-512. doi:10.1101/gad.258483.115

172. Murrell DS, Cortes-Ciriano I, van Westen GJP, et al. (2015) Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules J. Cheminform. 7:45. doi:10.1186/s13321-015-0086-2

173. Mussa HY, Marcus D, Mitchell JB, et al. (2015) Verifying the fully "Laplacianised" posterior Naive Bayesian approach and more. J. Cheminform. 7:27. doi:10.1186/s13321-015-0075-5

174. Naldi A, Monteiro PT, Mussel C, et al. (2015) Cooperative development of logical modelling standards and tools with CoLoMoTo. Bioinformatics 31:1154-1159. doi:10.1093/bioinformatics/btv013

175. Natarajan P, Punta M, Kumar A, et al. (2015) Structure and sequence analyses of Bacteroides proteins BVU_4064 and BF1687 reveal presence of two novel predominantly-beta domains, predicted to be involved in lipid and cell surface interactions. BMC Bioinformatics 16: doi:10.1186/s12859-014-0434-7

176. Neafsey DE, Waterhouse RM, Abai MR, et al. (2015) Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science 347: 1258522. doi:10.1126/science.1258522

177. Nguyen N, Hickey G, Zerbino DR, et al. (2015) Building a pan-genome reference for a population. J. Comput. Biol. 22:387-401. doi:10.1089/cmb.2014.0146

178. Nik-Zainal S, Kucab JE, Morganella S, et al. (2015) The genome as a record of environmental exposure. Mutagenesis 30:763-770. doi:10.1093/mutage/gev073

179. Nsamba P, de Beer TA, Chitray M, et al. (2015) Determination of common genetic variants within the non-structural proteins of foot-and-mouth disease viruses isolated in sub-Saharan Africa. Vet. Microbiol. 177:106-122. doi:10.1016/j.vetmic.2015.03.007

180. Núñez de Villavicencio-Díaz T, Ramos Gómez Y, Oliva Argüelles B, et al. (2015) Data for comparative proteomics analysis of the antitumor effect of CIGB-552 peptide in HT-29 colon adenocarcinoma cells. Data Brief. 4:468-473. doi:10.1016/j.dib.2015.06.024

181. Ochoa D, Beltrao P (2015) Kinase-two-hybrid: towards the conditional interactome. Mol. Syst. Biol. 11:798. doi:10.15252/msb.20156107

182. Oellrich A, Collier N, Smedley D, et al. (2015) Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. PLoS ONE 10:e0116040. doi:10.1371/journal.pone.0116040

183. Oprea TI, Overington JP (2015) Computational and practical aspects of drug repositioning. Assay Drug Dev. Technol. 13:299-306. doi:10.1089/adt.2015.29011.tiodrrr

184. Orchard S, Hermjakob H (2015) Shared resources, shared costs-leveraging biocuration resources. Database (Oxford) 2015. doi:10.1093/database/bav009

185. Papadatos G, Davies M, Dedman N, et al. (2015) SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. 44:D1220-D1228. doi:10.1093/nar/gkv1253

186. Papadatos G, Gaulton A, Hersey A, et al. (2015) Activity, assay and target data curation and quality in the ChEMBL database. J. Comput. Aided Mol. Des. 29:885-896. doi:10.1007/s10822-015-9860-5

187. Partridge L, Thornton J, Bates G (2015) The new science of ageing. Philos Trans R Soc. Lond. B Biol. Sci. 370:1676. doi:10.1098/rstb.2015.0249

188. Patel S, Roncaglia P, Lovering RC (2015) Using Gene Ontology to describe the role of the neurexin-neuroligin-SHANK complex in human, mouse and rat and its relevance to autism. BMC Bioinformatics 16: doi:10.1186/s12859-015-0622-0

189. Pavlopoulos GA, Malliarakis D, Papanikolaou N, et al. (2015) Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. Gigascience 4:38. doi:10.1186/s13742-015-0077-2

190. Pedro H, Maheswari U, Urban M, et al. (2015) PhytoPath: an integrative resource for plant pathogen genomics. Nucleic Acids Res 44(D1):D688-D693. doi:10.1093/nar/gkv1052

191. Perez-Riverol Y, Uszkoreit J, Sanchez A, et al. (2015) ms-data-core-api: An open-source, metadata-oriented library for computational proteomics. Bioinformatics 31:2903-2905. doi:10.1093/bioinformatics/btv250
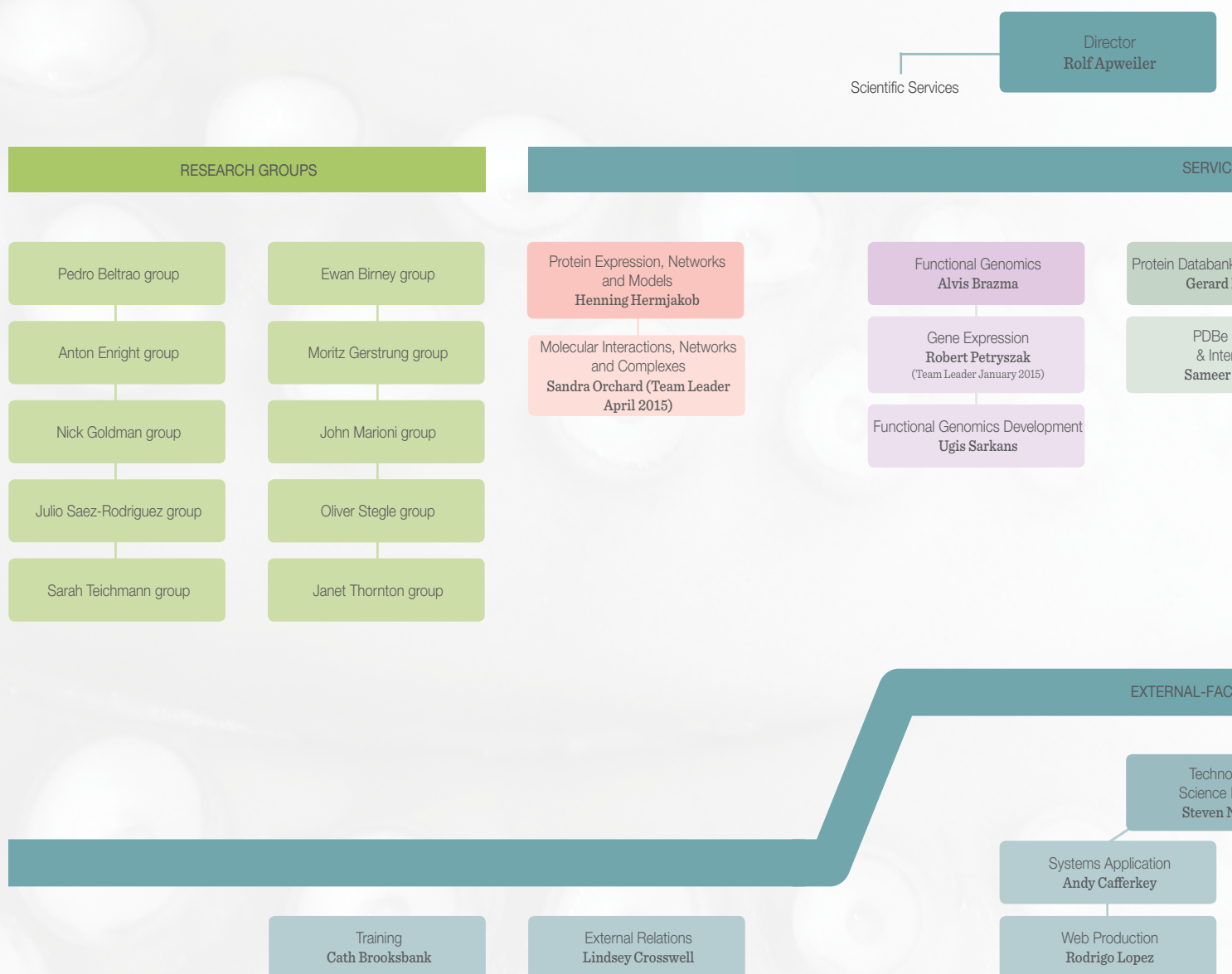
192. Perez-Riverol Y, Xu QW, Wang R, et al. (2015) PRIDE Inspector Toolsuite: Moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. Mol. Cell. Proteomics 15:305-317. doi:10.1074/mcp.O115.050229

193. Perna D, Karreth FA, Rust AG, et al. (2015) BRAF inhibitor resistance mediated by the AKT pathway in an oncogenic BRAF mouse melanoma model. Proc. Natl. Acad. Sci. U.S.A. 112:E536-45. doi:10.1073/pnas.1418163112

194. Philippakis AA, Azzariti DR, Beltran S, et al. (2015) The Matchmaker Exchange: A platform for rare disease gene discovery. Hum Mutat 36:915-921. doi: 10.1002/humu.22858

195. PLoS Medicine Editors, Beck A, Birney E, et al. (2015) Progress in medicine: experts take stock. PLoS Med. 12: e1001933. doi:10.1371/journal.pmed.1001933

196. Porras P, Duesbury M, Fabregat A, et al. (2015) A visual review of the interactome of LRRK2: Using deep-curated molecular interactions data to represent biology. Proteomics 15:1390-1404. doi:10.1002/pmic.201400390

197. Porta-Pardo E, Garcia-Alonso L, Hrabe T, et al. (2015) a pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Comput. Biol. 11:e1004518. doi:10.1371/journal.pcbi.1004518

198. Prakash A, Bateman A (2015) Domain atrophy creates rare cases of functional partial protein domains. Genome Biol. 16:88. doi:10.1186/s13059-015-0655-8

199. Proserpio V, Lönnberg T (2015) Single-cell technologies are revolutionizing the approach to rare cells. Immunol. Cell Biol. (in press); doi:10.1038/icb.2015.106

200. Proserpio V, Mahata B (2015) Single-cell technologies to study the immune system. Immunology 147:133-140. doi:10.1111/imm.12553

201. Pundir S, Magrane M, Martin MJ, et al. (2015) Searching and navigating UniProt databases. Curr Protoc Bioinformatics 50:1.27.1-1.27.10. doi:10.1002/0471250953.bi0127s50

202. Punta M, Simon I, Dosztányi Z (2015) Prediction and analysis of intrinsically disordered proteins. Methods Mol. Biol. 1261:35-59. doi:10.1007/978-1-4939-2230-7_3

203. Rahman SA, Singh Y, Kohli S, et al. (2015) Reply to '"Mycobacterium indicus pranii" is a strain of Mycobacterium intracellulare': "M. indicus pranii" is a distinct strain, not derived from M. intracellulare, and is an organism at an evolutionary transition point between a fast grower and slow grower. MBio 6: doi:10.1128/mbio.00352-15

204. Raposo AA, Vasconcelos FF, Drechsel D, et al. (2015) Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. Cell Rep (in press); doi:10.1016/j.celrep.2015.02.025

205. Ravenhall M, Škunca N, Lassalle F, et al. (2015) Inferring horizontal gene transfer. PLoS Comput. Biol. 11: e1004095. doi:10.1371/journal.pcbi.1004095

206. Rawlings ND (2015) Bacterial calpains and the evolution of the calpain (C2) family of peptidases. Biol. Direct 10:66. doi:10.1186/s13062-015-0095-0

207. Rawlings ND (2015) Peptidase specificity from the substrate cleavage collection in the MEROPS database and a tool to measure cleavage site conservation. Biochimie (in press); doi:10.1016/j.biochi.2015.10.003

208. Rawlings ND, Barrett AJ, Finn R (2015) Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 44:D343-D350. doi:10.1093/nar/gkv1118

209. Rebollo-Lopez MJ, Lelièvre J, Alvarez-Gomez D, et al. (2015) Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery. PLoS One 10: e0142293. doi:10.1371/journal.pone.0142293

210. Reisinger F, Del-Toro N, Ternent T, et al. (2015) Introducing the PRIDE Archive RESTful web services. Nucleic Acids Res. 43:W599-W604. doi:10.1093/nar/gkv382

211. Ring N, Meehan TF, Blake A, et al. (2015) A mouse informatics platform for phenotypic and translational discovery. Mamm. Genome 26:413-421. doi:10.1007/s00335-015-9599-2

212. Roberts AM, Ware JS, Herman DS, et al. (2015) Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. Sci. Transl. Med. 7:270ra6. doi:10.1126/scitranslmed.3010134

213. Rocca-Serra P, Salek RM, Arita M, et al. (2015) Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics 12: doi:10.1007/s11306-015-0879-3

214. Rodriguez N, Thomas A, Watanabe L, et al. (2015) JSBML 1.0: providing a smorgasbord of options to encode systems biology models. Bioinformatics 31:3383-3386. doi:10.1093/bioinformatics/btv341

215. Rohwer N, Bindel F, Grimm C, et al. (2015) Annexin A1 sustains tumor metabolism and cellular proliferation upon stable loss of HIF1A. Oncotarget 7:6693-6710. doi:10.18632/oncotarget.6793

216. Rompp A, Wang R, Albar JP, et al. (2015) A public repository for mass spectrometry imaging data Anal. Bioanal. Chem. 407:2027-2033. doi:10.1007/s00216-014-8357-8

217. Ruiz Hernandez SE, Streeter I, de Leeuw NH (2015) The effect of water on the binding of glycosaminoglycan saccharides to hydroxyapatite surfaces: a molecular dynamics study. Phys. Chem. Chem. Phys. 17:22377-22388. doi:10.1039/c5cp02630j

218. Saez-Rodriguez J, MacNamara A, Cook S (2015) Modeling signaling networks to advance new cancer therapies. Annu. Rev. Biomed. Eng. 17:143-163. doi:10.1146/annurev-bioeng-071813-104927

219. Salek RM, Arita M, Dayalan S, et al. (2015) Embedding standards in metabolomics: the Metabolomics Society data standards task group. Metabolomics 11:782-783. doi:10.1007/s11306-015-0821-8

220. Salek RM, Neumann S, Schober D, et al. (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics 11:1587-1597. doi:10.1007/s11306-015-0810-y

221. Sali A, Berman HM, Schwede T, et al. (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. Structure 23:1156-1167. doi:10.1016/j.str.2015.05.013

222. Saraiva LR, Ahuja G, Ivandic I, et al. (2015) Molecular and neuronal homology between the olfactory systems of zebrafish and mouse. Sci. Rep. 5:11487. doi:10.1038/srep11487

223. Saraiva LR, Ibarra-Soria X, Khan M, et al. (2015) Hierarchical deconstruction of mouse olfactory sensory neurons: from whole mucosa to single-cell RNA-seq. Sci. Rep. 5:18178. doi:10.1038/srep18178

224. Sariyar M, Schluender I, Smee C, et al. (2015) Sharing and reuse of sensitive data and samples: supporting researchers in identifying ethical and legal requirements. Biopreserv. Biobank. 13:263-270. doi:10.1089/bio.2015.0014

225. Schmid M, Smith J, Burt DW, et al. (2015) Third report on chicken genes and chromosomes 2015. Cytogenet. Genome Res. 145:78-179. doi:10.1159/000430927

226. Schoenfelder S, Furlan-Magaril M, Mifsud B, et al. (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements Genome Res. 25:582-597. doi:10.1101/gr.185272.114

227. Schoenfelder S, Sugar R, Dimond A, et al. (2015) Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat. Genet. 47:1179-1186. doi:10.1038/ng.3393

228. Schwarz RF, Branicky R, Grundy LJ, et al. (2015) Changes in postural syntax characterize sensory modulation and natural variation of C. elegans locomotion. PLoS Comput. Biol. 11: e1004322. doi:10.1371/journal.pcbi.1004322

229. Scialdone A, Howard M (2015) How plants manage food reserves at night: quantitative models and open questions. Front Plant Sci 6:294. doi:10.3389/fpls.2015.00204

Publications in 2015

230. Scialdone A, Natarajan KN, Saraiva LR, et al. (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 85:54-61. doi:10.1016/j.ymeth.2015.06.021

231. Scoriels L, Salek RM, Goodby E, et al. (2015) Behavioural and molecular endophenotypes in psychotic disorders reveal heritable abnormalities in glutamatergic neurotransmission. Transl. Psychiatry 5:e540. doi:10.1038/tp.2015.26

232. Scruggs SB, Watson K, Su AI, et al. (2015) Harnessing the heart of big data. Circ. Res. 116:1115-1119. doi:10.1161/circresaha.115.306013

233. Senger S, Bartek L, Papadatos G, et al. (2015) Managing expectations: Assessment of chemistry databases generated by automated extraction of chemical structures from patents. J. Cheminform. 7: 49. doi:10.1186/s13321-015-0097-z

234. Skinner BM, Sargent CA, Churcher C, et al. (2015) The pig X and Y Chromosomes: structure, sequence, and evolution. Genome Res. 26:130-139. doi:10.1101/gr.188839.114

235. Smedley D, Haider S, Durinck S, et al. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43:W589-W598. doi:10.1093/nar/gkv350

236. Smith JA, Leonardi T, Huang B, et al. (2015) Extracellular vesicles and their synthetic analogues in aging and age-associated brain diseases. Biogerontology 16:147-185. doi:10.1007/s10522-014-9510-7

237. Sousa FL, Parente DJ, Shis DL, et al. (2015) AlloRep: a repository of sequence, structural and mutagenesis data for the LacI/GalR transcription regulators. J. Mol. Biol. (in press); doi:10.1016/j.jmb.2015.09.015

238. Spjuth O, Krestyaninova M, Hastings J, et al. (2015) Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. Eur. J. Hum. Genet. (in press); doi:10.1038/ejhg.2015.165

239. Squizzato S, Park YM, Buso N, et al. (2015) The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI. Nucleic Acids Res. 43:w585-w588. doi:10.1093/nar/gkv316

240. Stanford NJ, Wolstencroft K, Golebiewski M, et al. (2015) The evolution of standards and data management practices in systems biology. Mol. Syst. Biol. 11:851. doi:10.15252/msb.20156053

241. Stauch B, Fisher SJ, Cianci M (2015) Open and closed states of Candida antarctica Lipase B: Protonation and the mechanism of interfacial activation. J. Lipid Res. 56:2348-2358. doi:10.1194/jlr.M063388

242. Stavrakas V, Melas IN, Sakellaropoulos T, et al. (2015) Network reconstruction based on proteomic data and prior knowledge of protein connectivity using graph theory. PLoS ONE 10: e0128411. doi:10.1371/journal.pone.0128411

243. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 16:133-145. doi:10.1038/nrg3833

244. Stephan J, Stegle O, Beyer A (2015) A random forest approach to capture genetic effects in the presence of population structure. Nat. Commun. 6:7432; doi:10.1038/ncomms8432

245. Strumillo M, Beltrao P (2015) Towards the computational design of protein post-translational regulation. Bioorg. Med. Chem. 23:2877-2882. doi:10.1016/j.bmc.2015.04.056

246. Stubbington MJ, Mahata B, Svensson V, et al. (2015) An atlas of mouse CD4(+) T cell transcriptomes. Biol. Direct 10:14. doi:10.1186/s13062-015-0045-x

247. Su J, Ekman C, Oskolkov N, et al. (2015) A novel atlas of gene expression in human skeletal muscle reveals molecular changes associated with aging. Skelet. Muscle 5:35. doi:10.1186/s13395-015-0059-1

248. Sudmant PH, Rausch T, Gardner EJ, et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75-81. doi:10.1038/nature15394

249. Surakka I, Horikoshi M, Mägi R, et al. (2015) The impact of low-frequency and rare variants on lipid levels. Nat. Genet. 47:589-597. doi:10.1038/ng.3300

250. Suzek BE, Wang YQ, Huang HZ, et al. (2015) UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926-932. doi:10.1093/bioinformatics/btu739

251. Swat MJ, Moodie S, Wimalaratne SM, et al. (2015) Pharmacometrics Markup Language (PharmML): Opening new perspectives for model exchange in drug development. CPT Pharmacometrics Syst. Pharmacol. 4:316-319. doi:10.1002/psp4.57

252. Tan G, Gil M, Löytynoja AP, et al. (2015) Simple chained guide trees give poorer multiple sequence alignments than inferred trees in simulation and phylogenetic benchmarks. Proc. Natl. Acad. Sci. U.S.A. 112:e99-100. doi:10.1073/pnas.1417526112

253. Tan G, Muffato M, Ledergerber C, et al. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst. Biol. 64:778-791. doi:10.1093/sysbio/syv033

254. Tello-Ruiz MK, Stein J, Wei S, et al. (2015) Gramene 2016: comparative plant genomics and pathway resources. Nucleic Acids Res. 44:D1133-D1140. doi:10.1093/nar/gkv1179

255. Ten Hoopen P, Pesant S, Kottmann R, et al. (2015) Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Stand. Genomic Sci. 10:20. doi:10.1186/s40793-015-0001-5

256. Terfve CD, Wilkes EH, Casado P, et al. (2015) Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. Nat. Commun. 6:8033-8033. doi:10.1038/ncomms9033

257. The Cancer Cell Line Encyclopedia Consortium & The Genomics of Drug Sensitivity in Cancer Consortium (2015) Pharmacogenomic agreement between two cancer cell line data sets. Nature 528:84-87. doi:10.1038/nature15736

258. The UniProt Consortium (2015) UniProt: a hub for protein information Nucleic Acids Res. 43(D1):D204-D212. doi:10.1093/nar/gku989

259. Thiele S, Cerone L, Saez-Rodriguez J, et al. (2015) Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. BMC Bioinformatics 16:345. doi:10.1186/s12859-015-0733-7

260. Thornton J (2015) What you need to know to make the most of big data in biology. Lancet 385:S5-6. doi:10.1016/s0140-6736(15)60321-x

261. Touloumis A, Tavare S, Marioni JC (2015) Testing the mean matrix in high-dimensional transposable data. Biometrics 71:157-166. doi:10.1111/biom.12257

262. Truszkowski J, Goldman N (2015) Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. Syst. Biol. 65(2):328-333. doi:10.1093/sysbio/syv089

263. Tsang JC, Yu Y, Burke S, et al. (2015) Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biol. 16:178. doi:10.1186/s13059-015-0739-5

264. Tyagi N, Farnell EJ, Fitzsimmons CM, et al. (2015) Comparisons of allergenic and Metazoan parasite proteins: Allergy the price of immunity. PLoS Comput. Biol. 11: e1004546. doi:10.1371/journal.pcbi.1004546

265. Tyler-Smith C, Yang HM, Landweber LF, et al. (2015) Where next for genetics and genomics? PLoS Biol 13: e1002216. doi:10.1371/journal.pbio.1002216

266. UK10K Consortium, Walter K, Min JL, et al. (2015) The UK10K project identifies rare variants in health and disease. Nature 526:82-90. doi:10.1038/nature14962

267. Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian Analysis of Single-Cell Sequencing data. PLoS Comput. Biol. 11: e1004333. doi:10.1371/journal.pcbi.1004333

268. van de Wetering M, Francies HE, Francis JM, et al. (2015) Prospective derivation of a living organoid biobank of colorectal cancer patients. Cell 161:933-945. doi:10.1016/j.cell.2015.03.053

269. van den Berg BA, Reinders MJ, de Ridder D, et al. (2015) Insight into neutral and disease-associated human genetic variants through interpretable predictors. PLoS ONE 10: e0120729. doi:10.1371/journal.pone.0120729

270. van Roosmalen W, Le Devedec SE, Golani O, et al. (2015) Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant. J. Clin. Invest. 125:1648-1664. doi:10.1172/JCI74440

271. Vaudel M, Verheggen K, Csordas A, et al. (2015) Exploring the potential of public proteomics data. Proteomics 16:214-225. doi:10.1002/pmic.201500295

272. Velankar S, van Ginkel G, Alhroub Y, et al. (2015) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res. 44:D385-D395. doi:10.1093/nar/gkv1047

273. Videla S, Guziolowski C, Eduati F, et al. (2015) Learning Boolean logic models of signaling networks with Asp. Theor. Comput. Sci. 599:79-101. doi:10.1016/j.tcs.2014.06.022

274. Videla S, Konokotina I, Alexopoulos LG, et al. (2015) Designing experiments to discriminate families of logic models. Front. Bioeng. Biotechnol. 3:131. doi:10.3389/fbioe.2015.00131

275. Villar D, Berthelot C, Aldridge S, et al. (2015) Enhancer evolution across 20 mammalian species. Cell 160:554-566. doi:10.1016/j.cell.2015.01.006

276. Villaveces JM, Jimenez RC, Porras P, et al. (2015) Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. Database 2015. doi:10.1093/database/bau131

277. Villaverde AF, Henriques D, Smallbone K, et al. (2015) BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology BMC Syst. Biol. 9:8. doi:10.1186/s12918-015-0144-4

278. Viti C, Decorosi F, Marchi E, et al. (2015) High-throughput phenomics. Methods Mol. Biol. 1231:99-123. doi:10.1007/978-1-4939-1720-4_7

279. Vitsios DM, Enright AJ (2015) Chimira: Analysis of small RNA sequencing data and microRNA modifications. Bioinformatics 31:3365-3367. doi:10.1093/bioinformatics/btv380

280. Vitsios DM, Psomopoulos FE, Mitkas PA, et al. (2015) Inference of pathway decomposition across multiple species through gene clustering. Int. J. Artif. Intell. Tools 24: doi:10.1142/S0218213015400035

281. Vizcaíno JA, Csordas A, Del-Toro N, et al. (2015) 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44:D447-D456. doi:10.1093/nar/gkv1145

282. Vuckovic D, Gasparini P, Soranzo N, et al. (2015) MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. Bioinformatics 31:2754-2756. doi:10.1093/bioinformatics/btv222

283. Wagih O, Parts L (2015) Genetic interaction scoring procedure for bacterial species. Adv Exp Med Biol 883:169-85. doi:10.1007/978-3-319-23603-2_10

284. Wagih O, Sugiyama N, Ishihama Y, et al. (2015) Uncovering phosphorylation-based specificities through functional interaction networks. Mol. Cell. Proteomics 15:236-245. doi:10.1074/mcp.m115.052357

285. Warren AS, Aurrecoechea C, Brunk B, et al. (2015) RNA-Rocket: An RNA-seq analysis resource for infectious disease research. Bioinformatics 31:1496-1498. doi:10.1093/bioinformatics/btv002

286. Warren WC, Jasinska AJ, Garcia-Perez R, et al. (2015) The genome of the vervet (Chlorocebus aethiops sabaeus). Genome Res. 25:1921-1933. doi:10.1101/gr.192922.115

287. Wickramasinghe VO, Gonzàlez-Porta M, Perera D, et al. (2015) Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength. Genome Biol. 16:201. doi:10.1186/s13059-015-0749-3

288. Wilkes EH, Terfve C, Gribben JG, et al. (2015) Empirical inference of circuitry and plasticity in a kinase signaling network. Proc. Natl. Acad. Sci. U.S.A. 112:7719-7724. doi:10.1073/pnas.1423344112

289. Wimalaratne SM, Bolleman J, Juty N, et al. (2015) SPARQL-enabled identifier conversion with Identifiers.org. Bioinformatics 31:1875-1877. doi:10.1093/bioinformatics/btv064

290. Witte S, Bradley A, Enright AJ, et al. (2015) High-density P300 enhancers control cell state transitions. BMC Genomics 16:903. doi:10.1186/s12864-015-1905-6

291. Woo YH, Ansari H, Otto TD, et al. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites eLife 4: e06974. doi:10.7554/eLife.06974

292. Wood C, Burnley T, Patwardhan A, et al. (2015) Collaborative computational project for electron cryo-microscopy. Acta Crystallogr. D Biol. Crystallogr. 71(Pt 1):123-126. doi:10.1107/s1399004714018070

293. Xie X, Stubbington MJ, Nissen JK, et al. (2015) The regulatory T cell lineage factor Foxp3 regulates gene expression through several distinct mechanisms mostly independent of direct DNA binding. PLoS Genet. 11: e1005251. doi:10.1371/journal.pgen.1005251

294. Yachdav G, Goldberg T, Wilzbach S, et al. (2015) Anatomy of BioJS, an open source community for the life sciences. Elife 2015:4. doi:10.7554/elife.07009

295. Yates A, Akanni W, Amode MR, et al. (2015) Ensembl 2016. Nucleic Acids Res. 44:D710-D716. doi:10.1093/nar/gkv1157

296. Zepeda-Mendoza CJ, Mukhopadhyay S, Wong ES, et al. (2015) Quantitative analysis of chromatin interaction changes upon a 4.3 Mb deletion at mouse 4E2. BMC Genomics 16:982. doi:10.1186/s12864-015-2137-5

297. Zerbino DR, Wilder SP, Johnson N, et al. (2015) The Ensembl regulatory build. Genome Biol. 16:56. doi:10.1186/s13059-015-0621-5

298. Ziehm M, Ivanov DK, Bhat A, et al. (2015) SurvCurv database and online survival analysis platform update. Bioinformatics 31:3878-3880. doi:10.1093/bioinformatics/btv463

299. Zirbel CL, Roll J, Sweeney BA, et al. (2015) Identifying novel sequence variants of RNA 3D motifs. Nucleic Acids Res. 43:7504-7520. doi:10.1093/nar/gkv651

SERVICE TEAMS

TRAINING PROGRAMME

RESEARCH GROUPS

INDUSTRY PROGRAMME

EXTERNAL-FACING ACTIVITIES

CONTENTS

Organisation of EMBL-EBI Leadership in 2015

Scientific Services

Director
Rolf Apweiler

RESEARCH GROUPS

Pedro Beltrao group

Ewan Birney group

Anton Enright group

Moritz Gerstrung group

Nick Goldman group

John Marioni group

Julio Saez-Rodriguez group

Oliver Stegle group

Sarah Teichmann group

Janet Thornton group

SERVIC

Protein Expression, Networks and Models
Henning Hermjakob

Molecular Interactions, Networks and Complexes
Sandra Orchard (Team Leader April 2015)

Functional Genomics
Alvis Brazma

Gene Expression
Robert Petryszak
(Team Leader January 2015)

Functional Genomics Development
Ugis Sarkans

Protein Databank
Gerard

PDBe
& Inte
Sameer

EXTERNAL-FAC

Techno
Science
Steven N

Systems Application
Andy Cafferkey

Training
Cath Brooksbank

External Relations
Lindsey Crosswell

Web Production
Rodrigo Lopez

Director
Ewan Birney

Adminnistration
Mark Green

Finance

Human Resources

Grants

Strategic Projects

E TEAMS

...k in Europe (PDBe)
...Kleywegt

Protein Sequence Resources
Alex Bateman

Vertebrate genomics
Paul Flicek

Chemogenomics
**John Overington
(October 2008 to May 2015)
Anne Hersey
(Acting Team Leader)**

...Content
...gration
...Velankar

Protein Sequences, Families
and Motifs
Rob Finn

Gene Annotation and
Comparative Genomics
Bronwen Aken

Literature
Services
**Johanna
McEntyre**

Samples,
phenotypes
& ontologies
Helen Parkinson

UniProt Development
Maria-Jesus Martin

European Nucleotide Archive
Guy Cochrane

Cheminformatics and Metabolism
Christoph Steinbeck

UniProt Content
Claire O'Donovan

Non-vertebrate genomics
Paul Kersey

Genetic Variation
**Justin Paschall
(August 2012 to January 2016)**

...ING ACTIVITIES

Vertebrate Genomics Production
**Andy Yates
(Team Leader April 2015)**

...logy and
...ntegration
...Newhouse

Functional Annotation
Daniel Zerbino

Systems Infrastructure
Petteri Jokinen

Web Development
Brendan Vaughan

Industry Programme
Dominic Clark

Innovation & translation
Birgit Kerber, EMBLEM

Services

Team summaries

European Nucleotide Archive

Our team builds and maintains the European Nucleotide Archive (ENA), an open, supported platform for the management, sharing, integration, archiving and dissemination of public-domain sequence data. ENA comprises both the globally comprehensive data resource that preserves the world's public-domain output of sequence data, and a rich portfolio of services that support the research community in handling and using sequence data.

As nucleotide sequencing becomes increasingly central to applied areas such as healthcare and environmental sciences, ENA has become a foundation upon which scientific understanding may be assembled. Our users comprise data submitters, data coordinators for sequence-based studies, direct data consumers and secondary service providers (e.g., UniProt, RNAcentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress) that build on ENA services and content.

By extending and adapting software and services that started within ENA, we provide technology and data-repository support beyond ENA, including the CRAM sequence-data compression technology and the Webin submissions application.

We nurture communities of scientists and engineers in building and improving data infrastructure, helping them incorporate richer, more discoverable content and enable higher-impact, sequence-based science. The ENA and its partner databases represent the engine room of global data sharing in the life sciences, and our team members help drive progress in our domain. We focus on data standards driven by domain experts, technical interoperability and social aspects of data sharing such as policy and regulatory requirements.

Major achievements

Throughout 2015 we provided services to more than 2000 sequence data submitters. We made the ENA's content discoverable and accessible to a user base many times this size. We improved many of our services, including the Webin technical infrastructure, which provides the means for users to transfer and describe their newly generated data sets. For example, we made it easier for users to describe the materials they sequenced, and updated the programmatic interface to support both tabular and XML formats for metadata files. We enhanced the ENA website and associated Web Services through greater data integration and usability. Improvements to the presentation of assembly data include options to browse those sets of raw reads that have been aligned to a given reference. This represents a major joining together of the reads and sequence domains of ENA.

We continued to provide support for the global programmes for which we offer data coordination, including those handling marine (e.g. *Tara* Oceans, Ocean Sampling Day) and, in the context of the COMPARE project, pathogen-related data.

Pathogen surveillance and genomics

There is growing global enthusiasm for whole-genome sequencing as a platform for pathogen surveillance. It offers opportunities to apply a single, affordable technology across multiple pathogen groups, and promises to cater for vital analyses spanning public, animal, food and environmental health/safety areas. The ENA provides the robust data infrastructure required for whole-genome sequencing to come to fruition, supporting the global sharing of sequence data and providing new technologies, tools and services for pathogen surveillance. As a partner on the EU-funded COMPARE project, we are working actively to help speed up the detection of and response to infectious disease outbreaks using genome technology.

In 2015 we updated many of the core technologies required to adapt ENA to support rapid sharing of pathogen data, for example deploying a dedicated viral assembly data submission workflow in Webin, accelerating the validation of incoming data, enhancing performance in INSDC global data sharing pipelines and optimising the efficiency of data structures for the presentation and sharing of assembled bacterial genomes. We developed data reporting standards for raw and assembled viral and bacterial pathogen data sets, such that the descriptive information supplied is well suited for data discoverability and reuse. We packaged core technologies, standards and data submission services into the COMPARE 'Data Hubs'. These hubs facilitate the immediate sharing of data amongst collaborating public-health scientists. Built on the ENA and the global INSDC system, datasets shared through these hubs are structured appropriately and described richly, ready for full public release as soon as any embargo lifts. To facilitate the systematic interpretation of pathogen surveillance data flowing into ENA, we made available an Embassy Cloud compute environment to support computational workflows.

Guy Cochrane

European Nucleotide Archive
PhD University of East Anglia, 1999.

At EMBL-EBI since 2002.
Team Leader since 2009.

Quality through standards

The ENA team tracks growth in the uptake of sequencing and the emergence of innovations in the field, as these directly impact the growth and evolution of our services. In 2015 we entered the final stages of a transition from direct, manual submission processing to a system that focuses curator input using formally defined checklists and validation rules. Checklists offer a structured way to collect information about a sample, presenting attribute names alongside their definitions, usage conventions and syntactic rules (e.g. conventional expression, controlled vocabulary). This approach allows us to optimise datasets for discoverability and reanalysis across classes of data submission. It also allows ENA curators to create and edit attributes efficiently. This supports concurrent working on the system and allows for safety operations such as rollback. The system makes it possible for a single editing event to drive a change to the attribute across all checklists in which it appears, where appropriate. Such normalisation ensures a consistent experience for the data submitter, supports the capture of consistent and reliable data and, ultimately, improves the presentation of search services.

Data compression

We advanced our CRAM reference-based sequence data compression technology in 2015. We continued to offer and support CRAM as a public software package for its broadest possible use, extended the technology itself and adopted it more deeply across ENA services. We transitioned to CRAM v. 3; extended the software to include more effective, faster compression; adopted new compression codecs; improved the treatment of unmapped reads; established greater controls on data integrity under random access; and provided more support for external tools such as the widely used hts-jdk. We enriched services for CRAM as a core data format within the Webin and ENA systems, providing full support across the Webin interfaces for CRAM submission and the systematic reference indexing of all submitted raw read CRAM data files to make these reads available through genomic coordinate-based queries.

Future plans

In 2016 we will continue to work with user communities on data standards, for example extending the established Marine Microbial Biodiversity, Bioinformatics and Biotechnology (M2B3) standard, including coverage of aquaculture and blue biotechnology-related studies. We also expect further work on pathogen-related standards. We will actively seek to collaborate with further communities to target coverage gaps, with a view to having checklist coverage across all classes of incoming data. Curation of data submissions representing non-assembly annotated sequence will become a fully autonomous strand of activity in 2016, which will complete our transition to having all major submission workflows operating in a scalable, quality-assured mode.

We will implement specific computational workflows in the COMPARE Embassy Cloud system, initially covering bacterial assembly and functional annotation and typing/resistance profiling. We will further develop the COMPARE Data Hub concept to allow simpler user management and more integrated access. We will begin to construct a data portal for the pathogen surveillance community, with tailored search, browse and visualisation tools, and will continue to support data sharing and analysis efforts around emerging outbreaks.

We will extend the existing ENA system for structured analysis output data, for example for antimicrobial drug-resistance profiles and abundance profiles from ecological studies. This will allow for the agile response to submissions and data presentation for as-yet-unsupported data types. It has already been used as the basis for assembly and variation data in the EVA, submission and indexing support. Extending this system to serve as data infrastructure for EBI Metagenomics will help us improve submission and retrieval flexibility.

Selected publications

Gibson R, et al. (2016) Biocuration of functional annotation at the European nucleotide archive. *Nucleic Acids Res.* 44:D58-D66. doi:10.1093/nar/gkv1311

Cochrane G, et al. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 44:D48-D50. doi:10.1093/nar/gkv1323

Ten Hoopen P, et al. (2015) Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. *Stand Genomic Sci.* 10:20. doi:10.1186/s40793-015-0001-5

Ip CL, et al. (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res.* 4:1075. doi:10.12688/f1000research.7201.1

Mitchell A, et al. (2016) EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* 44:D595-D603. doi:10.1093/nar/gkv1195

Vertebrate Genomics

The Vertebrate Genomics team works together with the Vertebrate Annotation, Genome Analysis and Genomics Technology Infrastructure teams to create and deliver resources of the Ensembl project. It is responsible for data management for large-scale projects, including BLUEPRINT. It maintains the International Genome Sample Resource, which incorporates 1000 Genomes Project data, and the curated NHGRI-EBI GWAS Catalog. These resources are publicly available and widely used by the scientific community, including members of our team who conduct research into evolution, epigenetics and transcriptional regulation.

Vertebrate Genomics is responsible for the scientific leadership of the Ensembl project, the development of the Ensembl website and Ensembl outreach and training. We are also actively involved in the Global Alliance for Genomics and Health (GA4GH).

Variation Annotation activities in the Vertebrate Genomics team, coordinated by Fiona Cunningham, include enriching our part of the GWAS Catalog, which is created jointly with the Samples, Phenotypes and Ontologies team. This resource grew substantially upon migration to EMBL-EBI in early 2015.

Resequencing Informatics, coordinated by Laura Clarke, provides data coordination activities for many projects and consortia, including the European Bank for induced pluripotent Stem Cells (EBiSC), BLUEPRINT and HipSci. This part of our team also leads the International Genome Sample Resource (IGSR), which is a continuation of the 1000 Genomes Project.

Our research projects focus the evolution of transcriptional regulation, understanding tissue specificity and annotation of the non-coding genome. Based on comparative regulatory genomics techniques, our work provides some of the most definitive results on how transcription factor binding evolves across the vertebrate lineage. Our studies of tissue-specific gene regulation, published in 2015, included clarifying the role of CTCF, cohesin and other genomic structural proteins.

Major achievements

Ensembl

We released five comprehensive updates to Ensembl in addition to a special update of our website supporting GRCh37, the previous version of the human genome assembly. In 2015 we focused on improving and extending four key areas of the Ensembl website: new views and tools, performance and usability of existing views, support for track hubs and mirror sites. We also introduced a new visualisation for long-range connections between genomic regions such as enhancer–promoter interactions.

We promoted and facilitated the use of Ensembl by researchers, clinicians and educators through 97 in-person workshops, conference attendance, high-quality online training, sustained social media engagemen and usability testing of new interfaces. We conducted workshops in several developing countries, including Colombia, Malawi and Tanzania. We also held workshops at major conferences including ASHG, ESHG, PAG and PAGAsia.

Data Management and Coordination

In 2015 we updated the data from several major projects to reflect the new GRCh38 human reference assembly. This involved realignment of more than 70 Terabases of DNA sequence from the 1000 Genomes Project for the IGSR and updating all BLUEPRINT analyses.

We performed a major upgrade of the HipSci website to improve the discoverability of individual cell lines and related information.

We helped launch the Functional Annotation of Animal Genomes (FAANG) project, which is producing comprehensive maps of functional elements in the genomes of domesticated animal species. In this project, we led efforts to define data and metadata standards to ensure the outputs will be reusable, well into the future.

Genome-Wide Association Studies Catalog

We helped relocate the GWAS Catalog software infrastructure from the National Human Genome Research Institute (NHGRI) in the US to EMBL-EBI. The new NHGRI-EBI GWAS Catalog website serves as a single point of access, and includes an improved search interface using SOLR technology and supporting ontology expansion queries. Our curation team adds new studies to the resource, which now offers data from more than 2000 studies representing nearly 15 000 SNP–trait associations.

Paul Flicek

Genes, Genomes and Variation Resources
DSc Washington University, 2004. Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008. At EMBL-EBI since 2005.

Team Leader at EMBL-EBI since 2007, Senior Scientist since 2011.

Vertebrate Genomics Research

The most important result in 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 postdoctoral fellow Camille Berthelot. We mapped and analysed the evolution of genomic promoter and enhancer elements across 20 mammalian species, including whale and dolphin, to produce the most comprehensive view of shared and lineage-specific promoter and enhancer elements in the mammalian lineage. We found that approximately half of all enhancers active in liver tissue of any of our 20 assayed species are not found in any of the other 19 species, but that these 'recently evolved' enhancers tend to be exacted from ancestral DNA sequences shared across mammals. We also found that approximately 16% of active liver promoters and 1% of active liver enhancers are shared across the species, and that promoters have certain evolutionary characteristics similar to protein-coding genes.

We gained new insights into the role of cohesin in genome regulation in the context of a collaboration with the Merkenschlager group at Imperial College London and with former PhD student Andre Faure, now postdoctoral fellow at the CRG in Barcelona. EMBO Postdoctoral Fellow Emily Wong contributed to a study led by the Spector group at Cold Spring Harbor Laboratory in the US, investigating chromatin changes in response to a large genomic deletion.

Our research with the Spector group and the Marioni group at EMBL-EBI showed that random monoallelic gene expression, which refers to the transcription of a gene from one of two homologous alleles, increases upon embryonic stem-cell differentiation. These results support a model in which random monoallelic expression occurs stochastically during differentiation and, for some genes, is compensated for by the cell to maintain the required transcriptional output of these genes.

Future plans

In 2016 we will continue to expand the functionality and capacity of the Ensembl project, making our datasets more useful and stable to support reproducible genome interpretation. We will release new methods and tools for scaling up to accommodate ever-larger datasets and numbers of genomes. Using data from the Google Analytics Event Tracking feature, we will improve the Ensembl website experience for both first-time visitors and experienced users. We will expand our outreach and training, and continue to engage with researchers on diverse online and in-person platforms. We will grow and our major data resources, for example with new human populations for the IGSR including those from the Human Genome Diversity Project, new search interfaces for HipSci and the first release of FAANG data. In terms of GWAS, we will increase the depth of gene-mapping information to all available genes for each variant provided. To identify and display all known variants in the GWAS catalog that co-segregate with a variant of interest, we will integrate all linkage disequilibrium information.

We will build on our studies of closely related mouse species, continuing to investigate the function, inheritance and evolution of the intensity of the ChIP-seq signal (i.e. binding intensity). These studies must be performed using a higher number of experimental replicates than is standard to achieve an effective analysis. We will investigate whether ChIP-seq datasets could be used to uncover mitochondrial heteroplasmy. We will work on a comparative benchmark study between the primate and murine clades, leveraging genomes from the Mus genus.

Selected publications

1000 Genomes Project Consortium et al. (2015) A global reference for human genetic variation. *Nature* 526:68-74

Andersson L, et al. (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* 16:57

Ing-Simmons E, et al. (2015) Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Res.* 25:504-513

Kulis M, et al. (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature Genetics* 47:746–756

Villar D, et al. (2015) Enhancer evolution across 20 mammalian species. *Cell* 160:554-566

Zepeda-Mendoza CJ, et al. (2015) Quantitative analysis of chromatin interaction changes upon a 4.3 Mb deletion at mouse 4E2. *BMC Genomics* 16:982

Genome Analysis

The Genome Analysis team is responsible for the development of scalable solutions for the functional annotation of genomes, with a focus on genetics and epigenomics. Within Ensembl, we maintain the Variation and Regulation databases that collect all available data on vertebrate genetic markers and gene expression regulation.

Whereas the cost of sequencing a genome has dropped, the downstream analysis is still costly and error prone. Outside of the coding sequences, which are comparatively well understood, our knowledge of what the vast majority of the genome does is still hazy, and our ability to predict the functional consequence of most variants approximate. For example, most genetic markers that are revealed by genome wide association studies (GWAS) lie outside of coding regions. It is therefore very difficult to understand the corresponding disease mechanisms, let alone devise effective therapeutic strategies to counter them.

The first obstacle that needs to be addressed is easy access to the data. For a number of practical, commercial or ethical reasons, many genetic datasets are isolated in separate databases. It is now clear that to detect minute genetic associations researchers must pool their resources so as to create a greater common dataset. It is similarly clear that to understand the full dynamics of the epigenome, the world research community needs to coordinate a very large number of assays. With this in mind, we are active members in a number of international collaborations to collect and share data.

Having gained access to the relevant datasets, we develop technical solutions to efficiently process them, and cross-examine large numbers of experiments. We therefore develop specialised infrastructure to process multidimensional genome-wide datasets. This includes user-friendly webtools as well as more advanced command-line tools and micro-services. In particular, the Variant Effect Predictor (VEP) stands out as the most popular way of querying Ensembl for a set of DNA variants.

Downstream of our data and our tools, we deliver integrative analyses that summarise all the data that we processed. Our users can rely on these results, with the knowledge that we integrated as many experiments as possible to produce the most precise annotation possible. For example, the Ensembl Regulatory Build is the first published annotation that integrates ChIP-Seq and open chromatin experiments from multiple consortia into a single synthetic summary.

Major achievements

A lot of our efforts were focused on obtaining more datasets and encouraging healthy data sharing between research partners. We are active members of the Global Alliance for Genomics and Health (GA4GH), an organisation that strives to break the technical and legal barriers to genomic data sharing across the world. This year, the Global Alliance finished its first standard exchange protocols. We also play an important role in the International Human Epigenome Consortium (IHEC) and the Functional Annotation of Animal Genomes (FAANG) consortium, that set the standards and collect references to epigenomic datasets produced by all major consortia. Finally, we are collaborating with the GWAS Catalog project to help collect and curate available GWAS results.

We also continued to improve access to our data. We continued to improve the VEP, for example by speeding it up. We also helped develop and release user-friendly tools for epigenomics as produced by the BLUEPRINT consortium, GenomeStats and EpiExplorer. We also improved the querying of the datasets that we store locally, for example by acceleration linkage disequilibrium calculations and displaying them as Manhattan plots. Cis-regulatory interactions are making their way into Ensembl, and we are currently able to display user datasets (see Figure 1).

Finally, we kept abreast of technical development in the field. Our Ensembl Regulatory Build algorithm was published in Genome Biology. We collaborated with Peter Fraser and Mikhail Spivakov of the Babraham Institute on the downstream analysis of an exciting technology, called Promoter Capture Hi-C, that produces high-resolution maps of the 3 dimensional conformation of the chromatin. We are also working with the CTTV Epigenome Project, where we compare and contrast cell line epigenomic data.

Future plans

In the next few years, we aim to greatly expand the number of species, tissues and cell types covered by the Ensembl Regulatory Build. Having spent much of

Daniel Zerbino

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.
Team leader since 2015.

2015 to improve and automate our ChIP-Seq analysis pipelines, we will be processing most of the available datasets, as collected by IHEC. We will also integrate new data sources into Ensembl's annotation. In particular, DNA methylation datasets are very cost effective, and many experiments have been done across many species. Currently only available on human and mouse, we are expecting to expand the number of species covered by the Ensembl Regulatory Build by integrating datasets produced by the FAANG consortium. In addition to better detecting regulatory elements, we wish to better understand their components, and transcription factor binding motifs in particular. New technologies such as SELEX or Uniprobe have produced new collections of binding motifs that we want to use in our detection of possible binding sites.

Our annotation of variants will be greatly expanded by tying them empirically to nearby genes. First, we are very excited to release GTEx eQTL data in 2016. We have devoted a lot of efforts to develop appropriate HDF5-based infrastructure to store these large data matrices. Already, a prototype is running, and can provide any selection of the GTEx data in a fraction of a second. In parallel, we expect the first promoter capture Hi-C datasets to be made available around the same time. Our tissue specific annotations, as well as phenotype and disease associations, will all be enriched by the use of ontologies.

We are also preparing for the fundamental revolution brought about by CRISPR technologies. In a first pass, we will be displaying the CRISPR target sites, as computed by the WGE tool developed at the Sanger Institute. We will also be laying the groundwork to store external CRISPR screens in view of creating a central archive.

In terms of tools we will continue to improve on our current set, with a focus on the VEP and GenomeStats. We are also currently collaborating with CTTV to produce a GWAS functional analysis pipeline, which will integrate all available cis-regulatory data to impute causal genes from GWAS summary results. This new service should be initially released in 2016, and will eventually be made available through various entry points, such as the Ensembl webpage, the RESTful endpoints or as a standalone program.

Selected publications

Cunningham F, et al. (2015) Improving the Sequence Ontology terminology for genomic variant annotation. *J. Biomed. Semantics* 6:32

Nguyen N, et al. (2015) Building a pan-genome reference for a population. *J. Comput. Biol.* 22:387-401

Zerbino DR, et al. (2015) The Ensembl regulatory build. *Genome Biol.* 16:56.
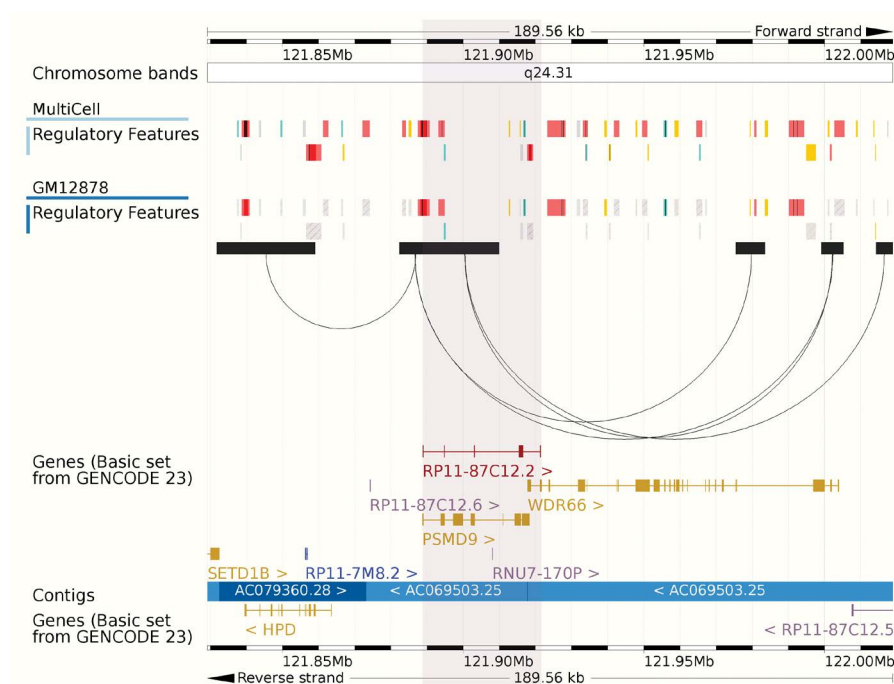
Experimental Promoter Capture Hi-C loaded and displayed on the Ensembl Genome Browser.

Vertebrate Annotation

The Vertebrate Annotation team aims to create comprehensive, up-to-date gene annotation and comparative genomics resources that further our understanding of biology, evolution and the mechanisms of disease. The data from these resources are distributed by the Ensembl project, and provide a foundation for clinical and research communities.

While the reference genome assembly is an increasingly important tool for research, most scientists working in this area need to link genomic sequence to biological function. This is made possible by gene annotations, which identify the location, structure and expression of genes. Valuable insights can be gained by comparing the annotated sequences of individuals of the same species, or across a wide range of species.

Our team produces reference gene annotation that is used by clinical, agricultural and research communities as well as by other data services at EMBL-EBI. Our gene annotation service provides high-quality gene sets for human, mouse and almost 100 other vertebrate species, including key model organisms and farmed animals. These are the primary annotations used in the initial genomic analyses of many international genome projects. In collaboration with GENCODE, we produce the gold-standard gene annotation for human and mouse.

For every genome assembly in Ensembl, we also produce comparative genomics resources that link diverse species at the DNA and gene level. These data are used for investigating gene function and evolution, adaptive traits and conservation biology.

We are responsible for TreeFam, which clusters similar gene sequences into homologous families and indicates gene history events such as duplication and speciation. Our comparative annotations include gene families, gene orthologues, whole genome multi-species alignments and conserved genomic regions across many species.

We develop alignment and annotation methods that integrate diverse data from the public archives. We collaborate closely with other service teams at EMBl-EBI including the ENA, UniProt, and Expression Atlas to annotate new assemblies and to update annotation on existing assemblies as new data become available.

Major achievements

We are proud to maintain the reference human and mouse gene sets through our GENCODE collaboration. In 2015, we improved and updated these gene resources by annotating assembly updates provided by the Genome Reference Consortium (GRC), incorporating new manual annotation, identifying gene alleles across the various haplotypes, and contributing to the CCDS project.

We released major updates to the rat and zebrafish resources. We made new assemblies available for both species, and produced genome-wide annotation for them using our evidence-based methods to identify protein-coding genes, noncoding RNA genes, and pseudogenes. We produced tissue- and sample-specific transcript sets from RNA-seq data in the public archives. We updated the multi-species whole-genome alignments, gene trees and orthologues in Ensembl to include these new assemblies.

We extended our gene-annotation methods to include annotation on lincRNA genes. We applied this method to human, mouse, rat and sheep and will be producing lincRNA annotation for more species in 2016.

TreeFam produces phylogenetic trees and orthology predictions for all Ensembl eukaryotes. The number of publicly available genomes is increasingly rapidly, providing an opportunity for new insights via comparative genomics. To achieve scalability, we designed a novel workflow that will classify protein sequences from thousands of genomes into gene families in a quick and robust manner. This workflow is now partially in production and uses our new library of Hidden Markov Model (HMM) profiles.

Our comprehensive genome annotations are the foundation for myriad downstream analysis tools and research, including the Ensembl Variation Effect Predictor (VEP). Access to consistent gene annotation for a wide range of vertebrates is important for evolutionary studies. Members of our team collaborated with others in studying gene families in the vervet (African green monkey) lineage, using freely available data from ten of our annotation projects.

Bronwen Aken

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.
Team leader since 2015.

Future plans

We will continue to develop methods for producing high-quality genome annotation, and to produce world-leading reference gene sets and comparative resources including TreeFam gene trees, orthologues, whole-genome multi-species alignments, and conserved regions.

In 2016 we plan to release the first genome assemblies annotated using our new large-scale annotation pipeline. Our goal is to improve scalability so that we can produce gene annotation in a fraction of the time it currently takes. This will accomodate the increased number of genomes being sequenced (Genome 10K Community of Scientists, 2009), which require consistent, efficient, highly automated annotation solutions that enable intra- and inter-species genome comparisons.

For human and mouse, we will update the gold standard annotation regularly, including producing gene annotation on new alternate sequences from the GRC as they arrive. The GRC has expanded the definition of the reference human genome to include genomic sequence for additional haplotypes and gene alleles, and releases new alternate sequences on a regular basis. We will provide access to the most up-to-date gene annotation, and identify genes and alleles on the new alternate sequence that could not otherwise be represented.

New technologies are giving rise to an increasing amount of genome-wide information on how transcript isoforms are expressed in various tissues, cells or developmental stages. Our transcript-reconstruction method builds transcript models using only genomic sequence and transcriptome reads as input, allowing us to identify novel genes.

We will refine this method and use it to annotate incoming genome assemblies. We will also develop methods for long-read transcriptome data such as

PacBio. This will allow us to better annotate full-length transcript isoforms, mapping them directly to the genome.

We will further develop our scalable TreeFam workflow to release gene families for all Ensembl eukaryotes, and update our gene trees, orthologues, and other comparative resources. We will explore future applications for our TreeFam HMM resource, including analysis and annotation of incoming genome assemblies. As Ensembl aims to incorporate externally annotated gene sets, our scalability enhancements to TreeFam will allow us to link a broad range of eukaryotic species.

As more species are added to a whole-genome alignment, scalability of storage and accessibility have become an issue. Together with colleagues at the University of California Santa Cruz (UCSC), who are developing a new aligner (Cactus) and a new file format (HAL) to address scalability issues, we are committed to creating a shared alignment process that will scale well and ensure consistent whole-genome alignment data between the UCSC browser and Ensembl.

As we develop our methods and workflows, we will continue to distribute our software to all groups who wish to run them on their species of choice and make the process of deploying these pipelines progressively easier.

Selected publications

Warren WC, et al. (2015) The genome of the vervet (*Chlorocebus aethiops sabaeus*). *Genome Res*. 25:1921-33

Eöry L, et al. (2015) Avianbase: a community resource for bird genomics. *Genome Biol*. 16:21.

Church DM, et al. (2015) Extending reference assembly models. *Genome Biol*. 16:13

Boeckmann B, et al (2015) Quest for orthologs entails quest for tree of life: In Search of the Gene Stream. *Genome Biol Evol*. 7:1988-99

Tan G, Muffato M, et al (2015). Current methods for automated filtering of Multiple Sequence Alignments frequently worsen single-gene phylogenetic inference. *Syst Biol*. 64:778-91

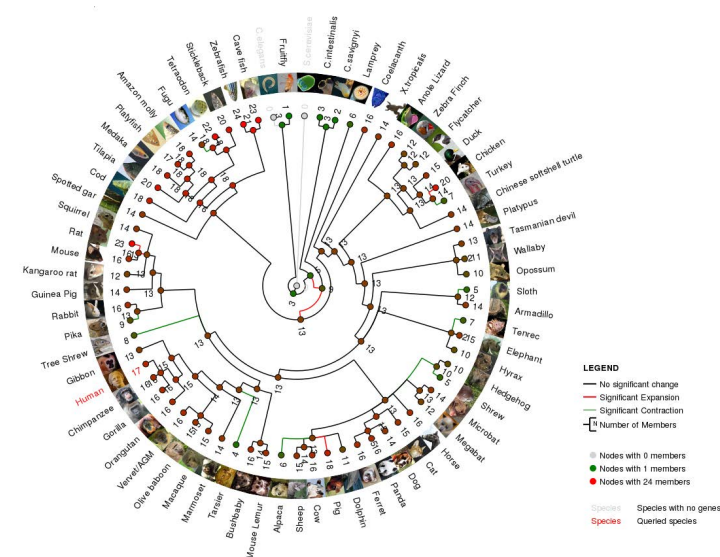Cunningham F, Amode MR, Barrell D, et al. (2015) Ensembl 2015. *Nucleic Acids Res*. 43:D662-9

Figure 1: We released a new, interactive gene gain/loss widget in the Ensembl browser. The coloured nodes can be clicked to reveal the time of divergence between taxa.

Genomics Technology Infrastructure

Our team is responsible for developing technology to store, process, access and disseminate genomic annotation. We provide Ensembl and Ensembl Genomes with its production infrastructure, and maintain database application programming interfaces (APIs), web services, schemas and data-mining tools.

Genomic annotation, a foundational service provided by EMBL-EBI, is fundamental to many downstream programs for interpreting genomic data sets. We provide a single software stack, empowering our users to explore multiple datasets on many species, in reference to several genome assemblies, and ensure a consistent interface-to-annotation experience with minimal recoding required.

For the past 15 years, Ensembl has developed APIs for data access through its MySQL databases. For many bioinformatics communities, these APIs, including the highly used Variant Effect Predictor (VEP), are a vital component of many analysis methods. They are essential to Ensembl's genome-annotation pipelines, allowing our solutions to be rolled out quickly to new species, and enable the Ensembl website to serve over 4 million users a year.

Our team is also responsible for developing new methods for data access. Our current focus is on providing efficient web service APIs to access genomic annotation. Bioinformatics is no longer a language dominated by a small set of programing languages; we can increase the value of our APIs by ensuring they appeal to as large a consumer base as possible. The Ensembl REST API, first released in 2012, was our first foray into providing these types of API and one we continue to improve upon.

We also provide access to genome annotation via our public FTP sites and data-mining tool, BioMart. Both solutions provide 'de-normalised' representations of Ensembl data customised for genome-wide analysis and batch processing. We build these resources for Ensembl and Ensembl Genomes using the same codebases, ensuring consistent, high quality and correct data across both projects.

BioMart, developed at EMBl-EBI, is a well-used service, especially by R programmers through the Bioconductor biomaRt project. EMBL-EBI projects (e.g. Gene Expression Atlas, UniProt, Reactome) build their own resources from annotations provided by these alternative, efficient access methods.

Major achievements

In 2015 we issued a beta release of a new Track Hub Registry service, which makes community-submitted datasets more discoverable. The Registry was developed as part of our BBSRC funded project ProteoGenomics. Track Hubs enable users visualise genomic annotation data within a genome browser, using a UCSC Genome Browser text format to group and configure these datasets into 'tracks'. Ensembl has supported Track Hubs since their use in the ENCODE project, when they required more user involvement. Initially, it was necessary to know the location of a dataset, then and manually attach it to the Ensembl Browser. Now, the Track Hub Registry enables users to publish Track Hub locations and provides a comprehensive search interface (see Figure). We see Track Hubs and the Track Hub Registry as the natural successor to the Distributed Annotation System (DAS).

In 2015 use of our REST API continued to grow, serving over 70 million requests for data. We made REST available for all Ensembl and Ensembl Genomes species, and provided access to annotation based on GRCh37.

We released our first Global Alliance for Genomics and Health (GA4GH) endpoints, which provide access to Ensembl-hosted variation data. We also expanded the amount of data returned from a number of endpoints, providing more comprehensive information about genes and variations. We improved the performance of our endpoints, notably of retrieving genomic variations in protein coordinates and the VEP.

We developed the first set of programmatic interfaces to access CRAM data from Perl. Our work, funded as part of the Crop Bioinformatics Initiative from the BBSRC, enabled CRAM data attachment into Ensembl; making Ensembl genome browsers among the first to support the next-generation format.

In 2015 we built new layers in our indexing schemes, FAIDX and Tabix, providing fast access to FASTA formatted sequence and tab-delimited data files, respectively. The library is available from CPAN under an Apache 2.0 license.

Andy Yates

Genome Technology and Infrastructure
MSc in Bioinformatics, University of
Manchester, UK.

At EMBL-EBI since 2006.
Team leader since 2015.

Future plans

In 2016 we will continue to develop REST into a major distribution method for genomic annotation. We will extend the REST API into a service capable of driving an Ensembl website. This requires creating extensions to all endpoints and new methods to uniquely identify both reference assembly systems and datasets that are accessible through the API. We aim to make REST the primary method for programmatically accessing Ensembl-hosted data.

As part of GA4GH, we will build on our relationship with Google Genomics, developing annotation endpoints capable of delivering genomic annotation in a single format. We see these REST APIs delivering data to a new set of client-side widgets, based on our own JavaScript components including our scrollable genome browser, Genoverse. In future, we hope to see REST applied in a distributed Ensembl annotation system, in which datasets can be hosted by third parties through compatible APIs, ready to attach into the web browser or enter our analysis pipelines.

We will begin to replace the popular BioMart data-mining tool in order to ensure we can provide a dataset-querying platform capable of handling ever-increasing volumes of data flowing into EMBL-EBI. We plan to launch our new method for fast query and retrieval of genes based on the Elastic Search tool. We will also repurpose existing APIs, such as those found in the European Variation Archive, to provide answers to complex cross-data-type queries, for example "All genes with protein-altering variations implicated in autosomal disease."

Selected publications

Cunningham F, et al. (2015) Ensembl 2015. *Nucleic Acids Res.* 43:D662-D669

Yates A, et al. (2015) The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics* 31:143-145

Smedley D, et al. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43:W589-W598

Figure: The Track Hub Registry search interface showing a search for "lymphoma" resulting in a hit to the Blueprint Epigenomics hubs for GRCh37. Additional facets are available on the left hand side of the interface with the ability to attach the hub to Ensembl on the right.

Non-vertebrate Genomics

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

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 (Giraldo-Calderon et al., 2015) for invertebrate vectors of human disease, WormBase (Howe et al. 2016) for nematode biology, PomBase (McDowall et al., 2015) for fission yeast *Schizosaccharomyces pombe*, and PhytoPath (Pedro et al. 2016) for plant pathogens. In the plant domain, we collaborate closely with Gramene in the US and with a range of European groups in the transPLANT and ELIXIR-EXCELERATE projects.

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— the type of work once the sole domain of large, internationally co-ordinated sequencing projects. We also work on large, complex genomes like hexaploid bread wheat, establishing informatics frameworks for the analysis of species for which genomic data is 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

In 2015 we issued six public releases of Ensembl Genomes. Ensembl Bacteria now includes almost 30 000 genomes from over 5000 distinct species; while the number of fungal and protist genomes included have increased approximately 10-fold and 5-fold, respectively, in one year. It is likely that we will deploy a similar, automated approach to that currently taken for incorporating microorganism genomes for those of multicellular species in 2016.

With each release we have updated cross-references and comparative genomics, introduced improved assemblies and annotations, and sourced additional

data sets, mapping them onto the relevant genomes and incorporating them into the resource.

We contributed to the regular data releases of and PomBase, VectorBase, WormBase and PhytoPath. As part of VectorBase, we contributed to the publications of the genome of *Anopheles stephensi*, the primary mosquito vector of malaria in urban India.

In WormBase, we made substantial progress towards the implementation of a new database framework that should allow for improved performance and more rapid updates to the public site. In both WormBase and PhytoPath, we released new data-mining solutions. In each project there are specific challenges, but by re-using infrastructural components in different contexts we have gained efficiencies of scale.

Community curation is a good way of capturing high-value data from the experts. We are also now running community curation portals for 30 insect vector species using the Web Apollo framework, allowing scientists to modify gene models directly for subsequent incorporation into VectorBase and Ensembl.

In PomBase, we collect functional annotations using the Canto tool. In 2015 we extended our use of Web Apollo to plant pathogens for the first time, working with the community to improve the annotation of the necrotrophic fungus *Botrytis cinerea*, and prepared to deploy Canto for these phytopathogenic species.

In December, we released a new "pre-site" offering access to a new genomic assembly for bread wheat. Bread wheat has a large, complex genome and we have been working as part of a BBSRC-funded project to develop and disseminate a new assembly through a collaboration with The Genome Analysis Centre, The John Innes Centre, and Rothamsted Research. The new assembly is the most complete, contiguous assembly yet released for this species and we will be working to annotate it fully over the course of 2016.

In the context of the transPLANT project, we continued to work with the plant science community to develop standards for phenotypic data, and set out our findings with a publication (Krajewski et al., 2015).

Paul Kersey

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.

Future plans

A major project is currently underway to automate the
identification and genomic alignment of RNA-seq data
submitted to the European Nucleotide Archive. A new
pipeline is expected to ensure that all appropriate data
sets present in the archives can be visualised through
the Ensembl interfaces. In 2016 we will address the
issues arising from these development. For example,
these efforts need to be matched with new approaches
for storing, annotating, searching and visualising
the resulting alignments. Indeed, there are similar
challenges in allowing users easy navigation of large
numbers of alignment "tracks" on an individual genome
as there are to allowing easy navigation of large numbers
of genomes. Neither problem is trivial to solve: likely
solutions involve the increased use of programmatic
access methods, coupled to tight collaboration with
archival resources to ensure the capture of the metadata
necessary to support the search functionality desired.

Selected publications

Ensembl Genomes 2016: more genomes, more
complexity. Kersey PJ, Allen JE, Armean I, et al. (2016)
Nucleic Acids Res. 44: D574-D580.

Krajewski P, Chen D, Ćwiek H, et al. (2015) Towards
recommendations for metadata and data handling in
plant phenotyping. *J. Exp. Bot.* 66:5417-5427.

Jiang X, Peery A, Hall AB, et al. (2015) Genome analysis
of a major urban malaria vector mosquito, Anopheles
stephensi. *Genome Biol.* 15:459.

Pedro H, Maheswari U, Urban M, et al. (2016) PhytoPath:
an integrative resource for plant pathogen genomics.
Nucleic Acids Res. 44:D688-D693.

WormBase 2016: expanding to enable helminth genomic
research. Howe KL, Bolt BJ, Cain S, et al. (2016) *Nucleic
Acids Res.* 44:D774-D780.
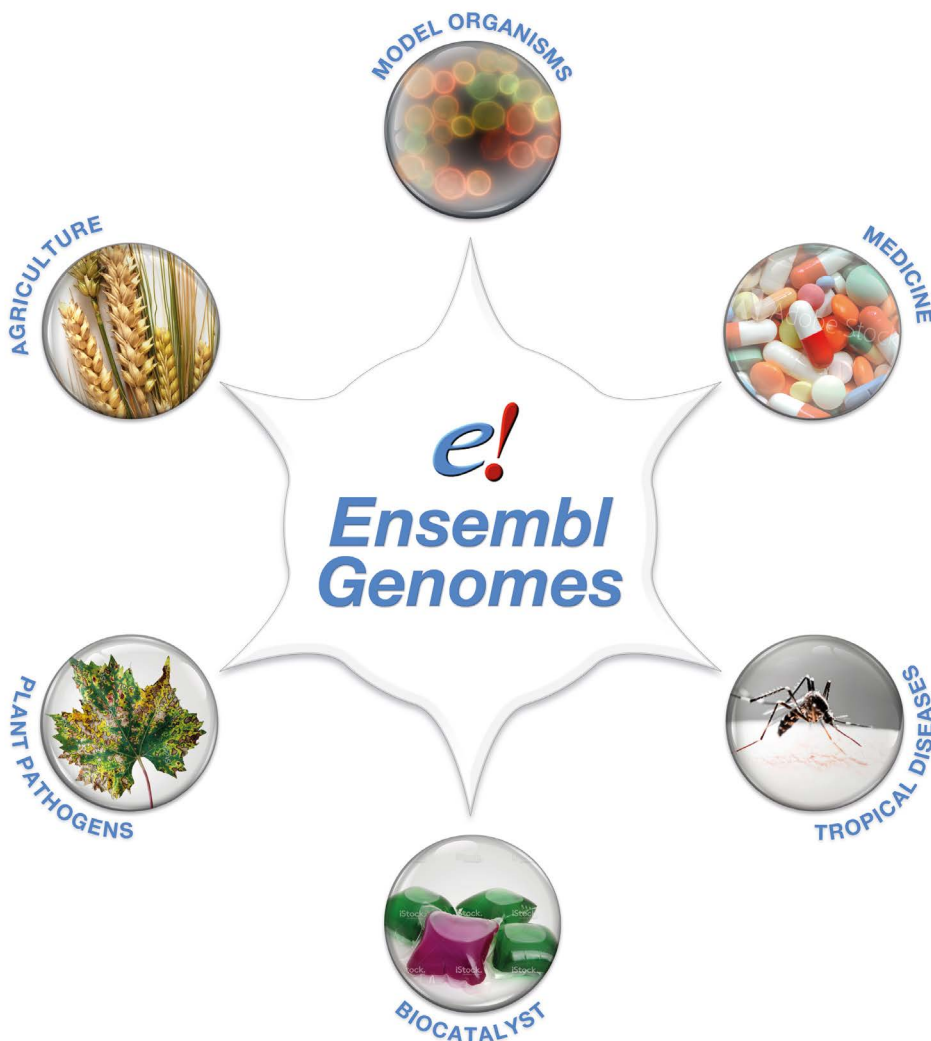
*Genomes for all: Ensembl
Genomes provides a central
resource for addressing all
areas of biological research.*

Variation Archive

Genetic variation represents the individual genomes of studied organisms and human patients as compared with one another and against the reference sequence for the species. It is a fundamental data type in molecular biology, population research and clinical investigation, and our team manages resources that make it available to researchers worldwide: the European Variation Archive (EVA) and the European Genome-phenome Archive (EGA).

Human genetic data presents particular challenges in terms of protecting participant privacy when individually unique genomes are archived for scientific research, often requiring controlled-access approval systems to ensure compliance with data access policies. The EGA supports secure, controlled-access data management for human genomes and variation data, providing a standard mechanism for providing access to data to a wide set of research users in a secure manner. It is jointly developed by EMBL-EBI and the Centre for Genomic Regulation (CRG) in Barcelona, Spain.

EMBL-EBI is a member of the Global Alliance for Genomics and Health (GA4GH), which partnered with ELIXIR in 2015 to provide genomics services that balance data protection and efficient data sharing. The project develops 'Beacons', which provide consent-based access to genomic data in the EGA as well as national resources in Finland, Sweden and the Netherlands.

Variation data is the primary analysis product of the sequencing, alignment and variant-calling pipeline to studies of population genetics, genotype-to-phenotype association and functional analysis linking the genome to molecular pathways. The EVA, the global reference catalogue of genetic variation, provides a basis for interpreting each new genome and variant observed in research and clinical studies. It includes the Database of Genomic Variant (DGVa) project, provides a primary archive service for genetic variation data and builds on EMBL-EBI's sequence-level archives (i.e. ENA and EGA), supporting value-added analysis and visualisation resources.

The European Genome-phenome Archive is now co-developed with the Centre for Genomic Regulation in Barcelona, Spain.

Together with international partners, the EVA provides a stable, accessioned database that catalogues and provides access to genetic variation in all species. This is a powerful tool for researchers working in clinical, agricultural, biotechnological and ecological research.

Major achievements

European Genome-phenome Archive

Our team handled a 50% increase in the volume of data archived in the EGA and a 65% increase in the number of files submitted; the resource grew to over three Petabytes of human genomic and variation data. We deployed a new EGA downloader service, which distributed over 1.7 Petabytes of data in 2015. In collaboration with the GA4GH, we implemented a tiered Beacon for the EGA, enabling both anonymous and registered access to a limited collection of variation data. By being a member of the Beacon Network, multiple institutions can now easily discover relevant datasets from a single access point.

We re-built the EGA pipeline to improve capacity and reliability, and achieved a reduction of the quarterly average processing time from three weeks to one and a half days. In the context of the NIH-funded Big Data to Knowledge (BD2K) project, we exposed variation data and delivered EGA content for the Omics Data Discovery Index.

Together with our federation partners and co-developers at the CRG in Barcelona, we substantially increased the resource's capacity to distribute data. The CRG now distributes files via FTP or Aspera, and EMBL-EBI distributes files via the EGA downloader. In addition, a new programmatic interface hosted at CRG provides access to publicly available metadata about studies, samples and datasets held in the Archive.

Together with our partners in ELIXIR, the GA4GH and CRG, we developed a three-tiered Beacon that provides a single point of access to datasets from multiple institutions. The Beacon Project allows users to make simple, anonymous queries on controlled-

access datasets, for example a simple yes/no question as to whether a certain allele exists at a certain position within a Beacon. We helped develop a tiered-access beacon that grants an authorised, registered user access to additional data, such as allele frequencies or extra datasets. We also co-developed a third level for controlled access, which serves users who have been granted full access to specific datasets.

European Variation Archive

In 2016 the number of species represented in the EVA increased to 22, representing substantial growth in variation data for animals and plants. Work supported by the BBSRC and the UK's Confederation of British Industry enabled us to offer datasets for plants and crops, including barley, maize, rice, sorghum and tomato.

Thanks to our participation in the NextGen project to preserve farm-animal biodiversity, the archive also offers variation data on animals including cow, goat and sheep. Other submitted datasets include chicken, chimpanzee, dog, mosquito, mouse and vervet monkey.

We made available the highly accessed human datasets from Phase 3 of the 1000 Genomes Project and from the Exome Aggregation Consortium (ExAC) v0.3.

We improved the Variant Browser in a number of ways, for example making datasets from 13 different species more accessible and integrating the variant annotations generated by the Ensembl's Variant Effect Predictor (VEP) tool. To help users detect infrequently occurring variants with potentially high impact, we updated the Variant Browser to filter by consequence type, protein-substitution score and minor allele frequency, and to display population statistics such as allele frequencies.

To guarantee the correctness of submitted data and improve data quality, we implemented a Variant Call Format (VCF) validator. Users will be able to download the results of website queries in VCF format in 2016.

We improved the representation of clinical information with a new display for data from ClinVar. To support drug discovery, we also helped standardise ClinVar data, providing ontology-based representation of disease to the new Target Validation Platform by Open Targets (formerly CTTV). In addition, as part of the GA4GH, we developed global standards and delivered a new API for archived variation data.

Justin Paschall

Variation Archive
MA 2008, Washington University St. Louis.

Team Leader at EMBL-EBI from 2012 to 2015.

Helen Parkinson

Head of Molecular Archival Resources
PhD Genetics, 1997. Research Associate in Genetics, University of Leicester 1997-2000.

At EMBL-EBI since 2000.

Future plans

We will complete our work to facilitate access to data in the EGA by authorised users, enabling them to analyse these datasets more easily by integrating them with genome viewers such as Ensembl and open-source biomedical research platforms such as Galaxy. We will also extend the downloader to allow access to indexed, encrypted files. As part of the Accelerating Medicines Partnership, we will provide federated access to the Type 2 Diabetes Portal. We also plan to develop a new front end for EGA, improve the submissions workflow, and optimise file quality control through better feedback processes with submitters and users.

Funding from the EU-funded CORBEL and ELIXIR-EXCELERATE grants will enable our team to carry out important work on establishing standards and best practice for the secure access to sensitive data. To allow federated access to the EGA, we will develop and implement standards for authentication and authorisation.

Integrating the EGA with EMBL-EBI services including Ensembl, BioSamples and the ENA will be an important part of our work in 2016. We will also integrate the IOBIO visualisation tool to provide more user-friendly display of file statistics.

The EVA will continue brokering submissions to dbSNP, saving users an extra step. Including dbSNP data on multiple species will enable us to connect users with a catalogue of variation data that submitted through external services, and all the reference SNP identifiers (rs) generated by dbSNP.

The EVA will move into a rolling release cycle in 2016, making datasets available as soon as their processing is finished. We will improve the performance of the processing pipeline and the website, creating a better experience for submitters and end users alike.

Selected publication

Lappalainen I, et al. (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics* 47:692-695

Functional Genomics

The Functional Genomics team provides bioinformatics services and conducts research in functional genomics data analysis, concentrating on high-throughput sequencing-based gene-expression and related proteomics data.

We are responsible for core EMBL-EBI resources including the Expression Atlas, which enables users to query for gene expression; the ArrayExpress archive of functional genomics data; and the emerging BioStudies database. We contribute substantially to training in transcriptomics and other EMBL-EBI bioinformatics tools.

The Brazma research group compliments the Functional Genomics service team, developing new methods and algorithms and integrating new types of data across multiple platforms. We are particularly interested in cancer genomics and elucidating relationships between transcriptomics and proteomics. We collaborate closely with several research groups at EMBL-EBI, including the Marioni, and Stegle groups.

Major achievements

BioStudies Database and the Expression Atlas

In 2015 we released BioStudies (McEntyre et al., 2015): a new database that holds descriptions of biological studies, links to data from these studies in other databases and data that do not fit in the structured archives at EMBL-EBI. BioStudies can accept a wide range of study types described using a simple format. Developed jointly with the Literature Services team, it enables authors to submit supplementary information and link to it from the manuscript publication. Data from 558 182 studies are available from BioStudies Database.

We increased the content of the RNA-sequencing-based Baseline Expression Atlas significantly, releasing the first large-scale proteomics data on protein expression in human tissues (Petryszak et al, 2015). Taken together, the Baseline and Differential Expression Atlases now offer data from 2620 studies and over 97 484 assays. In addition to the growth of data volume and diversity, we implemented many user interface improvements.
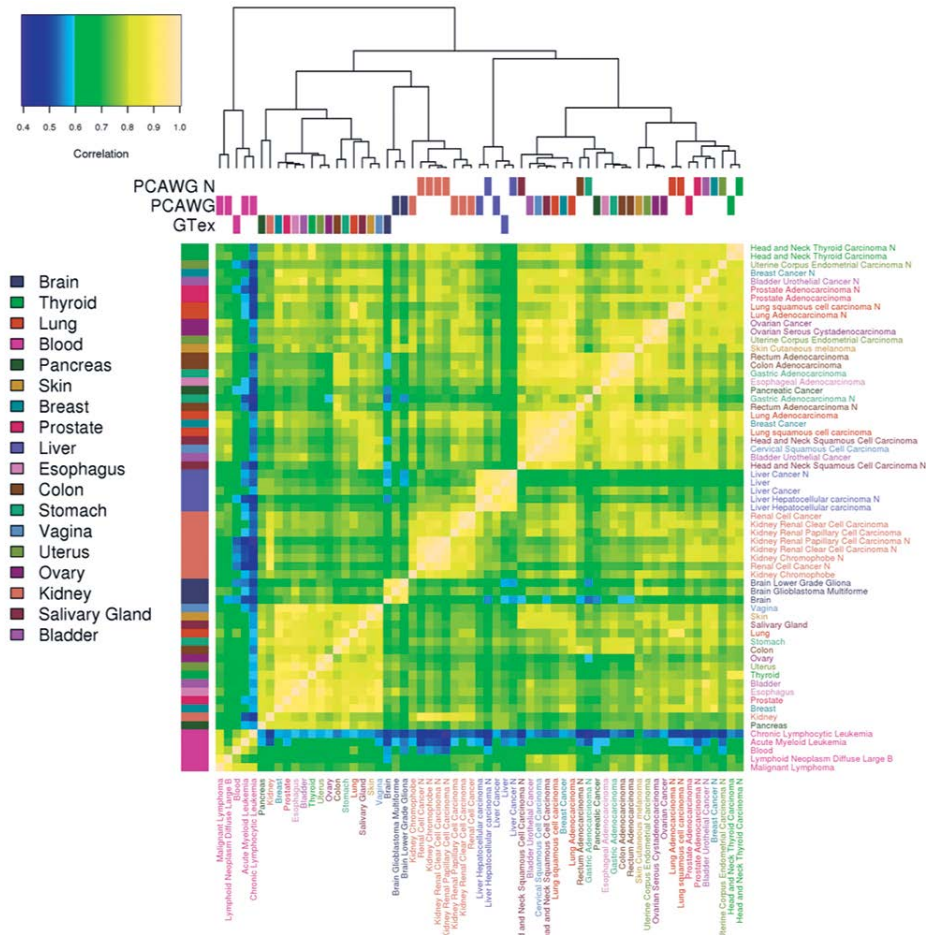
Research

In research we focused on two related areas: comparison of transcript and protein expression levels, and data integration for cancer genomics. We compared gene expression profiles, in both transcript and protein levels, across multiple human tissues, using several large-scale datasets. Overall we showed a higher level of correlation than has been reported previously, even in cases where different samples were used in transcriptomics and proteomics experiments. We continued our research into isoform-level gene expression, comparing data at transcript and proteome levels. A publication describing this research is under review, and several more are in preparation.

Together with colleagues at the University of California Santa Cruz and Memorial Sloan Kettering Cancer Center, we led a working group on RNA and DNA data integration for the Pan-cancer project of the International Cancer Genome Consortium. As a part of this work we also collaborated closely on a number of investigations with the Stegle group at EMBL-EBI and the Korbel group at EMBL Heidelberg, and prepared the results of these analyses for publication.

Future plans

Integration of baseline RNA sequencing gene expression and proteomics data will be the focus of our development of the Expression Atlas. The new BioStudies database will serve as the back-end for dealing with new types of data, including molecular imaging data.

Large-scale data integration and systems biology will remain the focus of our research. We will extend our work on cancer genomics as a part of the pan-cancer project of the ICGC, in which we are co-leading the transcriptomics/genomics integration working group that aims to study aberrant transcription patterns across many cancer types. We will also expand our research into dominant transcripts to protein abundance data.

Alvis Brazma

Functional Genomics
PhD in Computer Science, Moscow State
University, 1987. MSc in mathematics, University
of Latvia, Riga.

At EMBL-EBI since 1997.

*Joint clustering of cancer and normal tissue samples from RNA sequencing-based gene expression
datasets derived from over 3000 samples from the International Cancer Genome Consortium
Whole Genome Pan-cancer Project and the Genotype-Tissue Expression project.*

Selected publications

McEntyre J, Sarkans U, Brazma A (2015) The BioStudies
database. *Mol. Syst. Biol.*. 11:847. doi: 10.15252/
msb.20156658

Petryszak R, et al. (2016) Expression Atlas update-an
integrated database of gene and protein expression
in humans, animals and plants. *Nucleic Acids Res.*
44:D746-D752. doi: 10.1093/nar/gkv1045

Frankish A, et al. (2015) Comparison of GENCODE and
RefSeq gene annotation and the impact of reference
geneset on variant effect prediction. *BMC Genomics*
16:S2. doi: 10.1186/1471-2164-16-S8-S2

Functional Genomics Development

Our team develops software for ArrayExpress, a core EMBL-EBI resource, and the BioStudies database, a resource for biological datasets that do not have a dedicated home within the institute's services. We also contribute to the development of BioSamples, which centralises biological sample data.

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

BioStudies

The BioStudies database holds descriptions of biological studies, and links to data from these studies in other databases at EMBL-EBI and beyond. It is a repository for supplementary data files from published life-science experiments that do not fit in the structured archives at EMBL-EBI.

In 2015 we redeveloped the BioStudies user interface to reflect the recommendations of a user experience study. We drew on a wide range of information sources to compile data for a comprehensive initial release of the database. This included supplementary information from articles in Europe PMC, studies from the EurocanPlatform project, datasets from diXa and HeCaToS toxicogenomics studies, imaging data from the Image Data Repository project and original, unsolicited submissions to BioStudies. This initial work enabled us to further refine our data-management principles and fine-tune the functionality of the user interface. A simple data submission tool will be launched in early 2016.

ArrayExpress

We continued to maintain ArrayExpress data-management tools and access interfaces, focusing on a major redesign of the post-submission experiment management tool. We continued to improve and automate the management of sequencing data, and collaborated with the European Nucleotide Archive team on methods for storing raw sequence data and information.

We put significant efforts into further improving Annotare, our main data-submission tool (originally released in 2014). The emphasis of our work was on addressing the needs of curators in handling submitted datasets of varying levels of quality, and improving the file-upload functionality.

BioSamples

We continued to collaborate with the Samples, Phenotypes and Ontologies team on the basic infrastructure components of the BioSamples database, perhaps most notably in the context of the BioMedBridges project. We built a prototype system that links information from biobanks across BioSamples, the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) Hub and the Resource Entitlement Management System (REMS). The system now uses the Shibboleth system for authentication and access-rights control.

Toxicogenomics and biomedical informatics

Our work on the diXa toxicogenomics data warehouse—an Innovative Medicines Initiative (IMI) project—fed into the development of BioStudies from the outset. In 2015 we transferred datasets from the dedicated diXa data warehouse to the BioStudies database, and developed new functionality. We also started work on the HeCaToS project, which represents the continuation of diXa in terms of data management.

We participate in a number of medical informatics projects, for example providing data management solutions for the EU-AIMS project on autism spectrum disorder. As a part of the European Medical Information Framework (EMIF) project for the reuse of patient health records in clinical research, we are building a cloud environment for multi-omics data management and analysis, providing and integrating (among other components) our R-cloud scientific computation infrastructure, several Docker-enabled data-analysis pipelines, and the iRODS data-management system.

Ugis Sarkans

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.

Future plans

In 2016 we will release the BioStudies submission tool and, in the interests of providing users with a consistent experience, we will work with teams throughout EMBL-EBI to align further developments with those of other submission systems.

To make BioStudies an appealing destination for supplementary data files in the life sciences, we will work on the presentation of content. In particular, we will further improve the BioStudies 'project' feature, by which information on studies can be rendered in such a way that they reflect the project/domain to which they belong.

We will work with large projects that generate data on diverse 'omics platforms to ensure that the data ends up in the correct structured repositories. We will use the BioStudies database to tie various modalities together, and provide study metadata and robust links to the BioSamples database. This work will support to the institute's efforts to build a multi-omics atlas.

The BioStudies and BioSamples databases represent two orthogonal components for dealing with multi-omics data at EMBL-EBI. BioStudies contains overall study descriptions and grouping assays, and BioSamples groups samples and provides annotation. We will continue integrating the two systems, such that information on samples in BioSamples is readily available for BioStudy submissions.

We will further develop the Annotare tool for data submissions and, as for the BioStudies submission tool, will align it with other submission tools. We will also further simplify and automate the ArrayExpress data flow.

Our continued participation in medical informatics projects will provide us with a better understanding of the data types and data-management patterns in a range of life-science communities, and this knowledge will be essential for our work in developing the BioStudies database.

Selected publications

Hendrickx DM, Aerts HJ, Caiment F, et al. (2015) diXa: a data infrastructure for chemical safety assessment. *Bioinformatics* (Oxford, England) 31:1505-1507

Kirsanova C, Brazma A, Rustici G, Sarkans U (2015) Cellular phenotype database: a repository for systems microscopy data. *Bioinformatics* (Oxford, England) 31:2736-2740

Kolesnikov N, Hastings E, Keays M, et al. (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.* 43:d1113-d1116

McEntyre J, Sarkans U, Brazma A (2015) The BioStudies database. *Mol. Syst.Biol.* 11:847

The BioStudies database is a repository for supplementary data files from published life-science experiments that do not fit in the structured archives at EMBL-EBI.

Gene Expression

The Gene Expression team handles the acquisition, curation, quality control, statistical analysis and visualisation of functional genomics data at EMBL-EBI, focusing on microarray, high-throughput sequencing-based gene expression and related proteomics data.

We are responsible for several core EMBL-EBI resources, including the Expression Atlas, which enables users to query for information about gene expression, and the ArrayExpress archive of functional genomics data. We 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.

We are a centre of excellence for RNA-sequencing quality control and analysis, the results of which are used by numerous resources at EMBL-EBI and externally. We are increasingly interested in epigenetic analysis, for example methylation, and work towards placing transcriptomic data in a broader regulatory context.

We are part of Open Targets (formerly the Centre for Therapeutic Target Validation, CTTV) and the Cancer Genome Atlas Pan-Cancer analysis project. Analysis and visualisation on plant data is also a major component of our work through our involvement in Gramene project.

We collaborate closely with the Brazma, Marioni, Stegle and Teichmann research groups at EMBL-EBI and with the Choudhary group at the Wellcome Trust 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.

Major achievements

ArrayExpress, Expression Atlas and related projects

In 2015 we capitalised on the deployment and continual improvement of Annotare, the ArrayExpress submission tool, and focused our curation efforts on datasets in the Expression Atlas, which held 100 000 assays in December 2015 (a six-fold increase compared to 2014). These assays included 157 RNA-seq experiments, over 7000 differential comparisons across 26 organisms, and 568 plant experiments.

At the end of 2015 the Baseline Expression Atlas contained 46 RNA-seq studies, including data from many high impact studies (e.g. GTEx and FANTOM5) and its first proteomics study.

We improved the Expression Atlas interface substantially, applying many enhancements in its presentation of search results (e.g. faceting). We developed new functionalities that will be available to users in early 2016, for example gene co-expression and a new Bioconductor package for easy access to Atlas data in R language. The Expression Atlas now contributes transcriptomic data and visualisations to many resources, including the Open Targets (formerly CTTV), Ensembl, Reactome, Plant Reactome and International Mouse Phenotyping Consortium portals.

We developed an RNA-seq pipeline and adapted it to help analyse public RNA-seq data for major species in the European Nucleotide Archive's Sequence Read Archive. This functionality resulted in 148 000 processed sequencing runs in 85 species by the end of the year. Where applicable, this data is included in both the Expression Atlas and Ensembl.

Future plans

In 2016 our development efforts for the biology-centric Expression Atlas will centre on integration of baseline RNA-sequencing gene expression and proteomics data. The BioStudies database, developed by the Sarkans team, will serve as the back-end for dealing with new types of data, including molecular imaging data.

We will continue to expand our analyses and develop intuitive visualisation methods for both the existing data in Expression Atlas and for novel data types, such as epigenetic (methylation), genetic (eQTL), single-cell RNA-seq and smallRNA-seq. We will also complete the analysis of public RNA-seq data in major species and make the raw results available publicly.

As a part of the pan-cancer project of the ICGC, we will continue to investigate aberrant transcription patterns across many cancer types.

Robert Petryszak

Gene Expression
MPhil in Computer Speech and Language
Processing, University of Cambridge, UK.

At EMBL-EBI since 2003. Team Leader since 2015.

Selected publications

Frankish A, et al. (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* 16 Suppl 8:s2

Kolesnikov N, Hastings E, Keays M, et al. (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.* 43:D1113-D1116

Petryszak R, et al. (2016) Expression Atlas update-an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 44:D746-D752

Tello-Ruiz MK, et al. (2016) Gramene 2016: comparative plant genomics and pathway resources. *Nucleic Acids Res.* 44:D1133-D1140
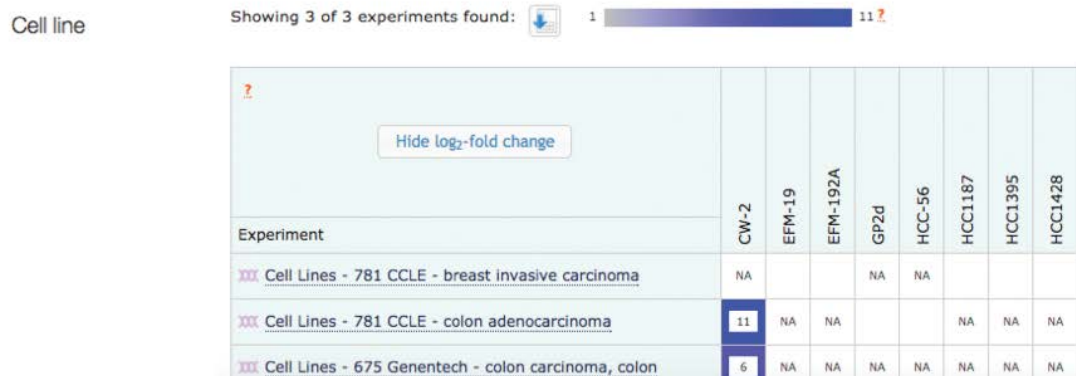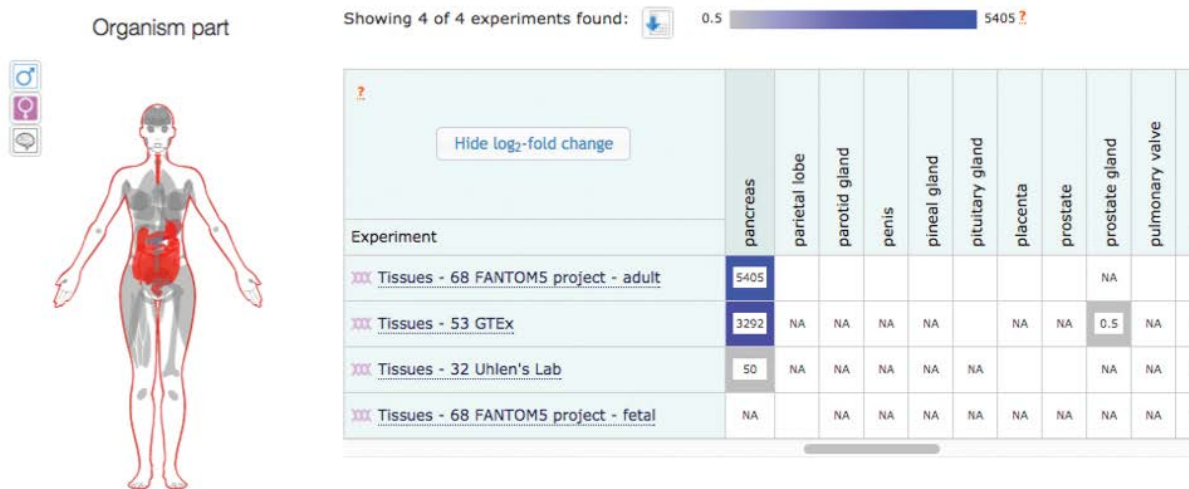
Expression Atlas: baseline expression in tissues and cell lines for human gene REG1B.

Sequence Families

The Sequence Families team is responsible for the InterPro, Pfam, RNAcentral and Rfam data resources, and coordinates the EBI Metagenomics project. We are also responsible for the HMMER web server, which performs extremely fast protein-homology searches.

InterPro integrates protein-family data from 11 major sources, and classifies hierarchically the different protein family, domain and functional site definitions to provide a unified view of diverse data. InterPro curators build on this, adding functional descriptions to the families and annotating entire datasets with structured Gene Ontology terms. The InterProScan tool allows users to identify InterPro entries on protein sequences. Pfam generates new protein family entries and has the largest sequence coverage of all the InterPro member databases. Both InterPro and Pfam have many important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome/metagenome annotation. HMMER, a sequence-similarity search tool for identifying distant relationships, enables users to infer function, annotate protein sequences and trace evolutionary histories.

Rfam classifies non-coding RNA sequences into families, using probabilistic models that take account of both sequence and secondary-structure information, termed covariance models. Rfam is used to annotate non-coding RNAs in genome projects and is a major contributor to RNAcentral, an integrating database of non-coding RNA sequences that serves as a single entry point for searching and accessing data from over 20 established RNA resources. RNAcentral gives users a unified view of all non-coding RNA types from all organisms.

Metagenomics is the study of the sum of genetic material found in an environmental sample or host species, using next-generation sequencing (NGS) technology. The EBI Metagenomics service enables researchers to submit sequence data and descriptive metadata to public nucleotide archives. We help ensure the data is functionally and taxonomically analysed, and the results primed for visualisation and download.

Major achievements

InterPro, Pfam and HMMER

In 2015 we issued six releases of InterPro. We integrated over 2000 new member database signatures, resulting in over 1800 new InterPro entries. Version 54.0, released in October, included an update to PANTHER 10.0, which involved a substantial rebuild of the database and enabled the addition of almost 60 000 new signatures and the removal of over 20 000. Although this rebuild

resulted in the loss of over 500 of the 3673 integrated InterPro entries, the curation team ensured the resource ultimately comprised over 4600 integrated PANTHER signatures by the end of the year.

Despite the removal of redundant bacterial proteomes from UniProtKB (see UniProt Development), InterPro v55.0 coverage remained substantial (79.7% of UniProtKB proteins). InterPro also continued to be a major provider of GO terms, with the latest release assigning almost 110 million terms to around 35 million proteins in UniProt release 2016_01.

InterPro's member databases use diverse analysis algorithms, and our team collaborates with several groups to improve analysis throughput. Release 50.0 saw the incorporation of a new version of PIRSF, which uses the HMMER3 analysis algorithm. HMMER3 runs about 1000 times faster than HMMER2.0, allowing InterPro to calculate the growing volumes of UniProtKB match data in a timely manner. InterPro release 51.0 debuted a sequence database pre-filtering heuristic to reduce the time it takes to calculate matches against the HAMAP database. This sped up our protein-match-generation process, and bolsters our capacity to handle future sequence data growth.

We completed the migration of Pfam from the Wellcome Trust Sanger Institute to EMBL-EBI in 2014, and in 2015 focused on issuing more frequent releases of the databases. We changed the underlying database to UniProt reference proteomes, rather than the whole of UniProt. Using this subset, which provides excellent representation of the whole of UniProt, brings significant efficiencies. Training Pfam probabilistic models on this representative set of sequences demonstrated no reduction in sensitivity. Furthermore, it reduces curation overheads, speeds up production and reduces the amount of data to present via the website. Pfam still calculates matches for all UniProt, and these remain accessible for download.

To streamline and improve our curation processes, we modified our quality-assurance measures to allow a limited overlap between Pfam entries. This new measure actually improves Pfam definitions, and helps speed up entry curation. In 2015, these improvements enabled the production of Pfam releases 28.0 and 29.0

We migrated the HMMER website from Janelia Research Campus to EMBL-EBI, where it provides

Rob Finn

Sequence Families
PhD in Biochemistry, Imperial College London;
Wellcome Trust Sanger Institute, 2001-2010;
Janelia Research Campus, 2010-2013.

At EMBL-EBI since 2014.

access to the HMMER algorithm in the context of global sequence and protein family resources. This task was non-trivial, as the HMMER search service relies on relational databases, a NoSQL resource, the HMMER search engine and associated web servers. We expanded the service to include PIRSF profile hidden Markov model libraries and to support the UniProt Reference Proteome sequence database.

Metagenomics and RNA resources

Our Metagenomics service grew to over 150 projects (up 60% over 2014), two of which were major marine metagenomics projects: Ocean Sampling Day (OSD) and Tara Oceans. Each contained roughly the same number of samples, but the Tara Oceans project was over 100 times bigger than OSD in terms of volume. We analysed over 100 billion raw nucleotide reads, more than five times the total number of reads in EBI Metagenomics at the start of 2015. This necessitated updating the analysis pipeline. We developed a pipeline-update mechanism with increased modularity; a facility to handle results from different versions within the website; and documentation and provenance of the pipeline and constituent tools. The updated pipeline increased throughput, thanks to improved speed and stability. To further aid navigation, we annotated our projects with the source biome, providing a coarse categorisation from which a user can delve into the results. Based on user feedback, we also introduced summary files for the taxonomic and functional analysis provided by EBI Metagenomics, which aggregate summary reports for each sample analysed in a single, convenient download.

We merged the Rfam and RNAcentral teams in 2015, as these resources are as synergistic as Pfam and InterPro. Late in the year, we prototyped software that will enable us to change Rfam to a more genome-centric database. RNAcentral grew substantially, with three data releases including 12 new member databases, 1.2 million new distinct RNA sequences and 9.3 million new cross references. To make RNAcentral identifiers more useful for biocuration, we developed new species-specific identifiers. We improved the website, implementing metadata search-result export functionality, and migrated the sequence search to use nhmmer, which provides a fast, sensitive sequence search.

Outreach and training

We participated in a wide range of training and outreach activities in 2015, including training workshops at EMBL-EBI highlighting the utility of EBI Metagenomics. These included workshops via the Training Programme; as part of the Micro B3 Ocean Sampling Day Analysis meeting; and in collaboration with the BBSRC-funded 'ComMet' metagenomics network. We gave several invited talks at national and international meetings, and delivered webinars on our services. We also developed training materials for new Train online courses and updates for InterPro, EBI Metagenomics, RNAcentral and Rfam. We hosted academic visitors from countries all over the world, including the US, Malawi and Australia.

Future plans

In 2016 we will continue to integrate InterPro and Pfam into a unified resource. We will also share prototypes for a new InterPro website, encompassing much of the functionality of Pfam. To support this new site we will design a new data warehouse, available to users wishing to perform more complex queries. We will integrate new InterPro member databases, extending functionality to include annotation of per-residue features such as active-site residues or metal-binding residues, thereby enabling InterProScan to perform more fine-grained functional annotation. Two releases of Pfam will incorporate several hundred new entries that will be fast-tracked into InterPro. We will improve the data flow between Rfam and RNAcentral, and will issue two releases of each database. We also plan to complete the move of Rfam to a more genome-centric view of non-coding RNA. We have laid the groundwork for absorbing massive growth of data in EBI Metagenomics, and will adapt this infrastructure proactively. We will expand the scope of the service to include public datasets housed in the ENA but not submitted directly to EBI Metagenomics. As the field shifts focus from bacteria to other kingdoms of life, we will assess and implement other tools for taxonomic assignment.

Selected publications

Finn RD, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 44:D279-D285

Finn RD, et al. (2015) HMMER web server: 2015 update. *Nucleic Acids Res*. 43:w30-w38

Mitchell A, et al. (2016) EBI Metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, 44:D595-D603

Nawrocki EP, et al. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*. 43:d130-d137

RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res*. 43:d123-d129

Protein Function Development

The work of our 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. We develop software and services for protein information in the UniProt, Gene Ontology (GO) annotation and enzyme data resources at EMBL-EBI. We are also responsible for developing tools for UniProt and GO Annotation (GOA) curation, and for the study of novel, automatic methods for protein annotation.

Major achievements

The UniProt website facilitates the search, identification and analysis of gene products. In 2015 our team released new web interfaces and functionalities, all built in response to user feedback gathered in a number of user workshops, usability interviews/sessions, helpdesk reviews and surveys. We now offer better ways to customise search and present results. A new UniProt course in Train online allows users to browse, explore and analyse the profoundly rich, integrated collection of protein sequence data in this resource.

Prior to the April 2015 release of UniProt, the UniProt Knowledgebase (UniProtKB) had doubled in size over the previous year to over 90 million entries, with a high level of redundancy. This was especially the case for bacterial species, where different genomes of the same bacterium have been sequenced and submitted independently (e.g. 4080 proteomes for *Staphylococcus aureus*, comprising 10.88 million entries). To deal with this redundancy, we developed a procedure to identify highly redundant proteomes within species groups. We implemented this procedure for bacterial species and the sequences corresponding to redundant proteomes (approximately 47 million entries) were moved from UniProtKB to the UniProt Archive (UniParc), where they are still available. This is the first concerted effort in a public protein database to deal firmly and effectively with redundancy in big data.

We released a new version of the UniProt Java API that improved several issues, for example frequent library updates, retrieval speeds and server availability. With the new API, users can create their own UniProt service, query and retrieve sets of proteins of interest, for instance all records updated in the past few months, or belonging to a particular family or species.

Our team worked with the UniProt user community as well as the NCBI RefSeq, Ensembl and Ensembl Genomes teams to provide a collection of non-redundant reference proteomes, and to maintain well-annotated organisms for biomedical and biotechnological research. New species released in 2015 include *Theobroma cacao* (cacao / cocoa), *Brassica napus* (rapeseed) and *Papio anubis* (olive baboon), among others.

In collaboration with genomics resources Ensembl and COSMIC, we created data links between DNA sequences and the functional proteins they encode. Cross-references to specific genomic sequences are now provided for each protein isoform. We also began distributing variants with consequences at the protein level for human and other species, and released variants from external resources including the Exome Aggregation Consortium (ExAC) and the Exome Sequencing Project (ESP) in the protein context.

We introduced new genome annotation track files in two formats, BED and bigBed, which allows users to map and visualise UniProtKB sequence feature annotations including domains, sites and post-translational modifications as genome browser tracks. These can be visualised in Ensembl, the UCSC Genome Browser and NCBI Genome. This beta release of the UniProt genome annotation tracks resource contains sequence annotations only for human; other species will be added in future.

We worked with the ProteomeXchange resources such as PeptideAtlas and MaxQB to provide experimental peptides from publicly available mass-spectrometry studies for UniProt proteins for several reference species.

In 2015 our team extended the functionality of our automated annotation system, which assists in the curation of the
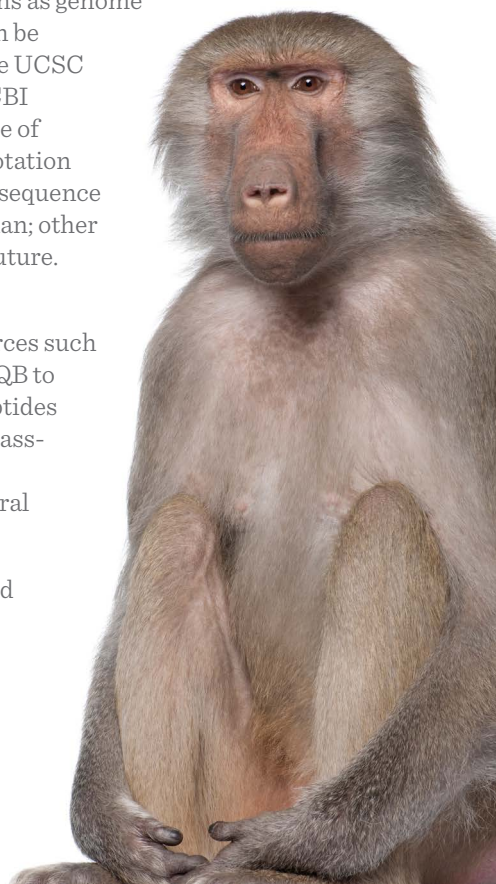
Maria-Jesus Martin

Protein Function Development
BSc In Veterinary Medicine, University
Autonoma in Madrid. PhD in Molecular Biology
(Bioinformatics), 2003.

At EMBL-EBI since 1996.
Team Leader since 2009.

millions of proteins in UniProt. Informed by specialist biocurators, the automated system adds as much useful information as possible to imported sequences, which now include domains, signal, transmembrane and coil regions. We extended UniRule and the Statistical Automatic Annotation System (SAAS), two systems for the automatic annotation of large volumes of uncharacterised proteins. These are now available through newly implemented interactive web pages, allowing our users to browse annotation rules. We also started to work in a service to download and/or use these rules as a system for genome annotation. We extended our collaborations with external automatic annotation communities including the Biofunction Prediction and Critical Assessment of Function Annotation initiatives, which will expand our knowledge and use of functional prediction methods.

In 2015 we further extended the scope of GO annotation to support annotations to RNA, identified by RNAcentral identifiers. We made significant changes to our database and Protein2GO, the web-based GO curation tool used by UniProt and GO Consortium curators to contribute annotations to the GOA project, in order to support a number of changes to annotation format and rules agreed by the GO Consortium. We re-engineered our pipeline that verifies the taxonomic correctness of GO annotations using a much-extended set of taxonomic constraints that originate from both GO and other ontologies, principally UBERON.

To make GO protein–protein interaction annotations available for visualisation in tools such as Cytoscape, we implemented a PSICQUIC (Proteomics Standards Initiative Common Query Interface) server, available through EMBL-EBI's PSICQUIC portal.

Our team maintains the Enzyme Portal, a resource that integrates enzyme-related data for all relevant EMBL-EBI resources and the underlying functional and genomic data. We re-launched the service, which now features improved interfaces and functionalities and provides a one-stop shop for all information available on enzymes. To further improve the discoverability of enzyme data, we collaborated with the Web Production team to refine the enzyme search within the EBI-Search and EBI Blast sequence search tools.

Future plans

In 2016 we plan to release a protein-sequence feature viewer that summarises functional sites in the UniProt web site. We will continue to engage with user communities working in functional prediction, and explore methods and data-exchange mechanisms to improve accuracy and coverage of protein annotations. We will maintain our focus on usability and engage

with our users to ensure we maintain a global genome/ proteome- and gene-product-centric view of the sequence space. We aim to expand our collaboration with the ProteomeXchange resources in the integration of post-translational modifications in UniProtKB, and in the provision of experimental, unique peptide mappings for reference species. We will continue to co-operate with variation projects such as ExAC to integrate relevant genome and proteome information.

Restructuring GO electronic annotation pipelines, principally those based on orthology supplied by Ensembl, will help us improve the quality of the projected annotations. We will continue the work undertaken on behalf of the GO Consortium in 2015 to transition from using UniProt cross-references rather than MOD-supplied mapping files to map from "foreign" identifiers to UniProtKB accessions. We also plan to revise the set of annotation files that we publish and submit to the GO Consortium. We plan to release a new QuickGO with re-designed interfaces and new features to improve the overall user experience. We will also continue to develop Protein2GO to keep it in line with changes in annotation strategy agreed by the GO Consortium, and to introduce additional function to enhance curators' workflow. We will make use of the enhanced set of Web Services provided by the new QuickGO to provide improved searching capabilities, and the ability to use any available ECO evidence code.

Following the successful relaunch of the Enzyme Portal in 2015, we will expand its functionalities in response to user needs and create new training activities.

Selected publications

Alpi E, Griss J, et al. (2015) Analysis of the tryptic search space in UniProt databases. *Proteomics* 15:48-57

Huntley RP, et al. (2015) The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res*. 43:d1057-d1063

Pundir S, Magrane M, Martin MJ, O'Donovan C, UniProt Consortium (2015) Searching and navigating UniProt databases. *Curr. Protoc. Bioinform.* 50:1.27.1–1.27.10

UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*. 43:d204-d212

Protein Function Content

One of the central activities of the Protein Function Content team is the biocuration of our 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 this 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

As a core contributor to the Consensus CDS project, UniProt is creating an authoritative complete proteome set for *Homo sapiens* in close collaboration with the RefSeq annotation group at the National Center for Biotechnology Information (NCBI) and the Ensembl and HAVANA teams at EMBL-EBI and the Wellcome Trust Sanger Institute. A component of this effort involves ensuring a curated and complete synchronisation with the HUGO Gene Nomenclature Committee (HGNC), which has assigned unique gene symbols and names to 39 000 human loci (19 003 of which are listed as coding for proteins). Information on the reviewed set of 20 199 entries is available on the UniProt website.

We play a major role in establishing minimum standards for genome annotation across the taxonomic range, largely thanks to collaborations arising from the annual NCBI Genome Annotation Workshops, which are attended by researchers from life science organisations worldwide. These standards have contributed significantly to the annotation of complete genomes and proteomes and are helping scientists exploit these data to their full potential.

The UniProt Automatic Annotation effort made great strides in 2015. We increased the number of UniRules significantly, with an emphasis on enzymes across the taxonomic space to enable us to respond to the need for annotation of uncharacterised genomes. We began establishing relationships with sequencing and annotation centres such as Genoscope to share these rules and to expand into new approaches.

The UniProt GO annotation program provides high-quality GO annotations to proteins in UniProtKB. The assignment of GO terms to UniProt records is an integral part of UniProt biocuration. UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users. Our curators are key members of the GO Consortium Reference Genomes Initiative for the human proteome and provide high-quality annotations for human proteins. In 2014, we provided a manually curated set of human proteins for the validation of the computational approaches submitted to for the Critical Assessment of Function Annotation experiment (CAFA) and presented a guide to how best to use and interpret Gene Ontology data at the Automated Function Prediction SIG at the International Conference on Intelligent Systems for Molecular Biology (ISMB).

Claire O'Donovan

Protein Function Content
BSc (Hons) in Biochemistry, University College
Cork, 1992. Diploma in Computer Science
University College Cork, 1993.

At EMBL since 1993, at EMBL-EBI since 1994.
Team Leader since 2009.

Future plans

In 2016 we will continue work on a 'gold-standard' dataset across the taxonomic range, with a particular focus on the UniProt proteomes set to fully address the requirements of the biochemical community. We will also continue to expand and refine our Ensembl and Genome Reference Consortium collaborations to ensure that UniProtKB provides the most appropriate gene-centric view of the protein space, allowing a cleaner and more logical mapping of gene and genomic resources to UniProtKB. We will continue to co-operate with diverse data providers (e.g., Ensembl, RefSeq, PRIDE) to integrate relevant genome and proteome information, and will import variation information from COSMIC. We also plan to extend our nomenclature collaborations to include higher-level organisms.

We will prioritise the extraction of experimental data from the literature and extend our use of data-mining methods to identify scientific literature of particular interest with regard to our annotation priorities. We are committed to expanding UniRule by extending the number and range of rules with additional curator resources, both internal and external, and providing these rules to external collaborators for use in their systems.

In 2016 we also plan to extend the scope of GO annotation to encompass entities other than proteins, in particular RNA and protein complexes.

Selected publications

Alam-Faruque Y, Hill DP, Dimmer EC, et al. (2014) Representing kidney development using the gene ontology. *PLoS One* 9:e99864

Alpi E, Griss J, da Silva AW, et al. (2014) Analysis of the tryptic search space in UniProt databases. *Proteomics* 15:48-57

Huntley RP, Sawford T, Martin MJ and O'Donovan C (2014) Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* 3:4

Huntley RP, Sawford T, Mutowo-Meullenet P, et al. (2014) The GOA database: gene ontology annotation updates for 2015. *Nucl Acids Res* 43(database issue):d1057-63

Poux S, Magrane M, Arighi CN, et al. (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database* (Oxford) 2014: bau016

UniProt, the Universal Protein Resource, is integrated with data resources spanning all of molecular biology.
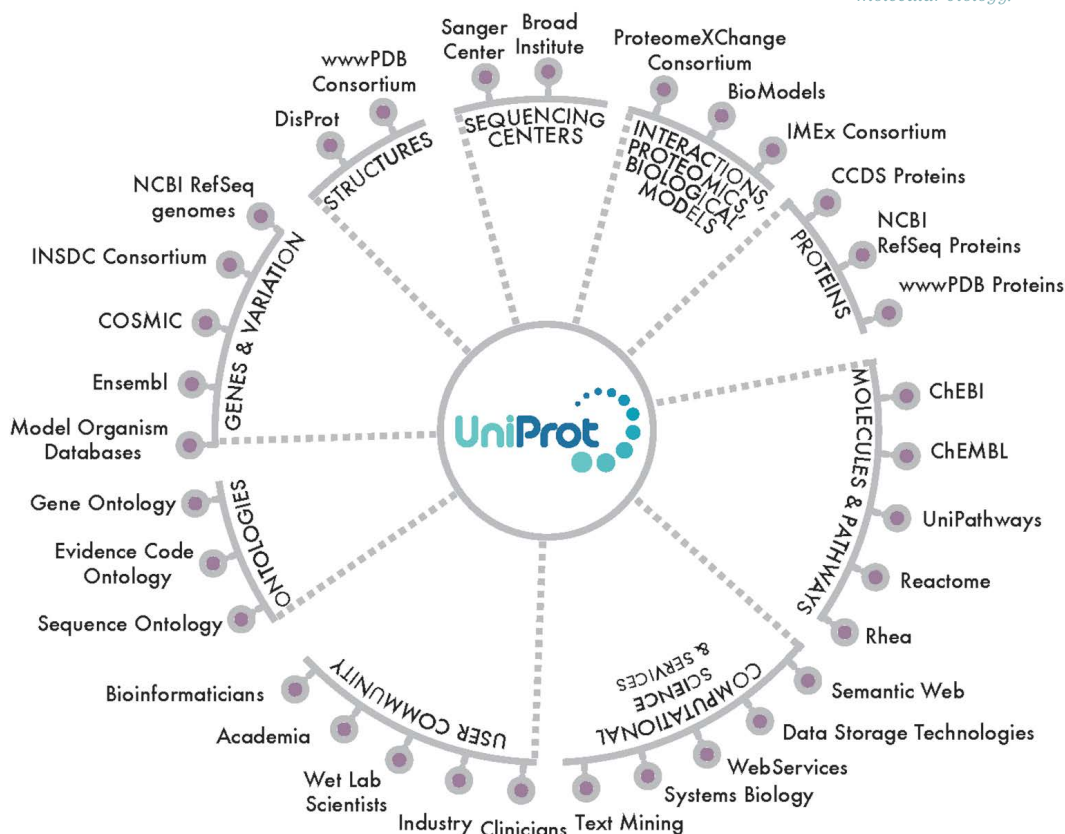
Protein Data Bank in Europe

The Protein Data Bank in Europe (PDBe) is an integrated structural data resource that aims to evolve with the science of structural biology to serve the needs of biologists.

PDBe handles the deposition and annotation of three-dimensional (3D) structural data, provides integrated, high-quality macromolecular and (sub-)cellular structures and related data and maintains in-house expertise in X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and 3D cryo-Electron Microscopy (3DEM). We provide advanced services, integrate structural and other information, and deliver ligand-related, validation and experimental data.

Our mission is to bring structure to biology, and our goal is to make PDBe the logical first stop on any quest for information about 3D molecular and cellular structure.

Major achievements

The main highlight of 2015 for PDBe was the April launch of the redesigned website (http://pdbe. org). Several years of usability research, designing, implementing, testing and incorporating user-experience feedback culminated in a new website that raises the standard for the delivery of 3D structural information. The design of the website is in line with the style of the main EMBL-EBI website. In addition, we redesigned the individual pages for PDB and EMDB entries from the bottom up, which resulted in an intuitive layout and organisation, new categories of value-added information (e.g. citation information from Europe PMC and validation information based on the wwPDB validation files), and a rich palette of informative images. We also released a new, powerful search system, which benefitted from input from our BBSRC-funded BioSolr project. To provide programmatic access to all our data, we developed an application-programming interface (API) that supports our new entry pages and can be used freely by external software developers. Further details on the redesign are described in this report by the PDBe Content and Integration team.

The international wwPDB partners have been working for several years to develop a joint, integrated software system for deposition and annotation of structural data from X-ray, neutron and electron diffraction, NMR and 3DEM. The new system (D&A 2.0), launched in January 2016, handles data to be deposited in the PDB, EMDB and BMRB archives. In addition to improved integration of data archiving for a variety of experimental techniques, appropriate validation reports have been developed for all relevant techniques. These reports are produced by validation-software pipelines, which have implemented recommendations from community-led Validation Task Forces (VTFs), convened by wwPDB. In 2015, our team put considerable efforts into the final development and testing of D&A 2.0 before its public release.

Since early 2015, PDBe has handled all D&A depositions to the PDB from European laboratories and companies, as well as all depositions that come through previous deposition systems. The PDBe annotation team also handles a substantial number of additional tasks, including D&A testing, outreach (including social media) and training. The number of PDB depositions our highly skilled annotators processed in 2015 (4007 of 10 886 total worldwide, ~37%) is approximately double the number we handled in 2013, both in absolute numbers and in share of the worldwide depositions. In the same period, they annotated 452 EMDB entries (~58% of the worldwide total). Despite this high workload, we processed over 90% of all depositions at PDBe within 48 hours.

Electron Microscopy

2015 was a very significant year for the field of 3DEM. More laboratories and national facilities gained access to state-of-the-art microscopes and detectors, which generate datasets that require processing using advanced software to optimize their utility. As a result, high-resolution structures (i.e. better than ~4Å resolution) started to be published so regularly that the change was dubbed "the resolution revolution". A record resolution of 2.2Å was set in 2015, for a map and model of β-galactosidase, both of which were deposited and annotated at PDBe (PDB entry 5a1a and EMDB entry EMD-2984). The journal Nature Methods named 3DEM its "method of the year".

PDBe is well positioned to support the 3DEM user communities during this phase of very rapid development, both as a founding member of the wwPDB (where 3DEM molecular models are archived) and as the birthplace of EMDB, where 3DEM maps and tomograms are archived (EMDB passed the milestone of 3,000 entries in 2015). Moreover, in 2014 PDBe started a new archive called EMPIAR, which serves as an archive for raw image data (mostly related to structures in EMDB). As the field expands and matures, the wide availability of such data is very useful to spur software and methods development, testing new approaches to validation, distributing data related to controversial studies, distributing data for community challenges

(such as the on-going EMDataBank Map Challenge), and training those new to 3DEM.

EMPIAR took off in 2015, with 37 released entries as of December 2015, taking up over 25 TB of disk space (the largest single entry is over 6 terabytes). We developed and put into operation a deposition and annotation system for EMPIAR data, as well as a volume-slice viewer that will become available for all EMDB entries in early 2016. Our team began work on adding support for imaging modalities other than 3DEM and electron tomography; we released the first such datasets in January 2016. In the context of the MRC/BBSRC-funded MOL2CELL project, of which EMPIAR is one component, we also organised a very productive expert workshop to discuss 3D segmentations and transformations.

Outreach

The new PDBe website offers several new useful and powerful tools, which are only useful if people know about them and use them. The launch was accompanied by communication, outreach and training activities, including: roadshows (in Newcastle, UK; Hamburg, Germany; Umeå, Sweden; Helsinki, Finland), presentations and posters at conferences (e.g. presentation and exhibition at the ECM conference in Croatia), webinars, engaging social media announcements, leaflets, newsletters, lectures in courses at universities, co-organisation of the EMBL-EBI Structural Bioinformatics training course, invited lectures in Europe and beyond, and new tutorials and modules for Train Online. We also hosted dozens of visitors and participants in workshops and meetings. In 2015, our Facebook following grew to over 4000, and our Twitter followers now number over 3300, an increase by around 50% for both in just one year. PDBe team members published six papers in peer-reviewed journals, and served as ambassadors for EMBL-EBI and PDBe.

Gerard Kleywegt

Protein Data Bank in Europe (PDBe)
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.

At EMBL-EBI since 2009.

Future plans

Feedback about the new PDBe website from the user community has been overwhelmingly positive. In 2016 we will build on this success and provide more data (especially pertaining to ligands and validation), tools (e.g., an easy-to-use viewer for X-ray electron-density maps for PDB entries, a volume-slice viewer for EMDB entries) and new search features. Following the public release of the joint wwPDB D&A system in January 2016, we will begin to implement the entire system at all wwPDB sites; the old and new systems will operate in parallel until all structure depositions can be handled via the new system. We will also work with our wwPDB partners to provide validation reports for all structures in the PDB archive (planned for public release in Q2 2016). In 2016 we will strive to reflect the rapid developments in the field of 3D bioimaging in our support of the archiving, analysis (including visualisation and validation) and dissemination of data related to 3D cellular structure at levels, from molecules, complexes, organelles and cells to small samples.

Selected publications

Gutmanas A, et al. (2015) NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.* 22:433-434

Lewis TE, et al. (2015) Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Res.* 43:D382-D3866

Sali A, et al. (2015) Outcome of the First wwPDB hybrid/integrative methods task force workshop. *Structure* 23:1156-1167

Wood C, et al. (2015) Collaborative computational project for electron cryo-microscopy. *Acta Crystallogr.* D71:123-126

EMPIAR, the archive for raw EM images data, had 37 entries as of December 2015.
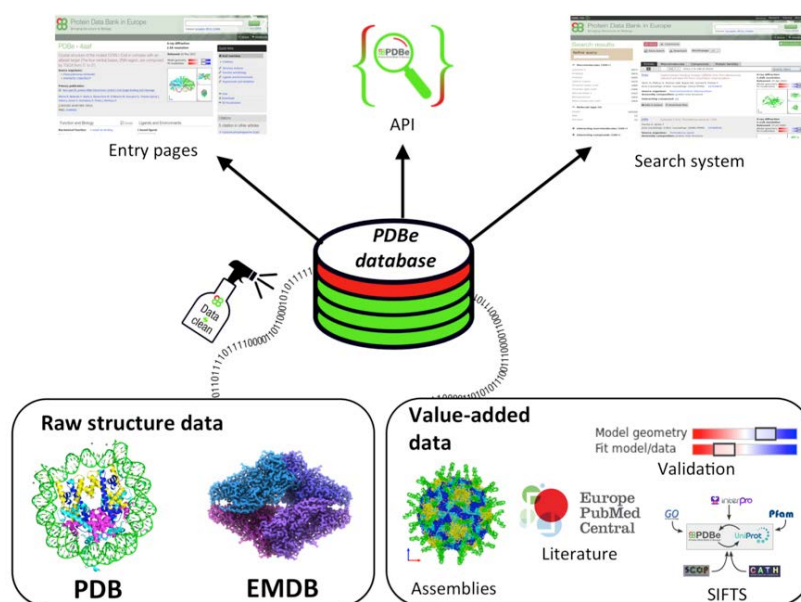
PDBe Content and Integration

PDBe aims to serve the biomedical community by providing easy access to macromolecular and cellular structure data. Our team is responsible for the curation of these data and for ensuring that the PDBe web resources serve the user community well. We also design tools to facilitate access to integrated, high-quality structural data.

Major achievements

Our major achievements in 2015 were completing the redesign of the PDBe website and annotating a record number of structures. In 2014, the wwPDB partners made the common deposition and annotation system available for X-ray crystal structures; as a consequence, the PDBe team has been handling all the European and African depositions to the archive, in addition to those made through the legacy deposition system. In addition, our curators annotate depositions to the EMDB archive, processing 4007 depositions to the PDB in 2015 (37% of 10 886 depositions worldwide) and annotating 452 EMDB depositions (58% of 780 worldwide depositions). This is more than double the PDB depositions handled by the PDBe team in 2013. We managed this increase efficiently, returning more than 90% of the depositions to the depositors within two working days.

As part of wwPDB development, we were involved in implementing the new common deposition and annotation system. This system provides, for the first time, a single interface for the deposition of data to the PDB and EMDB archives, and allows users to submit structures determined using X-ray, NMR and Electron Microscopy techniques. It was put into production in January 2016.

The completion of the PDBe redesign project was a major milestone. It addressed the need to provide long-term sustainability through a robust weekly production process and improved management of PDBe infrastructure. It also put into effect an efficient data-discovery mechanism, providing intuitive, consistent data presentation based on user-centric approaches as a major step towards fulfilling our mission of 'bringing structure to biology'.

The redesign of the PDBe infrastructure addressed data-quality issues and augmented the data with value-added annotations. It also incorporated a more robust system for the weekly release process. The PDBe database is now the central store for all information made accessible via the PDBe REST API. This API is used internally to generate the completely redesigned PDB and EMDB entry pages on our website. A powerful new search system, based on Apache Lucene Solr, provides basic data-analysis tools to identify the "best" quality structure for a given macromolecule or ligand-macromolecule complex or a particular protein-sequence family.

In 2014 and 2015, we redesigned the PDBe infrastructure to support a robust weekly production process. We reworked the database infrastructure with help from the Systems Infrastructure team, and redesigned the web infrastructure with help from the Systems Applications and Web Production teams. The new weekly production pipeline has proven to be more robust and easier to maintain than the old system.

A user survey conducted in 2012 highlighted issues related to the quality of data entries in the PDB archive, and since then we have carried out extensive remediation work to improve data quality. This gave rise to new quality-control measures that are now part of the weekly production process that we consider essential, as they improve data accessibility and discoverability. It has also resulted in a better user experience when querying the PDB archive. To enrich the PDB data with value-added information, we also worked on integrating annotations from the SIFTS resource as well as data analysis tools such as PISA for assembly information. Thanks to the availability of data-quality information based on the

Sameer Velankar

PDBe Content and Integration
PhD, Indian Institute of Science, 1997.
Postdoctoral researcher, Oxford University,
United Kingdom, 1997-2000.

At EMBL-EBI since 2000.
Team leader since 2011.

wwPDB validation pipeline, users can now identify "best quality" structures for given macromolecules based on our newly designed data-access mechanisms.

We implemented new data query systems and web pages in a user-centric approach, with user surveys and input at critical stages of the development. We initiated user involvement with a survey to establish essential requirements. The results of the survey and feedback on early prototypes informed the design of the new query system and web pages. We then developed new ways of presenting structure data to help non-expert biologists understand the structure information. A set of images providing rich insights into the quaternary structure or ligand-binding sites, or displaying sequence and structure domain annotations, now helps users with different levels of expertise understand the structure data available in the PDB. We integrated interactive web components showing annotations on sequence (1D) and structure (3D) data on the new entry pages. When user feedback suggested that general biologists struggled with these displays, we implemented an interactive topology (2D) viewer to provide a more seamless link between sequence-based data and 3D structure. The viewer displays such information on the topology diagram, which provides a schematic view of the arrangement of secondary structure elements in the tertiary structure. These diagrams are a result of a collaboration between PDBe and Dr Roman Laskowski from the Thornton research group.

The new query system provides an easy-to-use interface and basic data-analysis capabilities. It allows users to browse structure entries in the archive and offers filters for "drilling down" to a subset of entries. It also allows users to identify the most suitable structure from a given set by grouping entries based on unique macromolecules, small molecules and sequence families, displaying the "best" quality structure based on wwPDB validation data.

Our new search system offers extended functionality developed in the BioSolr project, a close collaboration between PDBe, the Samples, Phenotypes and Ontologies team and Flax, a Cambridge-based search technology company. A new plugin developed by BioSolr to integrate external programs now enables users to query PDB entries based on sequence. Following usability testing, the new feature will be released in 2016.

The team also publicly released a REST API to provide both query access and entry information for PDB and EMDB entries. The widely used 3D structure viewer Jmol/JSmol and the interactive sequence-analysis application JalView both make use of the API. It is used internally to ensure data consistency across all PDBe services, and provides access to value-added annotation from the SIFTS resource and data-quality information from the wwPDB validation data.

The wwPDB now uses the infrastructure developed in 2014 to integrate small-molecule crystal structures from the Cambridge Structural Database (CSD) into the PDB Chemical Component Dictionary (CCD). In 2015, a file containing information on 1418 chemical components found to be common between the CCD and CSD was released in the wwPDB public ftp area.

Future plans

We will continue to work on the wwPDB common deposition and annotation system to extend its functionality to include more experimental techniques. A major update, version 2.0, will be implemented at all wwPDB sites in 2016. This will allow PDBe to process all European and African depositions for both PDB and EMDB. We plan to continue improving the data quality of PDB and EMDB archive entries. We will also adapt PDBe services that provide value-added annotations and data analysis so that they can be integrated into the new infrastructure. We will implement major improvements to the PDBe query interface and entry pages, responding to user needs in terms of functionality and data access. As additional value-added information becomes available, we will extend the PDBe REST API so that the information can be accessed programmatically. To make our users aware of new developments, we will continue to participate in and organise international conferences, workshops and training events, engage on social media and publish scholarly articles.

Selected publications

Gutmanas A, et al. (2015) NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.* 22:433-434

Lewis TE, et al. (2015) Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Res.* 43:D382-D386

Meldal BH, et al. (2015) The complex portal--an encyclopaedia of macromolecular complexes. N*ucleic Acids Res.* 43:D479-D484

Sali A, et al. (2015) Outcome of the First wwPDB hybrid/integrative methods task force workshop. *Structure* 23:1156-1167

Westbrook JD, et al. (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* 31:1274-1278

ChEMBL

Drug discovery is more costly than ever, and innovation in efficacy and safety remains a significant challenge. Changes in the pharmaceutical industry over the past decade have led to an increase in drug-discovery activities in organisations that typically have access neither to large databases of legacy bioactivity data nor the experienced staff needed to manage them. Our team develops and manages ChEMBL, EMBL-EBI's database of quantitative small-molecule bioactivity data focused in the area of drug discovery; SureChEMBL, a 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.

ChEMBL contains data on curated chemical structures, bioactivity values and their relationship to biological targets and phenotypic assays. SureChEMBL combines full patent text and automatically data-mined chemical structures, significantly extending the speed and scope of public data available to drug-discovery researchers. The combination of structure–activity relationship (SAR) data from the scientific literature, deposited data from neglected disease high-throughput screens and now the patent literature all make ChEMBL an important and enabling resource for scientists working in pharmaceutical R&D.

Our research interests centre on data mining the ChEMBL database for applications relevant to translational drug discovery, including aspects of genetic variability, drug safety and neglected diseases.

Major achievements

In 2015 there was a major change in the ChEMBL Group when John Overington, who had been Team Leader since the database was taken on by EMBL-EBI in 2008, left to join the London-based biotech company Stratified Medical. Since April 2015 Anne Hersey has been Acting Team Leader.

We continued to expand the data coverage of ChEMBL to include drug-metabolism and pharmacokinetic (DMPK) data, and undertook extensive target and disease annotations on approved drugs and clinical candidates. We also developed methods to enhance and streamline the curation of data and significantly updated our Web Services as a flexible way for users to access ChEMBL data.

We further refined the SureChEMBL patent annotation pipeline to improve its robustness and provided and provided new methods to access the annotations.

The number of databases indexed in UniChem has increased to 27. We put in place a process to update the resource automatically every week.

During 2015 ChEMBL data content continued to expand, with the number of compounds reaching 1.7 million and the number of bioactivities nearly 14 million. Access to the full ChEMBL data continues to be freely available in a wide variety of technical formats including a web interface, data downloads, web services and Semantic Web technologies. During the year there were approximately 15,000 unique visitors per month to the web interface on average. There were substantial increases in the extraction of data from the scientific literature; in particular, we extracted data on drug metabolism and disposition and integrated it into the database. ChEMBL Web Services were significantly expanded and re-implemented to expose more data types and provide new functionality. In addition, we added cheminformatics Web Services based on RDkit that allow users to perform more complex queries and to combine data and chemistry-aware queries.

UniChem grew to contain links to over 100 million chemical structures from 27 source databases. For example, the UniChem web services are used on the ChEMBL web interface to provide dynamic links to other resources via the matching of the InChI/InChI Key. We fully automated the mechanism of updating and registering compounds in UniChem and since the start of 2015 weekly updates have been provided via the web interface, web services and as downloadable files.

At the end of 2015 the number of novel chemical entities annotated in SureChEMBL stood at approximately 17 million, growing at a rate of around 80,000 novel chemicals per month from roughly 50,000 new patents. Previously, the patent data in SureChEMBL was available only via a web interface. In 2015, in response to user demand, we increased options for users to access the data. We now provide a quarterly download of files

containing the mapping between compound structures and the patents they appear in. We also developed a data client feed that enables users to maintain a regular stream of the patent data behind a firewall and integrate the data with their in-house data. Using SciBite's Termite software, we annotated the patent corpus with dictionaries and ontologies for biological terms including gene and diseases. The mapping is available through the OpenPHACTS API.

In 2015 we intensified our work on the annotation of marketed drugs and clinical candidates with their intended therapeutic targets and diseases. This provides a rich source of data for researchers interested in validating therapeutic targets and identifying novel ones. This was carried out as part of the NIH-funded Illuminating the Druggable Genome (IDG) project and n Open Targets (formerly CTTV) project.

We continued to participate in two EU-funded projects, eTOX and HeCaTos, which aim to better identify and curate toxicity data and apply it to the prediction of toxicological endpoints. We continued our work on OpenPHACTS, an Innovative Medicines Initiative project that integrates pharmacological data across diverse resources. We also participated in the IDG project, Open Targets and Corbel, the European infrastructure project that follows on from BioMedBridges.

Research

We developed methods to predict molecular targets and off-targets using structural information from phenotypic screening data, an essential step in lead optimisation, polypharmacology and the study of side effects. Through collaborations, we validated three Mtb targets from our predictions for potential tuberculosis drug leads using biochemical and biophysical methods, and from the generation of target-ligand structures. In addition to developing a general theoretical model of drug resistance and drug combinations, we performed extensive data mining to retrieve information on the relationship between the type of target and the physicochemical properties of antibiotics – information that is important for the development of new drugs.

John Overington

Chemogenomics
BSc Chemistry, Bath. PhD in Crystallography, Birkbeck College, London, 1991. Postdoctoral research, ICRF, 1990-1992. Pfizer 1992-2000. Inpharmatica 2000-2008.

At EMBL-EBI from 2008 to 2015.

Anne Hersey

Acting Team Leader, ChEMBL
BSc Chemistry, University of Kent, PhD in Physical Chemistry, University of Kent, 1982, GlaxoSmithKline and former companies1982-2009.

At EMBL-EBI since 2009.

Future projects and goals

In 2016 we will continue to broaden the utility and content of ChEMBL and SureChEMBL by adding additional annotation, for example on diseases, targets and data measured using genetic variants of proteins. We will expand our use of ontologies to increase indexing of the ChEMBL data, particularly for complex and high-value endpoints such as ADMET, and in vivo pharmacology assays. We will develop technologies that enable us to build curation and data submission interfaces in a flexible and extendable way, and use text-mining methodologies to identify journal articles that enhance our coverage of chemical space. We will continue to develop automation methods for ChEMBL to enable the database to be updated more regularly and simply. We will also develop a sub-structure and similarity search facility for UniChem.

Selected publications

Davies M, et al. (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res*. 43: W612-W620

Gaulton A, et al. (2015) A large-scale crop protection bioassay data set. *Scientific Data* 2:150032

Mugumbate G, et al. (2015) Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. *PLoS One* 10:e0121492

Papadatos G, et al. (2016) SureChEMBL: a large-scale, chemically annotated patent document database *Nucleic Acids Res*. 44:D1220-D1228

Papadatos G, (2015) Activity, assay and target data curation and quality in the ChEMBL database. *J. Computer-Aided Mol. Design* 29:885-896

Cheminformatics and Metabolism

Our team works on methods to decipher, organise and disseminate information about the metabolism of organisms. We develop and maintain MetaboLights, a metabolomics reference database and archive, and ChEBI, the database and ontology of chemical entities of biological interest.

We develop algorithms to: process chemical information, predict metabolomes based on genomic and other information, determine the structure of metabolites by stochastic screening of large candidate spaces, and enable the identification of molecules with desired properties. This requires algorithms based on machine learning and other statistical methods for the prediction of spectroscopic and other physicochemical properties for compounds represented in chemical graphs.

Our research is dedicated to the elucidation of metabolomes, Computer-Assisted Structure Elucidation (CASE), the reconstruction of metabolic networks, biomedical and biochemical ontologies and algorithm development in cheminformatics and bioinformatics. The chemical diversity of the metabolome and a lack of accepted reporting standards currently make analysis challenging and time-consuming. Part of our research comprises the development and implementation of methods to analyse spectroscopic data in metabolomics.

Major achievements

ChEBI

We added over 5 200 fully curated entries to the ChEBI database during 2015 (total number of fully curated entries, over 47 500). Around 30% of new entries are from direct data submissions, and the remaining entries are from a variety of sources. Our manual curation concentrated on natural products, including compounds identified in submissions to the MetaboLights database, together with requests from individual users and research groups. We also put considerable curation efforts into the Species table, introduced to store taxonomy and citation information for natural products, which now contains over 11 000 entries.

On the ChEBI public website, we enhanced main entity pages to allow users to view information about an entity in a wider biological context. All biological reactions and pathways in which the entity participates displayed using Rhea and Reactome, respectively. ChEBI became entirely open source (again) in 2015 when we replaced the Marvin chemistry sketchpad (used for drawing chemical structures when performing exact structure, substructure, and similarity searches) with Ketcher.

From its first release in July 2004, ChEBI used SourceForge as an issue tracker for logging and tracking requests for new entries, updates to existing entries, new feature requests, bugs and other queries. During 2015, several press reports emerged regarding problems with SourceForge, which gave rise to worries about its future stability. In response to user requests, we migrated the ChEBI issue tracker to GitHub as it offers advanced features and flexibility.

MetaboLights

We developed new features to help researchers explore the reference compounds, studies and associated organism information in MetaboLights. These included species search functionality based on model organisms, free-text search and a browsable 'tree of life' with automatic classification compatible with any taxonomic identifier present in 89 different taxonomy sources. We linked over 18 000 compounds in MetaboLights to ChEBI, enriching compounds with 5355 spectra for Mass Spectrometry and 2890 for NMR. MetaboLights exported just over 150 datasets (265 total as of December 2015) to MetabolomeXchange, the global EBI Search and the Omics Discovery Index.

In 2015 we completely redesigned MetaboLights to accommodate a more flexible submission system, which now supports incremental submissions. The status of a study in our curation process is now clearly visible. We also designed a new online validation system that displays a detailed status of our current minimum reporting requirements. We incorporated new technology frameworks, for example AngularJS, and built new Web Services. We worked on integrating online analysis tools, for example adapting MetaboAnalyst (David Wishart group, University of Alberta). We incorporated tools into our architecture using R packages and the EMBL-EBI R-cloud. In tandem, we developed an authenticated Web Service that enables improved visualisations and response time, especially for larger datasets.

Community standards

The COSMOS consortium, an EU-funded endeavour for the coordination of standards for metabolomics, concluded its work in September 2015. The consortium, led by our team, combined the efforts

of 14 leading European labs in Metabolomics, and in 2015 successfully delivered new standards including the NMR markup language (NMR-ML) open standard for Nuclear Magnetic Resonance data. COSMOS also agreed procedures for the management and dissemination of data in metabolomics. The success of the project led to the formation of the MetabolomeXchange consortium, which continues the honoured tradition of global data exchange in biology. The MetabolomeXchange platform made more than 350 datasets publicly available in 2015.

PhenoMeNal e-infrastructure for metabolic computational workflows

A large volume of medical molecular phenotyping and genotyping data will be generated by metabolomics applications now entering research and the clinic. In 2015 the PhenoMeNal project was funded with €8 million by the European Commission's Horizon 2020 Programme to develop and deploy an integrated, secure, permanent, on-demand, service-driven, privacy-compliant, sustainable e-infrastructure for the processing, analysis, and information-mining of such data. This e-infrastructure will support the data processing and analysis pipeline for molecular phenotype data from the moment of its acquisition to high-level medical and biological conclusions and interpretations.

Outreach

We represented MetaboLights and ChEBI at 28 scientific conferences, training sessions and knowledge-exchange events in the UK and other European countries as well as China, India and the US. MetaboLights is now the recommended metabolomics repository several journals, including Nature Scientific Data, Metabolomics, PLOS and the EMBO journals.

Future plans

We select projects that are aligned with the goals of our services. Our research will focus on the development of efficient methods and algorithms for the assembly, analysis and dissemination of information on small molecules of relevance for biological systems. This includes information about primary and secondary metabolites, and also on xenobiotics and other molecules of relevance, such as epitopes. We will continue our work in related areas of ontology development, research on the computational representation of related data, inference of metabolomes from all types of available information, processing of metabolic and metabolomics information and reconstruction of metabolic networks. We will work on extending the ChEBI database to offer greater utility

Christoph Steinbeck

Cheminformatics and Metabolism
PhD Rheinische Friedrich-Wilhelm-Universität, Bonn, 1995. Postdoc at Tufts University, Boston, 1996-1997. Group leader, Max Planck Institute of Chemical Ecology, Jena, 1997-2002. Group leader, Cologne University 2002-2007. Lecturer in Cheminformatics, University of Tübingen, 2007.

At EMBL-EBI since 2008.

for metabolism and natural-products researchers, and to extend MetaboLights. We will enrich MetaboLights with curated knowledge, including reference spectra, pathways, protocols and references to a wider range of resources. We will develop new online data analysis capabilities to strengthen the position of MetaboLights as an important research tool.

PhenoMeNal will provide computational and data analysis services to the European biomedical community to improve understanding of aetiology, pathogenesis, pathways, and mechanisms of common human diseases. PhenoMeNal will create a federated, secure, high-performance e-infrastructure to handle and analyse large research datasets make their application available via national and European grid and cloud infrastructures where appropriate. By doing so, the 13 partners will address serious challenges arising from rapidly growing data volumes in molecular phenotyping. PhenoMeNal's cloud-based, virtualised compute engines will enable users to bring the compute to potentially sensitive data, rather than necessitating the movement of such data to potentially insecure environments.

Selected publications

Beisken S, et al. (2015) SpeckTackle: JavaScript charts for spectroscopy. *J. Cheminform.* 7:17

Hastings J, Jeliazkova N, Owen G, et al. (2015) eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *J. Biomed. Semantics* 6:10

Hastings J, Owen G, Dekker A, et al. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 44:D1214-D1219.

Moreno P, Beisken S, Harsha B, et al. (2015) BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics* 16:56

Salek RM, Neumann S, Schober D, et al. (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics* 11:1587-1597

Morgat A, Axelsen KB, Lombardot T, et al. (2015) Updates in Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res.* 43:d459-d464

Proteomic Services

The Proteomics Services team develops tools and resources for the representation, deposition, distribution and analysis of proteomics and systems biology data. We follow an open-source, open-data approach, and contribute to community standards, in particular the Proteomics Standards Initiative (PSI) of the international Human Proteome Organisation (HUPO) and systems biology standards (COMBINE Network).

Our team provides public databases as reference implementations for community standards: the PRIDE proteomics identifications database, the IntAct molecular interaction database, the Reactome pathway database, and the BioModels Database, a repository of computational models of biological systems. IntAct became an independent project under the leadership of Sandra Orchard in the Molecular Interactions team in April 2015.

As a result of long-term engagement with the community, journal editors and funding organisations, data deposition in our standards-compliant data resources is becoming a strongly recommended part of the publishing process. This has resulted in a rapid increase in the data content of our resources. Our curation teams ensure consistency and appropriate annotation of all data, whether from direct depositions or literature curation, to provide the community with high-quality reference datasets. We also contribute to the development of data-integration technologies, using protocols like the PSI Common Query Interface (PSICQUIC) and Semantic Web technologies, and provide stable identifiers for life science entities through Identifiers.org.

Major achievements

In 2015, the IntAct molecular interaction database team published a detailed visual analysis of the interactome of LRRK2, a complex, multidomain protein strongly implied in Parkinson's disease (Porras et al., 2015). Our study integrated data from consortium partners, large-scale experiments and targeted literature curation as well as pathway data from the Reactome database, providing open data supporting all figures as separate files for further analysis e.g. in Cytoscape. In April 2015, IntAct became an independent team led by Sandra Orchard.

The Reactome Knowledgebase (Fabregat et al., 2016) provides molecular details of signal transduction, transport, DNA replication, metabolism and other cellular processes as an ordered network of molecular transformations-an extended version of a classic metabolic map in a single, consistent data model.

Reactome functions both as an archive of biological processes and as a tool for discovering unexpected functional relationships in data, such as gene-expression pattern surveys or somatic-mutation catalogues from tumour cells. In 2015 we completely redeveloped the Reactome pathway diagram viewer to provide a faster, clearer interface and smooth zooming from the entire reaction network to view the details of individual reactions. All Reactome major components are available as web services or JavaScript-based widgets suitable for integration into third party applications; the new pathway diagram view has already been integrated into the Target Validation platform and ChEBI website.

The BioModels database (Chelliah et al. 2015), a standards-compliant resource for systems biology models, celebrated its tenth anniversary in 2015, and by the end of the year contained more than 1000 literature-based models. Our work on BioModels focused on targeted curation of disease-related models, and in 2015 we released a new disease summary page. We further developed the JUMMP software platform, a flexible infrastructure on which we will build the next version of BioModels and which will accommodate the IMI-funded Drug Disease Model Resources (DDMoRe) project's model repository. DDMoRe provides a public dissemination platform for models in the Pharmacometrics Markup Language (PharmML), which was released in 2015 (Swat et al., 2015).

The PRIDE database continued to grow rapidly as a key partner of the ProteomeXchange consortium of proteomics resources. In 2015, PRIDE processed over1500 submitted datasets (compared to around 900 datasets in 2014). PRIDE data downloads reached 200 Terabytes. To optimise support for both PRIDE data depositors and data users, we completely redeveloped the PRIDE Archive website, developed a REST-based web service and released a new version of the popular, stand-alone PRIDE Inspector tool suite (Perez-Riverol et al., 2015).

Generalising the principles of ProteomeXchange central, the central search interface for all ProteomeXchange partners, we developed the Omics Discovery Index (OmicsDI), a dataset-discovery tool providing access

to nine different data repositories across three 'omics types: proteomics, metabolomics and access-restricted studies in the European Genome-phenome Archive (EGA).

Outreach and training

In 2015 the Proteomics Services Team contributed to 18 training events, which collectively reached 510 scientists. Ten of these were face-to-face courses held in the UK, two were hosted elsewhere in Europe and six were held outside of Europe. We contributed content to one new course in Train online, participated in a webinar and presented in two Industry Programme events.

Future plans

In 2016 the Proteomics Services Team will continue to restructure, with PRIDE becoming an independent team led by Juan Antonio Vizcaino. The Reactome and BioModels teams will be more closely integrated to enhance synergy and share both expertise and user interface components.

We will build on the improvements to the Reactome code base to provide enhanced functionality both for web users and third-party software developers, with a major focus on the integration of molecular interaction data from IntAct and other sources into a pathway context. The new JUMMP platform will provide the underlying infrastructure for two modelling resources: the DDMoRe model repository for pharmacodynamic models and the new version of the BioModels database. These will provide a resource for systems biology model archiving and dissemination for multiple representation languages, such as the de facto standard SBML and the community-based COMBINE archive. In the context of the Identifiers.org and OmicsDI project, we will continue to pilot agile strategies for data discovery and integration.

Henning Hermjakob

Proteomic services
MSc Bioinformatics University of Bielefeld, Germany, 1995. Research Assistant at the German National Centre for Biotechnology (GBF), 1996.

At EMBL-EBI since 1997.

Selected publications

Porras P, Duesbury M, Fabregat A, et al. (2015) A visual review of the interactome of LRRK2: Using deep-curated molecular interaction data to represent biology. Proteomics 15:1390-404. doi: 10.1002/pmic.201400390

Fabregat A, Sidiropoulos K, Garapati P, et al. (2016) The Reactome pathway knowledgebase. Nucleic Acids Res. 44:D481-D487. doi: 10.1093/nar/gkv1351

Chelliah V, Juty N, Ajmera I, et al. (2015) BioModels: ten-year anniversary. Nucleic Acids Res. 43:D542-D548. doi: 10.1093/nar/gku1181

Swat MJ, Moodie S, Wimalaratne SM, et al. (2015) Pharmacometrics Markup Language (PharmML): Opening new perspectives for model exchange in drug development. CPT Pharmacometrics Syst Pharmacol. 4:316-319. doi: 10.1002/psp4.57

Perez-Riverol Y, Xu QW, Wang R, et al.. (2016) PRIDE Inspector Toolsuite: Moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. Mol Cell Proteomics 15:305-317. doi: 10.1074/mcp.O115.050229

Molecular Interactions

The Molecular Interactions Team provides public-domain interaction data, the use of which ranges from network analysis of large-scale datasets to obtaining a detailed understanding of specific protein-binding interfaces. The team also produces the Complex Portal, a reference resource for macromolecular complexes.

The Molecular Interactions Team, created in April 2015, was previously part of the Proteomics Services team. We have responsibility for the long-established IntAct molecular interaction database, and for the more recent Complex Portal. Our team will continue to support the molecular interaction community standards published by the HUPO Proteomics Standards Initiative, and contribute to their further development. We actively encourage the direct deposition of data by data producers as part of the publication process, as this provides the best opportunities to ensure the information is represented accurately in the database. IntAct is the major deposition database for groups generating high-throughput, two-hybrid and affinity-purified Mass Spectrometry data. We curate small-scale data from the scientific literature, with strict quality-control procedures to ensure we provide the highest quality reference datasets.

The Complex Portal has increased in content and coverage since its launch in 2014. Successful collaborations with groups such as the Saccharomyces Genome Database have led to a shared curation effort and broadened community access to domain-specific expertise.

Major achievements

A To meet the changing demands that arise from the ever-growing complexity of interaction data, we updated the standard format for molecular interaction data (PSI-MI XML3.0). The format now allows for the description of allosteric interactions, dynamic data and protein-complex data abstracted from multiple publications. It also offers improved representation of the effect of a mutant or variant on an interaction.

We are major contributors to the IMEx Consortium of interaction databases, which manages curation efforts over multiple resources. As of December 2015, IntAct database made almost 600,000 binary interaction evidences publicly available. Its database infrastructure is shared as a curation platform by 12 external collaborators.

The Complex Portal contained over 1400 manually curated complexes at the end of 2015. We launched an innovative graphical tool in the Complex Portal, which enables users to visualise complex topology and stoichiometry. This tool, originally developed by the Rapsilber group at the University of Edinburgh, was adapted by our developers to work with a new Java library (JAMI).

The Shelterin (Telosome) complex is a DNA-binding protein complex that associates with the telomeres that cap the ends of eukaryotic chromosomes and distinguishes them from sites of DNA damage thus sheltering chromosome ends from being inappropriately processed by the DNA repair machinery. Consequently it plays an essential role in maintaining telomere structure and integrity. Three subunits can interact directly either with single-stranded (POT1) or double-stranded telomeric DNA (TERF1 & TERF2).
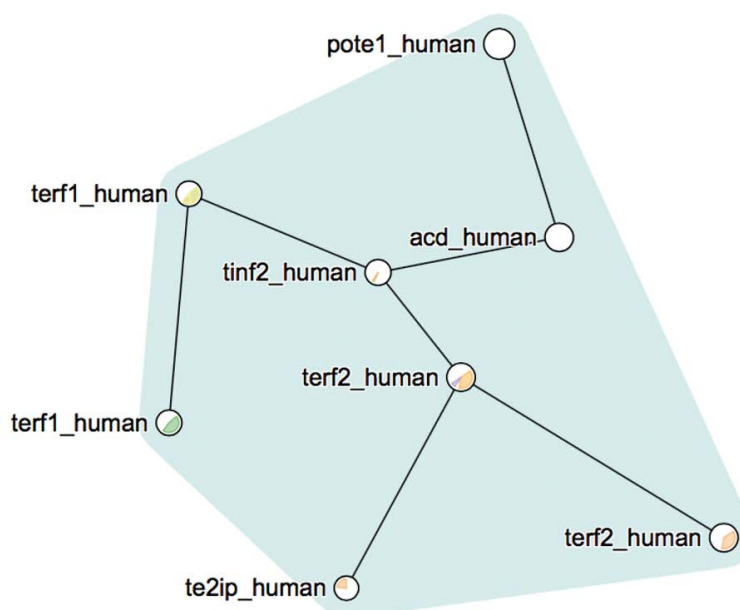
www.ebi.ac.uk/intact/complex/details/EBI-10887677

Sandra Orchard

Molecular Interactions
BSc (Hons) Biochemistry, University of Liverpool
1982. Team Leader Inflammatory Diseases Dept,
Roche Products Ltd UK, until 2001.

At EMBL-EBI since 2002.
Team Leader as of April 2015.

Outreach and training

In 2015, the Molecular Interaction team contributed to 17 training events, which collectively reached 713 scientists. Thirteen of these were courses held in the UK, one was hosted elsewhere in Europe and three were held outside of Europe. Our team also created a new online training course, participated in two webinars and presented in two Industry Programme workshops and events.
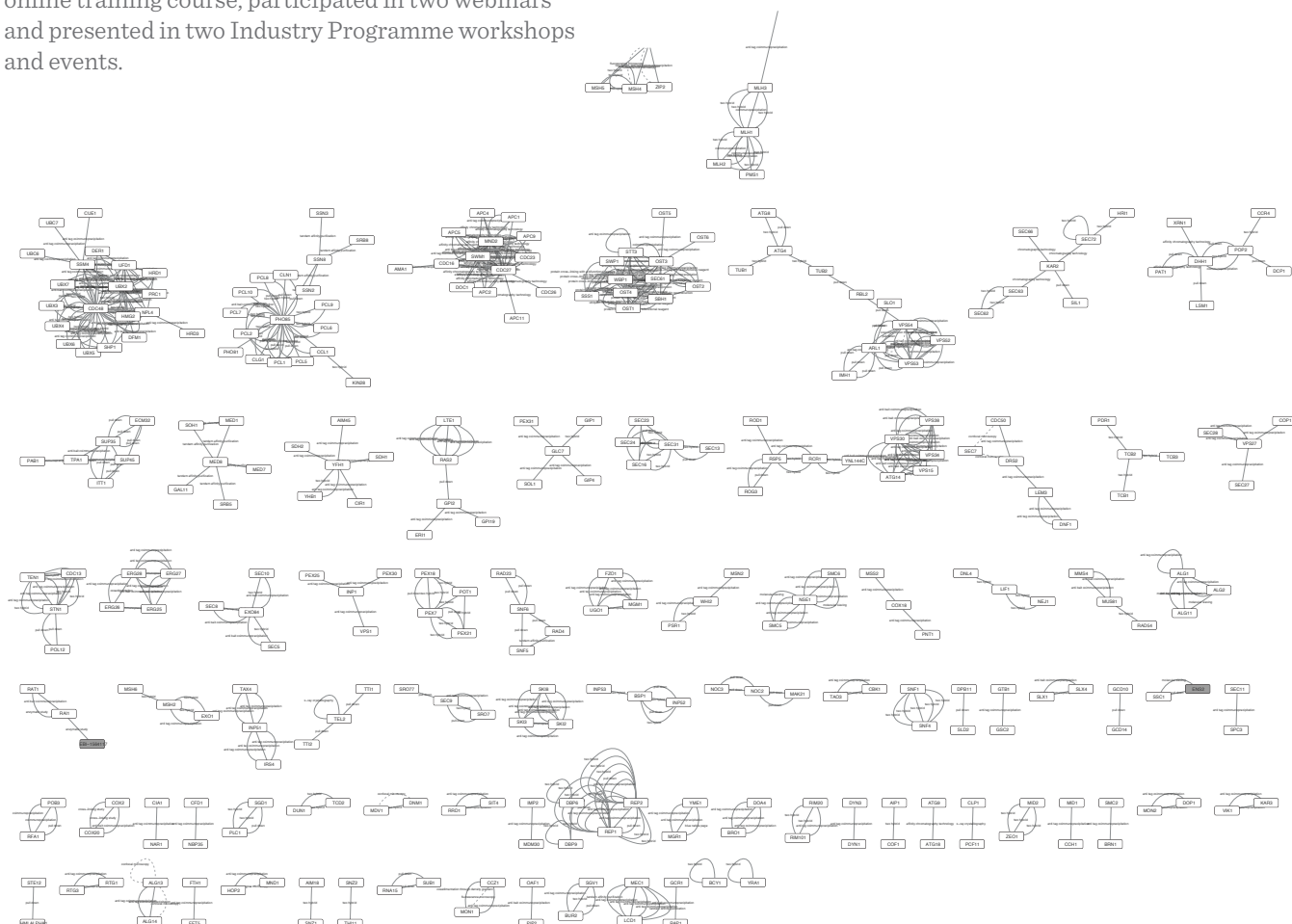
Future plans

All data resources that contribute to the IMEx databases will move to storing and maintaining their data in the IntAct database from 2016. DIP and MatrixDB data will be importing their data into the IntAct database during the year, which will result in presenting the user with a single, high-quality, merged dataset.

We will update the IntAct web pages to support new technologies, and improve data accessibility. We will develop tools, leveraging the newest data formats and standards, and incorporate these into our user interfaces to make the process of data analysis easier. We will also create a new online training module explaining the process of network analysis.

Selected publications

Porras P, Duesbury M, Fabregat A, et al. (2015) A visual review of the interactome of LRRK2: Using deep-curated molecular interaction data to represent biology. *Proteomics* 15:1390-1404

Villaveces JM, Jiménez RC, Porras P, et al. (2015) Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database* 10.1093/database/bau131

Literature Services

The EMBL-EBI Literature Services team runs Europe PMC, the database for full-text 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 critical to data-driven discovery. To achieve this, 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 the literature, and to place those 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 layers of value can be built upon the basic article content. Our team provides the infrastructure that enables individuals to enrich the 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, and as of December 2015 offered more than 30.5 million abstracts and around 3.5 million full-text research articles from PubMed and PubMed Central. It also includes metadata for Agricola (agricultural science records), biological patents and clinical guidelines. This year we added a new section to Europe PMC that contains a small collection of books and documents, extending the scope of full-text clinical guidelines in particular.

Europe PMC layers article-citation networks, links articles to underlying databases 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 via FTP (of which there are now over 1 million). Users can also search over 50 000 biomedical research grants that have been awarded to approximately 20 000 PIs supported by Europe PMC's 27 funders. These funders include the World Health Organisation, the European Research Council and many national funding agencies and charities, led by the Wellcome Trust. Europe PMC is developed by EMBL-EBI, the University of Manchester (Mimas and NaCTeM) and the British Library.

Major achievements

In 2015 use of Europe PMC continued to increase with more than 10 million unique IP addresses visiting the website during the year. Programmatic access via RESTful web services in particular also increased, serving on average 24 million requests per month in XML and JSON formats.

We focused on improving the user experience of Europe PMC, based on user research and direct feedback. We launched a new, responsive website design , which offer improved navigation and ease of use on mobile devices and smaller screens. We also introduced user accounts that enable users to log in with an ORCID, Twitter or Europe PMC account and save their favourite searches. Future developments will allow people to use their accounts to further customize their experience.

Our integration with ORCIDs – unique identifiers for researchers – continued to be central to developments on Europe PMC. We launched a new author profile page, which provides a graphical overview of an author's publications and their citation rate over time (see Figure), as well as citations over time of individual articles. The page can be downloaded for use in other documents, such as CVs. Over 200 000 unique ORCIDs are now associated with over 2.2 million articles in Europe PMC, many of whom used the Europe PMC claiming tool to add articles to their ORCID record. We improved ways to find ORCID-related information on the website: it is now possible to type in an ORCID in the main search box and find individual authors by name; in addition, the author field in the Advanced Search now includes suggestions for the author's ORCID and affiliation. All the data used to develop these features on the website are also available via web services. Via our collaboration in the EU-funded THOR project, we also began to explore how datasets may be claimed to an ORCID record.

In the area of text mining, we continued to improve access to content for text-mining groups, and our collaborations in this regard are informing the

Johanna McEntyre

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.

development of Europe PMC as a platform for application development based on full text. We contributed gene–disease associations text mined from both abstracts and open-access text articles to the Centre for Therapeutic Target Validation (Open Targets) platform, launched in December 2015. For any given gene–disease association, articles are provided ranked by relevance. We now text mine 19 accession number types, adding resources such as the Gene Ontology, European Genome-phenome Archive (EGA) and ProteomeXchange to the list. This latter work of detecting data citations is particularly important to support literature-data integration, is contributing to our understanding of the impact of data resources in research, and is a key indicator for defining ELIXIR Core Resources.

We now have 26 providers of external links from articles to websites of interest. This year we have added Publons, a small company that develops systems of credit for peer review, and Kudos, a company offering lay descriptions of research articles. The biggest supplier of links is now Wikipedia, with over 500 000 links to Wikipedia pages from articles in Europe PMC.

Future plans

In 2016 we plan to release a new system for displaying results from third-party application developers and text miners. Working with key communities across ELIXIR and beyond, we will develop the means to show text-mined annotations on open access articles on the Europe PMC website and share these results widely, for example to journals or other applications such as bibliographic reference managers. Using a

variety of user-based research approaches, we will continue to improve the experience on the website, and in particular we expect to change the article pages and search experience for the better over the course of the year. We will be begin to support an author manuscript submission system, which has previously been hosted by our Jisc collaborators. A component of this is an Awarded Grants database containing all awards from the 27 Europe PMC funders since they joined. A new, integrated grant search will be released, showing the crossover between grants and the research outputs (in the form of articles) they funded.

Selected publications

Europe PMC Consortium (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res.* 43:D1042-D1048

Kafkas Ş, Kim JH, Pi X, McEntyre JR (2015) Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. *J. Biomed. Semantics* 6:1

McEntyre J, Sarkans U, Brazma A (2015) The BioStudies database. *Mol. Syst. Biol.* 11:847

EuropePMC launched author profiles in 2015, which draws on ORCIDs to provide a graphical representation of an author's publications and citation rate over time.
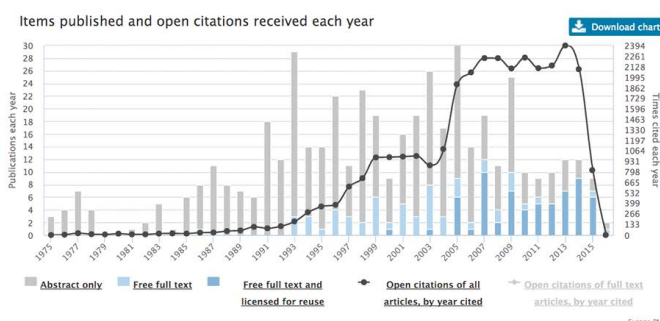http://europepmc.org

Samples, Phenotypes and Ontologies

The Samples, Phenotypes and Ontologies team has three major activity strands: BioSamples and semantic data integration, mouse informatics, and the Gene Ontology (GO) Editorial Office. We focus on metadata integration, ontology development and supporting tooling, as well as content development and delivery for the BioSamples database and mouse data for the biomedical research community.

Our team's activities diversified as the team grew in 2015, when we received new funding for two large-scale collaborative projects: ELIXIR Excelerate and CORBEL. Helen Parkinson leads informatics work packages in both these projects and co-leads the ELIXIR interoperability platform, which delivers cross-domain infrastructure for life sciences in Europe.

Major achievements

Supporting target validation with semantics and data integration

In 2015 we enhanced ontology services substantially and integrated them with industry efforts such as the Open Targets (formerly CTTV) and Roche drug discovery. We deployed the Experimental Factor Ontology in the first public release of the new Target Validation platform, and developed mapping tools and new ontology content to support its source disease annotations in EMBL-EBI databases, including the European Variation Archive, UniProt and Reactome. We also designed a phenotype–disease annotation representation that allows this public–private partnership to integrate rare and common diseases according to shared phenotype.

The BioSolr project brought Solr and ElasticSearch users together from EMBL-EBI and several other organisations, including the NCBI. Our team provided use cases, data and expertise to assist with the release of Solr and ElasticSearch expansion plugins that aim to bring ontology-enabled search to a wide range of users. The European Bank for induced pluripotent Stem Cells (EBiSC) project and may others have adopted these plugins.

Our outreach events in this domain included running a workshop at Semantic Web Applications and Tools for the Life Sciences (SWAT4LS) International in Cambridge, UK and presenting the BioSolr project at the Bioinformatics Open Source Conference in Dublin, Ireland.

The Gene Ontology now has content for apoptosis, cilia and viruses, and its representation of human intestinal parasites is much improved. An industry collaboration

with F. Hoffmann-La Roche AG to deliver improvements of an existing in-house mapping to GO led to the addition or revision of 200 terms. We also standardised protein-complex annotations based on consultation with experts from the IntAct resource..

NHGRI-EBI GWAS Catalog

In 2015 we relaunched the GWAS Catalog at EMBL-EBI based on a completely new infrastructure. We provided a new search interface, improved facilities for downloading data and a new curation platform to improve search functionality, enhance the presentation of the catalog content, improve links to other resources (especially Ensembl) and expand scientific content. The resource now provides enriched ontology-driven search capabilities, accurate display of complex interaction studies and haplotype analyses, and structured ancestry information for studies published from 2011 onwards.

Biosamples

The BioSamples database has grown to more than 4 million samples, and we have incorporated it into several projects that access biological sample data at the point of acquisition. We worked closely with developers and data providers in the EBiSC consortium to ensure that Biosamples can model induced pluripotent stem cells accurately, and that cell lines can be registered as soon as they are generated. This facilitates the submission of experimental data to the various assay databases at EMBL-EBI.

The Biosamples database has a new RESTful API to enable programmatic submission of sample data, and is more tightly integrated with the ENA to ensure that all sample data is accessioned and cross-referenced between these two resources.

In collaboration with colleagues Peter Robinson at the Institute for Medical Genetics and Human Genetics in Berlin, Germany and Tudor Groza at the Garvan Institute of Medical Research in Sydney, Australia, we performed text mining to associate diseases with phenotypes based on co-occurence in the scientific literature.

Helen Parkinson

Samples, Phenotypes and Ontologies
PhD Genetics, 1997. Research Associate in Genetics, University of Leicester 1997-2000.

At EMBL-EBI since 2000.

Mouse informatics

In 2015 we continued to support the archiving, analysis and dissemination of complex phenotype data being generated by the International Mouse Phenotyping Consortium (IMPC). We put considerable efforts into standardising high-throughput phenotyping data, which is pivotal to making the bridge between mouse biology and precision medicine. Outstanding features of the infrastructure provided by the IMPC, for example the versionable statistical analysis software package Phenstat, were highlighted in scholarly publications in 2015, underscoring their utility to the research community. The project was used as an example of how infrastructure can facilitate reproducibility and replicability of results by incorporating ARRIVE guidelines into data management.

The IMPC and its informatics platform were named as one of five case studies in a G7 report on global research infrastructures, which promoted the further development of a framework for transitioning national research infrastructures to international ones.

We also contributed further to other international collaborations in mouse biology, including INFRAFRONTIER, which coordinates the global distribution of mouse models produced in Europe, and PhenoImageshare, an online, cross-species, cross-repository tool enabling semantic discovery, browsing and complex annotations of phenotype images.

Future plans

In 2016 we will deliver a new version of the Ontology Lookup Service and develop this further to accommodate the needs of the CORBEL and EXCELERATE user communities.

The Gene Ontology Editorial Office at EMBL-EBI will continue and expand on existing work to consolidate across existing resources, improving consistency and reasoning. For example, taxon constraints integrated from the Uber Anatomy ontology can be directly used in the GO, lessening burden on editors and improving consistency for users. We will develop ontology content as well, with a focus on synaptic processes and components as well as microbial domains.

Our mouse informatics group expects to have phenotype data for over 5000 new knockout lines as the IMPC completes the first phase of their mission. They will develop new infrastructure for the second phase of phenotyping, including longitudinal analysis for ageing studies. A major publication centred on embryo phenotyping is in development.

Selected publications

Group of Senior Officials on Global Research Infrastructures, Progress Report 2015. Meeting of the G7 Science Ministers, 8-9 October 2015. Published online 14 December 2015 at https://www.bmbf.de/files/G7_Broschuere_BITV.pdf

Karp NA, Meehan TF, et al. (2015) Applying the ARRIVE guidelines to an in vivo database. PLoS Biol. 13:e1002151

Kurbatova N, Mason JC, Morgan H, et al. (2015) PhenStat: A tool kit for standardized analysis of high throughput phenotypic data. PLoS One 10:e0131274

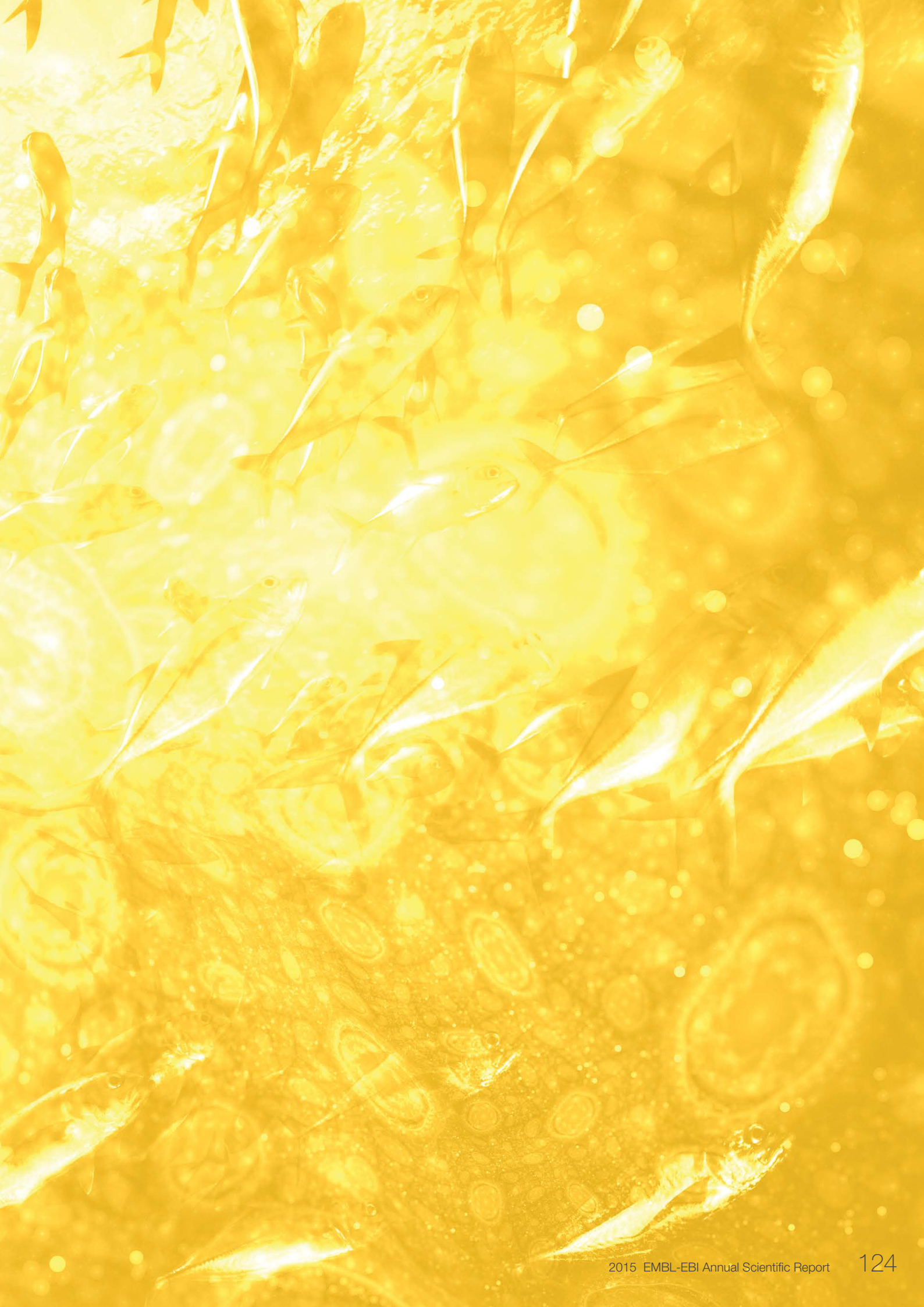Lloyd KC, Meehan T, Beaudet A, et al. (2015) Precision medicine: Look to the mice. Science 349:390

Ring N, et al. (2015) A mouse informatics platform for phenotypic and translational discovery. Mamm Genome. 26:413-421

Training Programme

Training Programme

EMBL International PhD Programme

EMBL-EBI Training Programme

EMBL-EBI provides an extensive bioinformatics training programme to ensure our users can get the most out of their own datasets by accessing public molecular data and services efficiently. Our programme is coordinated and funded centrally, and benefits from the regular input of scientific and technical experts throughout the institute.

This arrangement sets EMBL-EBI's courses apart. Our training 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

We are a partner in EMTRAIN, an Innovative Medicines Initiative project to establish a pan-European platform for professional development, covering the whole life cycle of medicines research. In 2015 we coordinated a successful workshop on continuing professional development in the biomedical sciences and continued to contribute to on-course®, EMTRAIN's comprehensive online course catalogue. Short courses continue to be the fastest growing part of the catalogue, and our team made a major contribution to this growth. Another new addition is the on-course toolkit for trainers, which provides training methods, tips and tricks for trainers in the biomedical sciences.

In 2015, EMBL-EBI staff orchestrated 164 events throughout the world and contributed to a further 98 organised by others. These included training courses at EMBL-EBI, off-site training events, conference exhibitions, career fairs and workshops, webinars and online courses. Our off-site workshops took us to Finland, Sweden, Colombia, Malaysia, Belgium, Brazil, the US and the Netherlands. The Training programme is made possible by the contributions of subject-matter experts throughout EMBL-EBI and beyond, and by the hosts of our external events, who put a huge amount of effort into ensuring that these run smoothly and meet the needs of their local trainees.

We continued to reach out to new audiences, for example clinical practitioners, who are becoming an increasingly important user group for EMBL-EBI as the need to analyse and interpret the data generated by large-scale clinical sequencing projects intensifies. In collaboration with the University of Cambridge, the Wellcome Trust Sanger Institute and Wellcome Genome Campus Advanced Courses Programme, we successfully tendered to deliver a new Master's syllabus in Genomic Medicine, tailored to the needs of current employees of the UK's National Health Service. We also contributed to the draft syllabus for a new senior role in the National Health Service: the consultant clinical scientist in clinical bioinformatics. Public consultation on the draft syllabus (see *http://bit.ly/NHSbioinformatics*) published in January 2016.

Working closely with the Web Development team and with input from many others, we redesigned the Training section of the EMBL-EBI website. Our main goal was to make it easier for our users to find all of our training, regardless of whether it is face-to-face or online. We improved the information about each course, providing clear, consistent information on learning outcomes and course prerequisites. We made extensive use of the EDAM ontology to make it easier to search courses by topic, and contributed to the extension of the EMBL-EBI content database, which now allows us to display the most relevant metadata in different contexts throughout the EMBL-EBI website. For example, centralised information about staff and visiting trainers can be displayed on different courses and events. This work fed into the broader initiative to improve the consistency and presentation of EMBL-EBI events, such that they are more attractive and may be shared easily with external partners.

Train online, EMBL-EBI's web-based training resource, is the fastest growing part of our training programme. In 2015 it had 214 470 visitors (compare with 132 890 in 2014), a figure based on unique IP addresses, which may represent any number of users at a single institute. The redesign had a positive impact on the discoverability of Train online courses: by the end of the year, between 25 000 and 30 000 unique IP addresses were accessing the resource every month. To keep pace with demand, we supported our colleagues throughout EMBL-EBI to develop or update 20 courses. We also launched a webinar series, helping our colleagues deliver 19 web-based presentations over the course of the year. These presentations were recorded and made available through Train online immediately after the live event.

Over the past three years we have developed, tested and implemented a process for supporting trainers outside EMBL-EBI to deliver high-quality training based on the EMBL-EBI model (Watson-Haigh et al., 2013). In May 2015 we launched our in-house Trainer Support Programme, basing it on similar principles. We developed a one-day workshop that introduces scientists with little training experience to EMBL-EBI's training methodology and philosophy, then pairs up new trainers with an experienced trainer, who mentors them and supports them as they develop and deliver their own training. The programme enabled us to train 19 new trainers in 2015 and was extended to include other EMBL sites and other institutes on the Wellcome Genome Campus. We continued our trainer-support collaborations Africa in a collaboration with the H3Africa initiative, and with Australia through the Australian Bioinformatics Network.

European training infrastructure

We are a partner in EMTRAIN, an Innovative Medicines Initiative project to establish a pan-European platform for professional development, covering the whole life cycle of medicines research. In 2015 we coordinated a successful workshop on continuing professional development in the biomedical sciences and continued to contribute to on-course®, EMTRAIN's comprehensive online course catalogue. Short courses continue to be the fastest growing part of the catalogue, and our team made a major contribution to this growth. Another new addition is the on-course toolkit for trainers, which provides training methods, tips and tricks for trainers in the biomedical sciences.

Future plans

Working with our training colleagues in ELIXIR and other European research infrastructures, we have secured funding to support some exciting new pan-European training projects, the following of which began in late 2015:

Cath Brooksbank

Training Programme
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.

RItrain, developing an executive curriculum in management and leadership of research infrastructure;

CORBEL, developing a curriculum for technical operators of biomedical research infrastructures, as part of a broader programme to develop shared services for life science;

ELIXIR-excelerate, developing ELIXIR's training platform, as part of ELIXIR's broader work plan to develop pan-European research infrastructure for biological data. Our focus is on a trainer support programme for ELIXIR trainers and on developing and collecting shared key performance indicators for training;

BioExcel, a pilot project to develop a new Centre of Excellence supporting academia and industry to make the most of high-end computing in biomolecular research. Our focus is on gathering training needs and developing a training programme for users of the centre.

These projects will enable EMBL-EBI to contribute towards shaping the continuing professional development of Europe's research community, in both bioinformatics and transferrable skills necessary for all successful research professionals.

Selected publications

Attwood T.K., Bongcam-Rudloff E., Brazas M.E., et al. (2015) GOBLET: the Global Organisation for Bioinformatics Learning, Education and Training. *PLoS Comput. Biol.* 11 p.e1004143 (DOI: 10.1371/journal.pcbi.1004143)

Brooksbank, C., Janko, C., Johnson, C., et al. (2015) LifeTrain: Driving lifelong learning for biomedical professionals. *J. Med. Dev. Sci.* 1, 41-47 (DOI: http://dx.doi.org/10.18063/JMDS.2015.02.001)

Brooksbank C, Johnson C. (2015) Europe: Lifelong learning for all in biomedicine. *Nature* 524, 415 (DOI: 10.1038/524415c)

Watson-Haigh NS, Shang CA, Haimel M, et al. (2013) Next-generation sequencing: a challenge to meet the increasing demand for training workshops in Australia. *Brief. Bioinform.* 14: 563-574

National School of Healthcare Science (2015) HSST Clinical Bioinformatics (Genomics) - Wider review. In: NHS Health Education England website. http://bit.ly/NHSbioinformatics

The EMBL International PhD Programme at EMBL-EBI

Students mentored in the EMBL International PhD Programme receive advanced, interdisciplinary training in molecular biology and bioinformatics. Theoretical and practical training underpin an independent, focused research project under the supervision of an EMBL-EBI faculty member and monitored by a Thesis Advisory Committee comprised of EMBL-EBI faculty, local academics and, where appropriate, industry partners.

EMBL-EBI is a University Partner Institute of the University of Cambridge, and in 2015 we benefited from the presence of 30 PhD students, welcoming five newcomers. Three students successfully defended their theses and were awarded PhDs, and one submitted his thesis. These are showcased below.

Programme events in 2015 at EMBL-EBI, through the Training Programme:

- *PhD Student Seminar Day at EMBL-EBI;*

- *Bioinformatics course for second-year EMBL PhD Programme students, organised and run by EMBL-EBI students through the Training Programme;*

- *EMBL-EBI–Sanger–Cambridge PhD Symposium (eSCAMPS)*

- *EMBL PhD Symposium at EMBL Heidelberg;*

- *Basic Teaching Module at EMBL Heidelberg;*

- *PhD student-led Lunchtime Seminar series;*

- *Statistics training course, jointly held with the Wellcome Trust Sanger Institute at EMBL-EBI.*

Selected theses

Ewan Johnstone

An evaluation of cancer subtypes and glioma stem cell characterisation

Supervisor: Paul Bertone

Ewan applied a novel co-expression analysis method to identify independent tumour cell subtypes within independent co-expression modules. These modules relate to intuitive biological features, enabling the user to identify module-specific variation independently of a transcriptome-wide subtype. Ewan used this methodology to evaluate established subtypes in breast ductal carcinoma and glioma. In breast carcinoma the basal, luminal, Her2-enriched and claudin-low cancer subtypes were replicated, revealing functional expression differences that denote each type. In glioma, dominant expression variation presented a grade-associated axis of proneural to mesenchymal expression. This axis was also presented within individual tumours, suggesting that one should not assume individual tumours are classified as discrete subtypes. Ewan's analysis of glioma-derived stem cell lines revealed distinct proneural and mesenchymal clusters in both gene expression and chromatin accessibility. Proneural signature genes suggested these lines are similar to normal glial progenitor cells, while mesenchymal expression largely relates in inflammatory and immune responses. Ewan was able to analyse epigenetic control of subtype–signature transcriptional networks thanks to differential chromatin accessibility of signatures genes. He also compared glioma-derived stem cells and neural stem cells to identify glioma-specific features. His thsis describes novel methods for ATAC-seq analysis are also described for the examination of chromatin accessibility. These findings will further assist the translation of tumour subtypes to the clinic alongside deeper characterisation glioma's persistent and heterogeneous cancer stem cell population.

Nils Koelling

Quantitative genetics of gene expression during fruit fly development

Supervisor: Ewan Birney

GWAS identify genetic variants associated with different phenotypes by exploiting naturally occurring genetic variation in large cohorts of individuals. The approach has also been applied to RNA-seq data to find loci associated with different levels of gene expression, called expression quantitative trait loci (eQTL). Most eQTL studies have been carried out in humans, but they are now feasible in higher model organisms, for example to investigate embryonic development. Nils and his colleagues performed a high-resolution eQTL study on 80 inbred fruit fly lines from the *Drosophila* Genetic Reference Panel, which represents naturally occurring genetic variation in a wild population of *Drosophila melanogaster*. Using a 3'Tag RNA-sequencing protocol they estimated the level of expression both of genes as well as of different 3' isoforms of the same gene. They estimated these expression levels for each line at three different stages of embryonic development, which improved their understanding of *D. melanogaster* gene regulation in general, and helped them investigate how gene regulation changes during development. Nils's thesis describes the processing of 3' Tag-Seq data into both 3' isoform and overall gene expression levels. He describes the design, challenges and results of performing a multivariate eQTL study in a higher model organism and provides new insights into gene regulation in *D. melanogaster* during embryonic development.

Michael Menden

In silico models of drug response in cancer cell lines based on various molecular descriptors

Supervisor: Julio Saez-Rodriguez

Michael's work focused on predicting drug responses in cancer cell lines from various molecular descriptors. Although cell lines are not perfect reflections of in vivo tumours, advances in cancer treatment have been achieved through cell-line models. To identify potential biomarkers of drug response, Michael analysed the two largest pan-cancer high-throughput screens available: the Genomics of Drug Sensitivity in Cancer project and the Cancer Cell Line Encyclopaedia. His project predicted the drug responses of different cell lines and analysed driving factors for drug sensitivities based on genomic, epigenomic and transcriptional features separately and in combination. Michael built models based on few features: just enough for a good prediction that will be useful in models for clinical medicine. He then explored the potential to increase predictive power by adding chemistry. Michael showed that combining genomic features of the cell lines and chemical properties of the compounds improved predictive power over using genomics alone. This model reproduced known biomarkers and showed superior predictability over models based solely on genomics. For leveraging germline mutations as sensitivity markers, he explored interactions with somatic mutations. Michael developed and applied methods to predict drug responses and studied biomarkers of drug sensitivity based on various descriptors. He furthermore analysed lead drug candidates regarding their mode-of-action and hypothesised future clinical applications.

Konrad Rudolph

Investigating the link between tRNA and mRNA abundance in mammals

Supervisor: John Marioni

Konrad's thesis explores the codon–anticodon interface in mRNA-to-protein translation, and how its stability is maintained during the life of the cell. His research into the control of the abundance of tRNAs by individual tRNA gene expression shows how this expression is subject to tight regulation, and that the abundance of tRNA molecules is thus kept highly stable even across vastly different cellular conditions. His work explores whether a cell benefits from expressing different sets of tRNAs, trading lower overall efficiency for high efficiency in translating the most important subset of genes. It examines the link between mRNA expression and tRNA abundance in a variety of biological conditions across several mammalian species, and establishes that changes in the pool of tRNAs are not correlated with changes in mRNA expression. Konrad's thesis provides insights into the interface between transcription and translation, suggesting that the regulation of translation is weaker than that of transcription in mammals.

Research Groups

Beltrao Group

Evolution of Cellular Networks
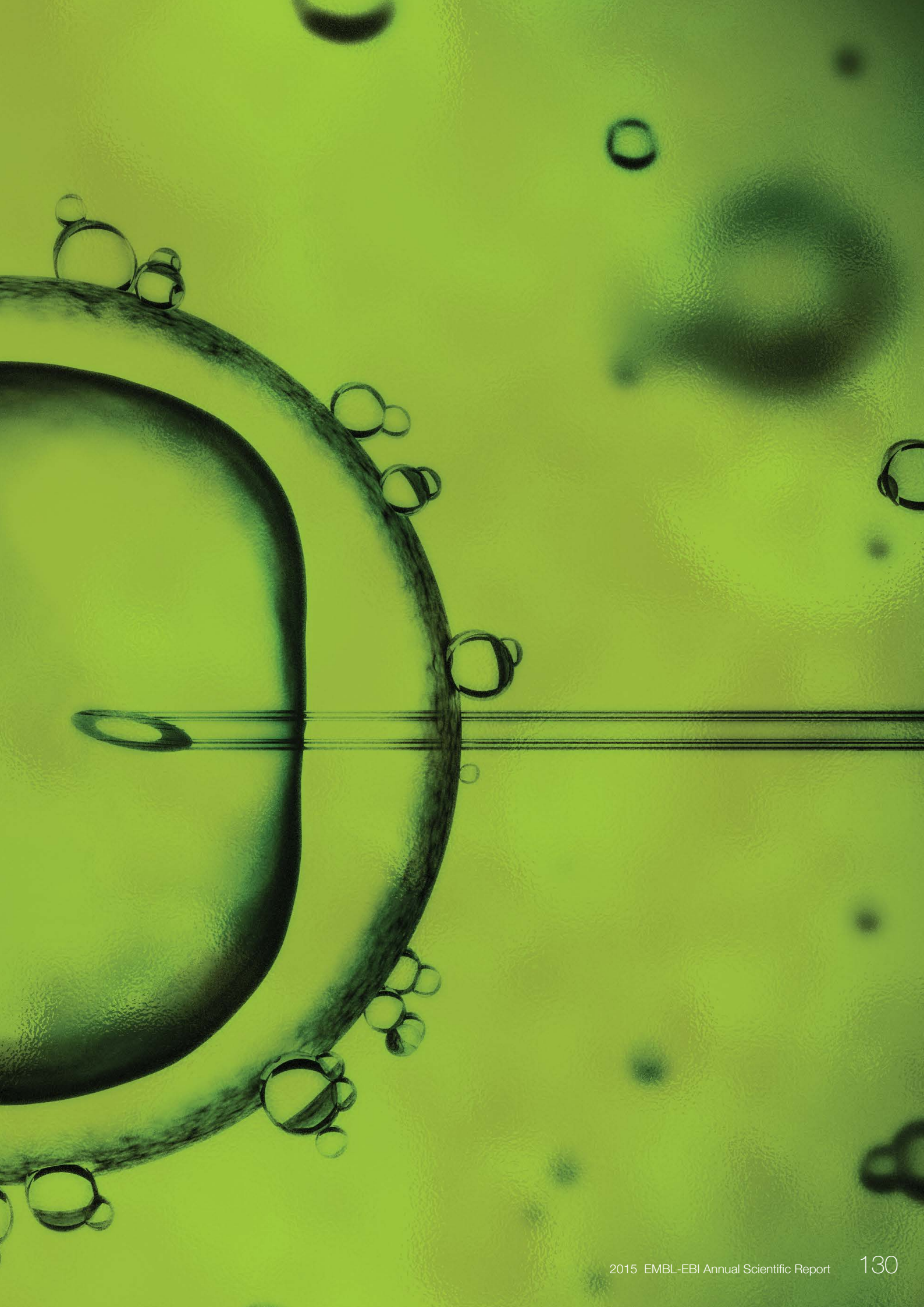
Our group is interested in understanding how novel cellular functions arise and diverge during evolution. We study the molecular sources of phenotypic novelties, exploring how genetic variability introduced at the DNA level is propagated through protein structures and interaction networks to give rise to phenotypic variability.

Within the broad scope of this evolutionary problem, we focus on two areas: the function and evolution of post-translational regulatory networks, and the evolution of genetic and chemical-genetic interactions. Looking beyond evolutionary process, we also seek to understand the genomic differences between individuals and improve our capacity to devise therapeutic strategies.

In collaboration with mass-spectrometry groups, we develop a resource of experimentally derived, post-translational modifications (PTMs) for different species in order to study the evolutionary dynamics and functional importance of post-translational regulatory networks. We use these data to create novel computational methods to predict PTM function and regulatory interactions. Our goal is to gain insights into the relationship between genetic variation and changes in PTM interactions and function.

Changes in cellular interaction networks underpin variation in cellular responses and sensitivity to environmental perturbations or small molecules. As we model and study the evolution of cellular interaction networks, we begin to see how different individuals or species diverge in their response to drugs. Understanding this relationship will enable us to develop methods to predict how genetic changes result in specific sensitivity to drug combinations.
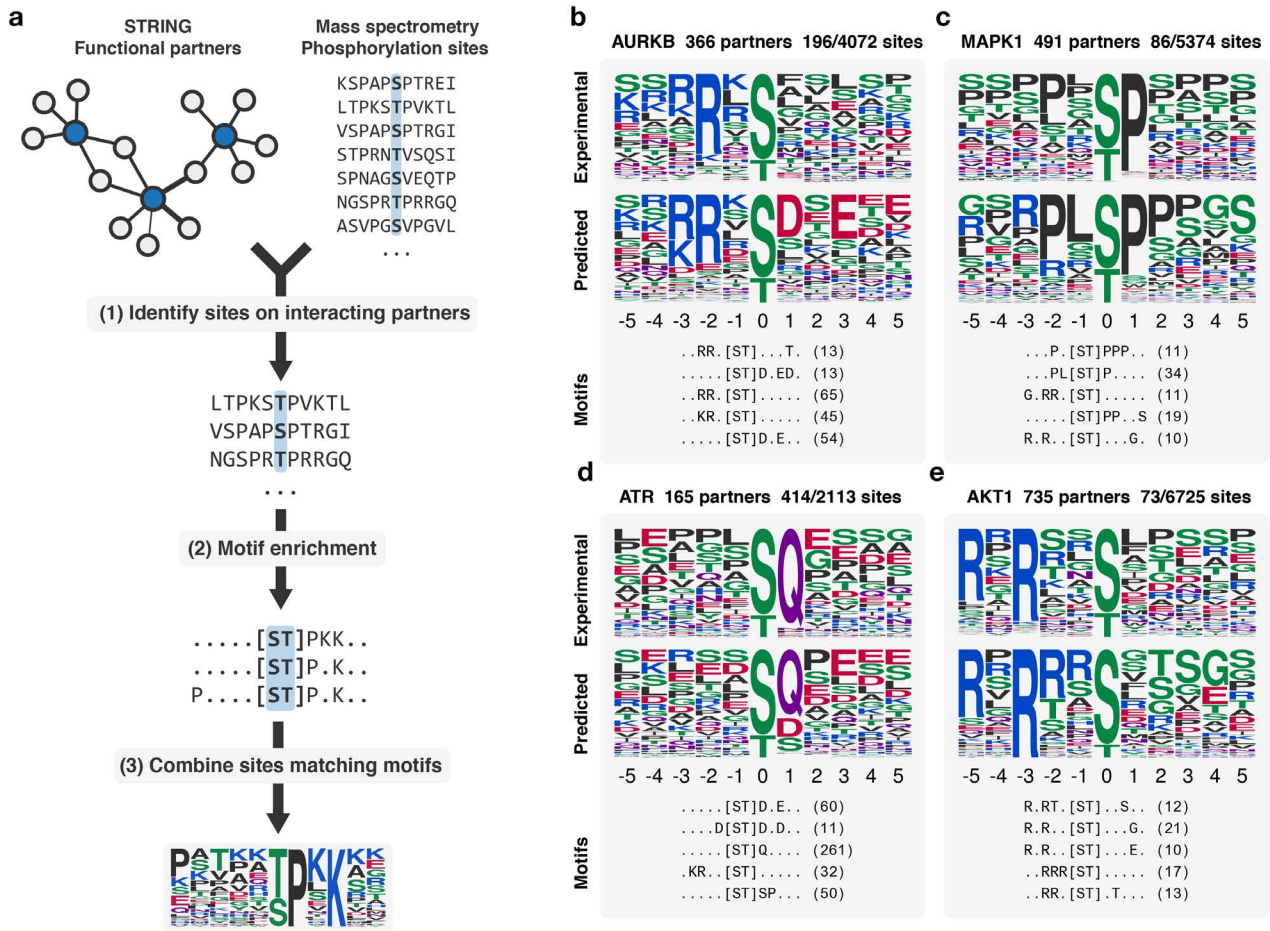
Major achievements

In 2015 we published a study on the conservation and structural properties of phosphosites in X. laevis (Johnson et al., 2015). We collected phosphoproteomics data for X. laevis and analysed the conservation of these phosphosites across 13 other species. We found that the degree of conservation of phosphosites and putative kinase–protein interactions is predictive of functionally relevant sites and interactions. We then used protein homology modelling to show that phosphosites tend to be in regions of proteins that are conformationally variable. This suggests that some of these sites might exert their function by controlling protein conformation. We believe that such studies of PTM function might, in future, lead to the engineering of PTM regulation by rational design.

Understanding how protein kinases identify their targets substrates is an open question in cell signalling. We developed a new approach that combines protein phosphorylation with interaction networks to predict what are the sequence determinants for kinase recognition. We used this method to predict the specificity of hundreds of human kinases (Wagih et al., 2016).

Changes in external conditions are sensed by the cell, which needs to trigger a response in order to adapt to the new environment. This adaptation can take many forms, such as changing protein post-translational modifications, their expression and interaction patterns. It is expected that the degree of functional association between pairs of genes will depend greatly on the environment. To study this, we measured the fitness, in different conditions, of yeast cells missing combinations of pairs of genes (Martin et al., 2015). In total we measured around 250 000 conditional gene-gene interactions. This large-scale endeavour showed that that a large fraction of genetic interactions are only observed under specific environmental conditions. Using these data, we were able to identify and validate novel condition-specific roles for several yeast genes.

Future plans

In 2016 we will continue our studies of the function and evolution of cellular networks with a strong emphasis on phospho-regulatory networks. Human cells have on the order of 500 kinases that are used by the cell to react to different stimuli and to reach decisions on what changes need to occur to cellular state. We seek to understand how these kinase-signalling networks are used, in different environmental conditions, to define specific cellular responses. To study this we have been developing approaches to study cell signalling states using comparative phosphoproteomics. We are also interested in understanding the structural properties that define protein-kinase specificity and continue to develop ways to combine different types of information to predict specificity from sequence. We aim to study the evolution of kinase specificity using these methods. We are also using some of the tools we have developed in applied research into protein kinases.

Pedro Beltrao

PhD in Biology, University of Aveiro (research conducted at EMBL-Heidelberg), 09/2007. Postdoctoral research at the University of California San Francisco.

At EMBL-EBI since 2013.

Prediction of kinase specificity from interaction networks and protein phosphorylation data. (a) Method description 1: Experimentally identified phosphosites on functional interaction partners of a kinase are collected. 2: The sites are then subject to motif enrichment to identify over-represented motifs, likely representing the kinase specificity. 3: Phosphosites matching the top k significant motifs are retained and used to construct a specificity model. (b-e) Examples of predicted specificity models. The top and middle panel of each example shows the specificity of the kinase as constructed from known substrates and as predicted by our method, respectively. The bottom panel shows the top five extracted motifs and the number of phosphosites matching them.

In collaboration with the Typas group at EMBL Heidelberg we are exploring how Salmonella cells subvert the normal functions of a human cell to their benefit. In collaboration with Oliver Billker and Jyoti Choudhary at the Wellcome Trust Sanger Institute, we are studying the phosphoregulation of the malaria parasite in the transition between the human and insect hosts.

Selected publications

Wagih O, Sugiyama N, Ishihama Y, Beltrao P (2016) Uncovering phosphorylation-based specificities through functional interaction networks. *Mol. Cell. Proteomics* 15:236-245

Johnson JR, et al. (2015) Prediction of functionally important phospho-regulatory events in *Xenopus laevis* oocytes. *PLoS Comp Biol.* 11:e1004362

Strumillo M and Beltrao P (2015) Towards the computational design of protein post-translational regulation. *Bioorg. Med. Chem.* 23:2877-2882

Martin H, et al. (2015) Differential genetic interactions of yeast stress response MAPK pathways. *Mol. Syst. Biol.* 11:800

Birney Group

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

Dr Birney's group has a long-standing interest in developing sequencing algorithms. Over the past few years they have focused on data compression, developing first theoretical then practical implementations of compression techniques. Dr Birney's 'blue skies' research includes collaborating with Dr Nick Goldman on a method to store digital data in DNA molecules. The Birney group continues to be involved in this area as new opportunities arise, for example the application of new sequencing technologies (in particular with Oxford Nanopore) and the integration of these methods with imaging techniques.

The other of strand of the group's research involves using natural DNA sequence variation to understand basic biology. Since 2010 there has been a tremendous increase in the use of genome-wide association studies to understand human diseases. This approach is very general and can be applied outside the arena of human disease. Association analysis can be applied to nearly

any measureable phenotype in a cellular or organismal system where an accessible, outbred population is available. We are pursuing association analysis for a number of molecular (e.g. RNA expression levels, chromatin levels) and basic biology traits in a number of species where favourable populations are available, including human and Drosophila.

We are also exploring molecular and physiological traits in human, for example 4D resolution of human hearts in healthy volunteers. We hope to expand this to a variety of basic biological phenotypes in other species, including establishing the first vertebrate near-isogenic wild panel in Japanese Rice Paddy fish (Medaka, Oryzias latipes).
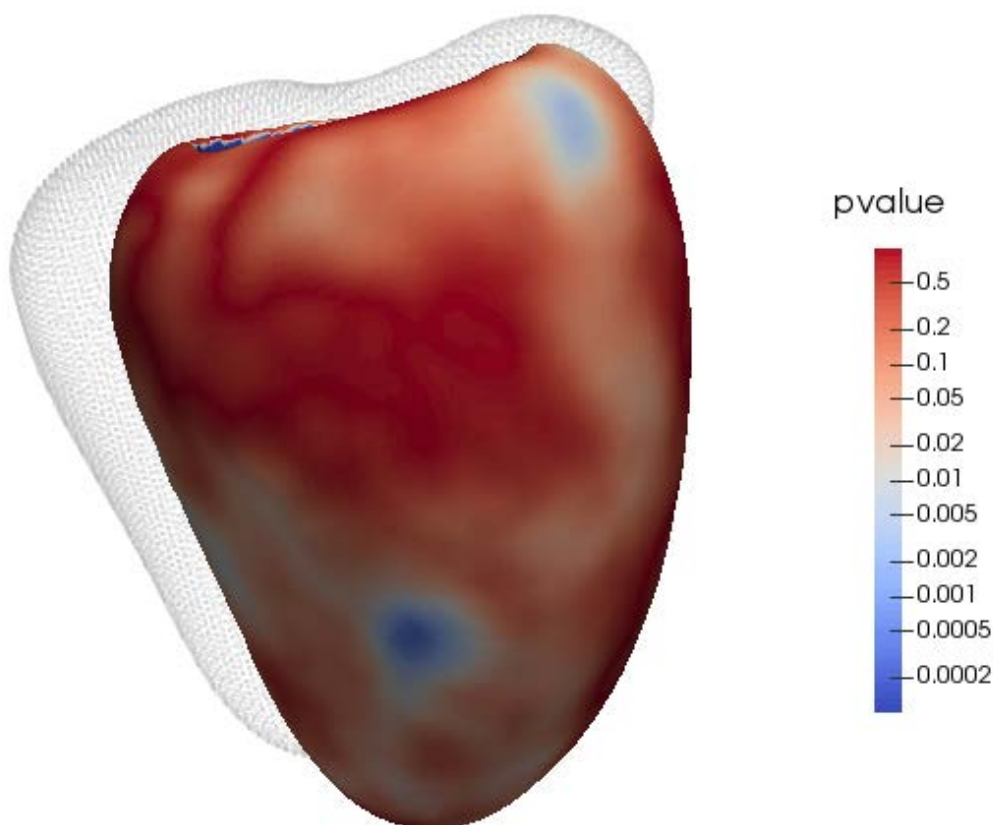
Visualisations of low dimensional projections of sources of variance in healthy human hearts on idealised 3D model of the human heart.

Ewan Birney

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

Director since June 2015.

Major achievements

In 2015 our group carried out two major types of projects: one a series of molecular studies, and the other a series phenotype-association studies. Working on a variety of human systems, we explored the association of molecular events in both normal and diseased samples. We worked closely with the Stegle research group at EMBL-EBI for both strands of research, using their new association methods that can handle both population confounders and other multiple-phenotype scenarios. Our molecular studies, in collaboration with MRI researchers and cardiologists at Imperial College, UK, focused on the structure and physiology of the human heart. Our disease-association work is in collaboration with Francis Collins at the NIH in the US, with whom we share a student.

We continued to develop the resources around Medaka fish, and demonstrated that our selected population would be appropriate for establishing a population reference panel.

We also worked on projects exploring the broader associations of molecular functional information, such as information generated by the ENCODE and Epigenome Roadmap projects, with cancer data (i.e. the BASIS project on Breast Cancers), and worked on leveraging human genetics to help improve drug discovery in the context of a project with Open Targets (formerly CTTV).

Dr Birney engaged in policy-level discussions on the use of genomic information in human health, and co-authored a paper on the risks and benefits of incidental findings in human genomics.

Future projects and goals

In 2016 the Birney group will continue to work on sequence algorithms and intra-species variation. Our work with human data will focus on molecular phenotypes in an induced pluripotent stem cell (iPSC) panel generated as part of the HipSci consortium, and on a project based on normal human cardiac data. Our work in Drosophila will investigate multi-time-point developmental biology measures. We will also assess the near isogenic panel in Japanese Rice Paddy fish for a number of molecular and whole-body phenotypes.

Figure. Region-specific changes in left ventricular wall thickness depending on SNP genotype for a locus on chromosome 6. The locus was discovered by a genome-wide association study of 9 Mio SNPs and latent factors of heart morpholgy in a cohort of 1000 healthy volunteers (Caucasians).

Selected publications

Birney E, Soranzo N. (2015) Human genomics: The end of the start for population sequencing. *Nature* 526:52-53

Huang J, Howie B, McCarthy S, et al. (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Commun.* 6:8111

Ip CL, Loose M, Tyson JR, et al. (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* 4:1075

Taylor PN, Porcu E, Chew S, et al. (2015) Whole-genome sequence-based analysis of thyroid function. *Nature Commun* 6:5681

UK10K Consortium, et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526:82-90

Enright Group

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. Our 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, piwiRNAs and long non-coding RNAs. We collaborate extensively with experimental laboratories on commissioning experiments and analysing experimental data. Some laboratory members take advantage of these close collaborations to gain hands-on experience in the wet lab.

## Major achievements

Long non-coding RNAs

Working closely with the Furlong group at EMBL Heidelberg and the O'Carroll group at the University of Edinburgh, we explored the roles of long non-coding RNAs in different systems. Through extensive, deep sequencing in mouse, we identified a catalogue of lncRNAs expressed through the murine germline. From this incredibly detailed atlas of transcription, we began to identify a set of lncRNAs whose functions may be extremely important to the maintenance of the germline and genomic integrity. With the Furlong group we began exploring the roles of lncRNAs in the development of the mesoderm in D. melanogaster embryos, and developed a range of new pipelines and techniques that greatly improved our ability to delineate real transcripts from artefacts and DNA contamination accurately. Current results are extremely encouraging, showing that we have discovered many novel lncRNAs with very specific and interesting expression patterns according to in situ hybridisation.

We performed two experiments in collaboration with the Mattick Lab at the Garvan Institute, Australia, exploring the use of a probe-based sequencing technology (i.e. CaptureSeq) in both human and mouse genomes, illustrating the tremendous complexity of the transcriptome in both species. This novel approach uses probes designed around transcript isoforms to perform extremely deep sequencing targeted at deciphering the appropriate transcript assembly for mRNAs or lncRNAs of interest.
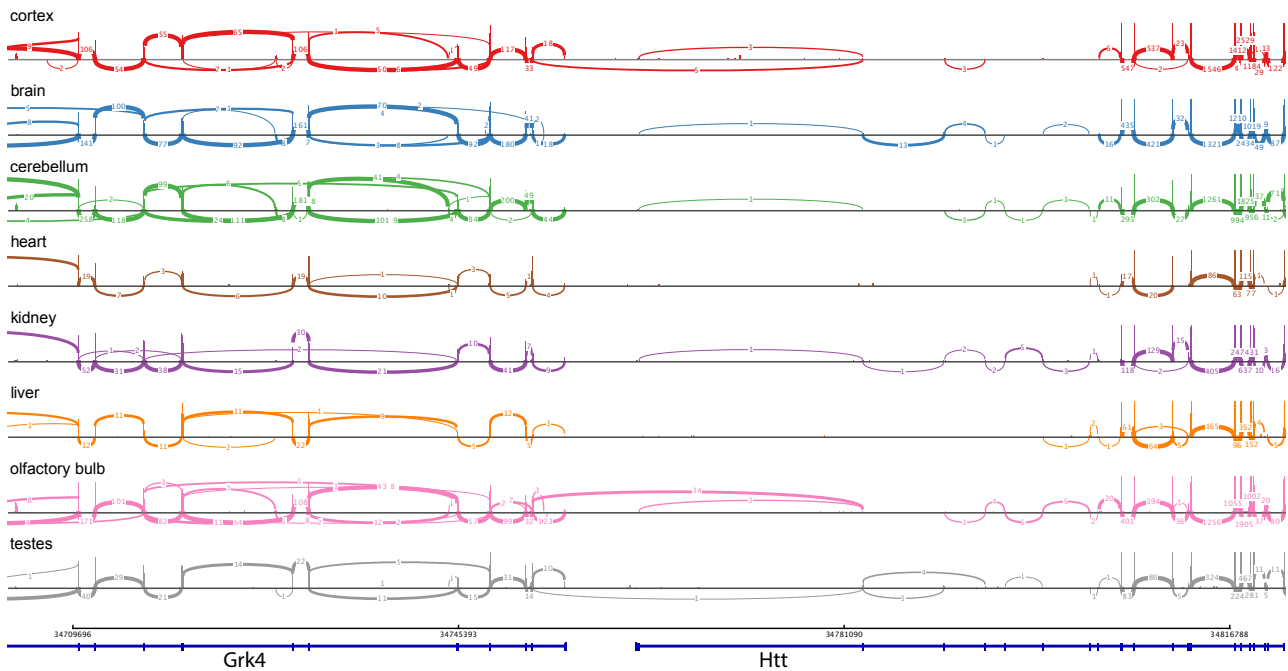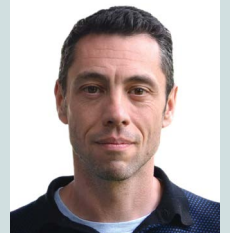
We worked with the laboratory of Tony Kouzarides at the Gurdon Institute in Cambridge to explore a class of lncRNAs which show conserved syntenic signatures between Human and Mouse. These positionally conserved RNAs appear to be essential to the function and expression of their neighbouring protein-coding genes and we are working closely with the Kouzarides lab to further isolate their specific biochemical or structural functions.

Clinical genomics

In our long-standing collaboration with the Department of Pathology at the University of Cambridge and Addenbrooke's Hospital, we previously explored the roles of microRNAs and other non-coding RNAs in paediatric germ cell tumours and solid tumours such as neuroblastoma. In 2015 we focused on the analysis of Human Papilloma Virus (HPV) integration sites in human cervical cell lines to explore how HPV integration leads to oncogenesis. This work involved large-scale RNA-seq, DNA sequencing and targeted HiC analysis to explore where in the genome HPV integrates across a clonal set of cervical cells and the effects that this integration has on gene-expression. We hope that this on-going work will shed light onto viral oncogenesis and cervical cancer.
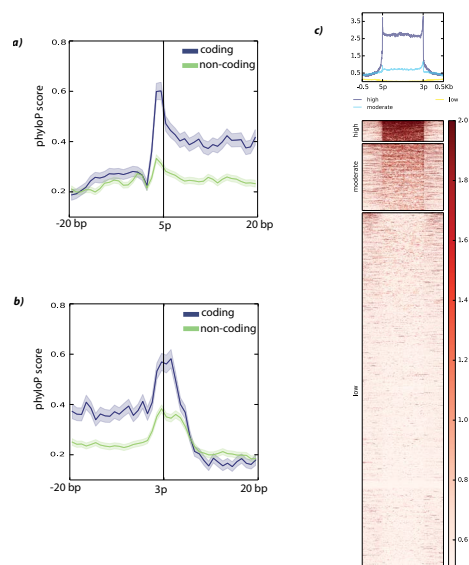
Analysis of spice site donor/acceptor conservation in coding versus non-coding sequences based on the CaptureSeq technology for extremely deep sequencing and isoform quantitation/discovery.

Anton Enright

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

At EMBL-EBI since 2008.

Different spliced isoforms detected across multiple murine tissues based on the CaptureSeq methodology.

MicroRNA evolution and function

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. We developed a new algorithm for the prediction of novel miRNAs from deep-sequencing data and applied this method to the de novo detection of miRNAs in a number of species. We worked closely with Elia Benito-Gutierrez from the Arendt group to define miRNAs in amphioxus species from around the world and identify miRNAs with distinct roles in developmental processes.

Computational methods

The epi-transcriptome is a growing area of interest and post-transcriptional modifications of a number of different types of RNA are becoming increasingly important. We developed Chimera, a recently published tool for the assessment of 3' Uridylation of miRNAs and other changes such as adenylation or ADAR editing. We began to use this system in collaboration with the O'Carroll laboratory to explore uridylation and RNA methylation in the murine male germline. Our data shows the many important epi-transcriptomic changes present in both microRNAs and mRNAs and how they respond when key proteins are conditionally knocked out.

Future plans

In 2016 we will complete our on-going long-term analysis on lncRNAs in the male germline and D. melanogaster mesoderm. We will continue to work on our project to explore the process of genomic integration of HPV and how this virus can lead to the development of cervical cancer through sequencing and clinical genomics. We also plan to release our new tool for the machine-learning-based prediction of novel microRNAs from large-scale sequencing data of both finished genomes and organisms for which a reference genome is not yet available.

Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. We are extremely interested in the evolution of regulatory RNAs and developing phylogenetic techniques appropriate for short, non-coding RNA and long non-coding RNAs. We will continue to build strong links with experimental laboratories that work on miRNAs in different systems, as this will allow us to build better datasets with which to train and validate our computational approaches. The use of visualisation techniques to assist with the interpretation and display of complex, multi-dimensional data will continue to be an important parallel aspect of our work.

Goldman Group

The diversity of all life has been shaped by its evolutionary history. Our 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 genomes but poses many new challenges.

Major achievements

In 2015 we investigated 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. We also studied the impact of different models on estimation of divergence time, as traditional substitution models tend to underestimate genetic distances between distantly related species, introducing large biases. By relaxing the assumption of site-invariant selective pressures, we demonstrated that longer distances are estimated for basal branches of species trees.

We published the results of a long-running comparative study of MSA filtering methods, which automatically identify and remove unreliable alignment sites, including data that may introduce errors into evolutionary analyses. The filtering step has been assumed to lead to better trees; however, our study showed that with most filtering methods and in most circumstances, alignment filtering worsens the resulting trees. These findings have implications for scientists working with difficult-to-align sequences.

We investigated how the presence of gaps in a MSA – for example those introduced by insertions or deletions of genetic material (indels) – affects the accuracy of inferred phylogenies. Standard phylogenetic methods do not attempt to model indels, rather treating gaps as missing data. To address the suggestion that this could lead to statistical inconsistency giving rise to an incorrect evolutionary tree, even as more and more data are collected, we derived a new, simple proof of statistical consistency of maximum likelihood phylogenetic reconstruction for un-gapped alignments. In so doing we showed that the suggested inconsistency only pertains to one, very specific, outlier case.

As most cell divisions introduce a small number of mutations into the genome, it is possible to use single-cell exome/genome sequencing data to infer somatic evolutionary histories of individual cells. These phylogenies, called 'cell lineage trees', are useful in developmental biology and cancer research but are difficult to resolve due to very low mutation rates and high sequencing error rates resulting from allelic dropout. The most commonly used algorithm does not distinguish between mutations and sequencing or sample amplification errors, so we developed a scalable algorithm based on a phylogenetic model that explicitly models mutations and sequencing errors as two separate processes. In addition to resolving a long-standing, tractable, mathematical problem, the new method gives more accurate trees than those produced using standard existing methods.

'Incongruence', increasingly common with the rise of multi-locus and whole-genome sequencing, occurs when an estimated tree varies depending on the particular set of sequences on which it was built. Incongruence can arise due to stochastic differences between loci, or to different sets of sequences having undergone different processes of evolution. To address the latter case we developed treeCl, a clustering method that groups loci that share a common evolutionary history and distinguishes sets that do not. We tested the method using both simulated data, a curated set of yeast proteins and RAD-sequenced DNA loci from globeflower flies (*Chiastocheta* spp.; Diptera, Anthomyiidae). We improved the curation of the yeast data by identifying several non-orthologous sequences,

Nick Goldman

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.

At EMBL-EBI since 2002. EMBL Senior Scientist since 2009.

and produced better-resolved evolutionary trees for the globeflower flies, gaining insights into the underlying causes of the flies' incongruent trees.

We continued to investigate structural and functional determinants of selective evolutionary constraint in mammals, focusing on the level of genes and domains. To facilitate the analysis of the mode of evolution, we developed a web service that integrates and displays structural information with selective constraints discovered using our Sitewise Likelihood Ratio (SLR) method.

We regularly share our expertise with experimental wet-lab biologists studying specific biological problems, and in such a collaboration in 2015 contributed analysis of the evolutionary dynamics of DNA regions differentially methylated between different tissues. DNA in blood has more sites that become methylated compared to other tissues but, contrary to expectation, we found that this subset of sites shows the same evolutionary patterns as those that exhibit higher fractional methylation.

Our work to re-purpose DNA as a medium for archiving digital information continues to be of great interest worldwide, and in 2015 the Biotechnology and Biological Sciences Research Council (BBSRC) supported our development of computational and laboratory DNA-handling technologies needed to bring DNA-storage closer to market. Extending state-of-the-art methods in coding theory, we began modelling the statistical properties of both the storage medium itself and the errors induced by the "DNA synthesis > processing > storage > sequencing channel", and developed algorithms to exploit the properties of this channel. We began developing a system that will allow mass storage of data on DNA with proven reliability and guaranteed efficiency at a level comparable to industry norms. Working closely with molecular biologists and DNA synthesis and sequencing specialists, we are developing solutions that will make DNA a viable choice for long-term, reliable and robust digital storage.

Future plans

We are dedicated to using mathematical modelling, statistics and computation to enable biologists to draw as much scientific value as possible from modern molecular sequence data. We will continue to improve and develop new methods for phylogenetic analysis, and new techniques to analyse incomplete datasets. We will apply of our cell lineage tree algorithm to single-cell sequencing data, and further develop the method to identify lineage divergences that are extremely difficult to pinpoint by manual analysis.

Past work in the group was some of the first in phylogenetics to be able to relate protein sequence evolution to features of the entire evolving protein, rather than assuming mutations at different locations had independent effects. We will further develop our method to model the evolutionary forces acting on proteins involved in cellular information processing, shifting focus from the 3D structure of the evolving proteins to the interactions between binding pairs of molecules in signalling networks. The method will enable us to infer how evolutionary pressures have impacted the evolution of sequences.

Clinicians are looking to genome sequencing to provide diagnostic aids and inform treatment decisions, for example in determining the correct antibiotic based on rapid determination of pathogen species and strain, or detecting mutations known to impact antibiotic resistance. We believe that state-of-the-art genomic analysis methods can assist clinicians and be further optimised to be fast and accurate. In collaboration with clinicians who have expertise in diagnostics and treatment policy, we will work on methods for informing their choices based on bacterial whole-genome sequencing. We will analyse the performance of existing methods for detecting antibiotic resistance using limited data sets, producing knowledge of value for linking the latest NGS technologies with the appropriate software for diagnostic and clinical applications.

Selected publications

Lowe R, Slodkowicz G, Goldman N, Rakyan VK (2015) The human blood DNA methylome displays a highly distinctive profile compared with other somatic tissues. *Epigenetics* 10:274–281

Schwarz RF, et al. (2015) Changes in postural syntax characterize sensory modulation and natural variation of *C. elegans* locomotion. *PLoS Comp Biol* 11:e1004322

Schwarz RF, et al. (2015) Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Medicine* 12:e1001789

Tan GM, et al. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.* 64:778–791

Truszkowski J and Goldman N (2015) Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. *Syst. Biol.* doi: 10.1093/sysbio/syv089

Marioni Group

Our research focuses on developing the computational and statistical tools necessary to exploit high-throughput genomics data, in order to understand the regulation of gene expression and model developmental and evolutionary processes.

We develop the computational and statistical tools necessary to exploit high-throughput genomics data, in order to understand the regulation of gene expression and model developmental and evolutionary processes.

Within this context, we focus on three specific areas: gene expression regulation, evolution of cell types and modelling variability in expression levels. To understand how the divergence of gene expression levels is regulated, we associate changes in expression with a specific regulatory mechanism. In so doing, we gain critical insights into speciation and differences in phenotypes between individuals. We study the evolution of cell types by using gene expression as a definition of the molecular fingerprint of individual cells. Comparing the molecular fingerprint associated with a particular tissue across species allows us to decipher whether specific cell types arise de novo during speciation, or whether they have a common evolutionary ancestor. We model spatial variability in gene-expression levels within a tissue or organism to identify heterogeneous patterns of expression within a cell type. This potentially allows us to uncover new cell types, perhaps with novel functions. We use similar approaches to study the extent of heterogeneity present throughout a tumour.

These strands of research are brought together by single-cell sequencing technologies. Studying variability in gene expression (and other genome-wide characteristics) at a single-cell level is revolutionising our ability to assay regulatory variation, molecular fingerprints and spatial patterns of expression. As founding members of the Sanger Institute – EMBL-EBI Single Cell Genomics Centre, and with group leader John Marioni as co-ordinator, we are closely involved in the centre's efforts to improve data generation and analysis methods, especially single-cell RNA-sequencing, and in using them to answer numerous exciting biological questions. We see the development of appropriate statistical and computational tools as critical to the full exploitation of these data, and will focus on these challenges over the next few years.

Major achievements

In 2015 we built on our previous work in the field of single-cell transcriptomics, collaborating with the Richardson group at the Medical Research Council (MRC) Biostatistics Unit to develop a hierarchical Bayesian approach for identifying highly variable genes (Vallejos et al., 2015). This approach builds on our earlier work (Brennecke et al., 2013) by jointly inferring normalisation and noise parameters for both technical and biological genes. We recently extended this model to identify genes that show different noise profiles between conditions. Moreover, we developed approaches for robustly identifying stochastic allele-specific expression (Kim et al., 2015), providing an important tool for dissecting

A new method for analysing RNA sequence data allows researchers to identify new subtypes of cells, creating order out of seeming chaos This novel technique represents a major step forward for single-cell genomics.
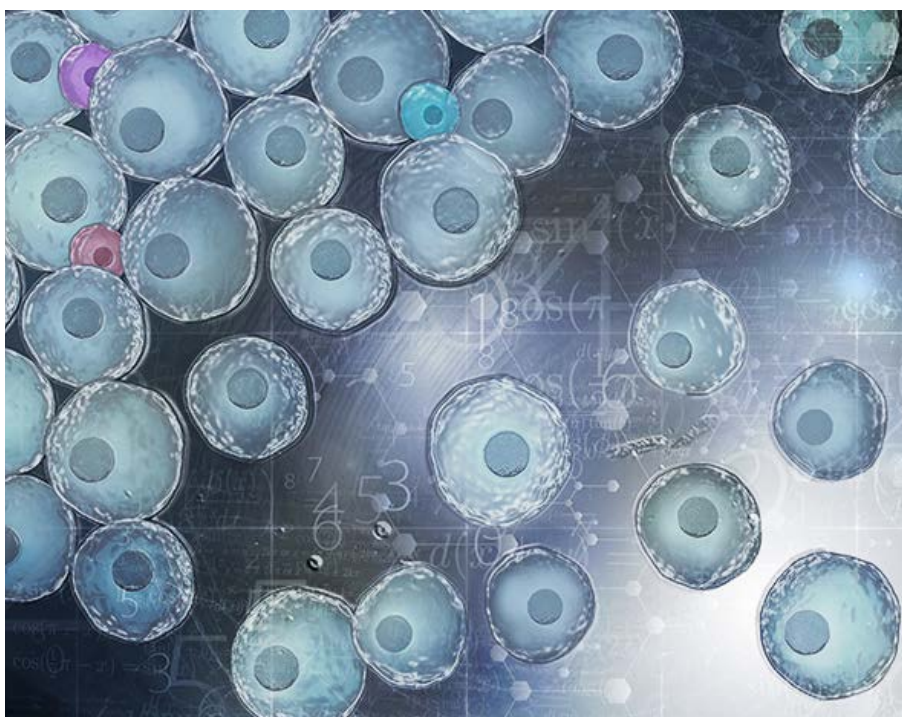
John Marioni

PhD in Applied Mathematics, University of Cambridge, 2008. Postdoctoral research in the Department of Human Genetics, University of Chicago.

At EMBL since 2010.

this source of heterogeneity in gene expression. Finally, motivated by developments in scRNA-seq technology, we are developing more appropriate methods for normalising scRNA-seq data, which are robust to the presence of 'stochastic' zeros in the data due to technical drop-outs.

We developed and validated a high-throughput method to identify the precise spatial origin of cells assayed using scRNA-seq within a tissue of interest (Achim, Pettit et al., 2015). This approach compares complete, specificity-weighted mRNA fingerprints of a cell with positional gene-expression profiles derived from a gene-expression atlas (e.g. generated via in situ hybridization experiments).

We applied our novel computational approaches to the study of heterogeneity in populations of mouse Embryonic Stem Cells in collaboration with the Teichmann group (Kolodziejczyk, Kim et al., 2015), and demonstrated that globally, gene-expression noise levels in stem cells cultured in different media show no difference, yet specific sets of genes vary systematically. Additionally, in collaboration with the Zernicka-Goetz group at the University of Cambridge, we applied scRNA-seq to study symmetry breaking in pre-implantation mammalian development. We showed that, at the four-cell stage, there exists substantial heterogeneity in gene expression levels between cells, which can bias cells towards particular cell fates in a non-deterministic fashion (Goolam et al., in press). This provides important insights into when cell fate decisions are first determined.

Our group was strengthened by the appointment of John Marioni as a Senior Group Leader at the Cancer Research UK Cambridge Institute at the University of Cambridge. This position will enable more direct interactions between members of the group and empirical researchers, especially in the area of cancer biology, an exciting future direction. In addition, the group received funding through two Wellcome Trust Strategic Awards that will enable the appointment of two fully funded three-year postdoctoral researcher positions.

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 tumourigenesis. We will actively 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. We will also generate methodology for handling dropSeq data, focusing particularly on how to model errors in the cellular and molecular barcodes, which can impede data interpretation. Our group also plans to build on spatially-resolved single-cell transcriptomic data, gained by using novel scRNA-seq analysis methods (Achim, Pettit et al., 2015), using these data to identify cell types and examine heterogeneity in expression at the spatial level. This project will require the development of new computational approaches than can cluster multiple data modalities simultaneously.

From the biological perspective, we will use our new methods to obtain insights into cell fate decisions during gastrulation – arguably the most important time in our lives. Preliminary results are extremely promising, suggesting that we can define at the highest possible resolution cell fate decisions during this key developmental period. Moreover, we will continue to apply our models in numerous biological contexts, such as the study of heterogeneity in different populations of neurons, cancer biology and non-model systems to study evolution and selection.

Selected publications

Achim K, Pettit JB, Saraiva LR, et al. (2015) Single-cell expression profiling and spatial mapping into tissue of origin. *Nature Biotechnol.* 33:503-09

Goolam M, Scialdone A, Graham SJL, Heterogeneity in Oct4 and Sox2 targets biases cell fate in four-cell mouse embryos. *Cell* (in press)

Kim JK, Kolodziejczyk AA, Tsang JCH et al. (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17(4):471-85

Kolodziejcyzk AA, Kim JK, Ilicic T et al. (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Commun.* 6:8687

Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian Analysis of single-cell sequencing data. *PLoS Comput Biol* 11:e1004333

Saez-Rodriguez Group

Our group aims to achieve a functional understanding of signalling networks and their deregulation in disease, and to apply this knowledge to the development of novel therapeutics.

Human cells are equipped with complex signalling networks that allow them to receive and process the information encoded in myriad extracellular stimuli. Understanding how these networks function is a compelling scientific challenge and has practical applications, as alteration in the functioning of cellular networks underlies the development of diseases such as cancer and diabetes. Considerable effort has been devoted to identifying proteins that can be targeted to reverse this deregulation; however, their effect is often unexpected. It is hard to assess their influence on the signalling network as a whole and thus their net effect on the behaviour of the diseased cell. Such a global understanding can only be achieved by a combination of experimental and computational analysis.

Because our research is hypothesis-driven and tailored to producing mathematical models that integrate diverse data sources, we collaborate closely with experimental groups. Our models integrate a range of data, from genomic to biochemical, with various sources of prior knowledge and with an emphasis on providing both predictive power of new experiments and insights into the functioning of the signalling network. We combine statistical methods with models describing the mechanisms of signal transduction, either as logical or physico-chemical systems. To do this, we develop tools and integrate them with existing resources. We then use these models to better understand how signalling is altered in human disease and predict effective therapeutic targets.

Major achievements

In 2015 we further developed and applied various methods to analyse large drug screenings in cancer cell lines, in collaboration with the McDermott and Garnett groups at the Wellcome Trust Sanger Institute. A major question we addressed was consistency among drug screenings, jointly with colleagues at Sanger Institute, Broad Institute, Novartis, and Massachusetts General Hospital.

We continued to develop approaches to model signalling networks. We developed a novel methodology to build logic signalling networks from phosphoproteomic data generated from mass spectrometry shotgun data (PHONEMeS, http://www.cellnopt.org/PHONEMeS), and used it to study the effect of cancer drugs in breast cancer cells. This work was carried out in collaboration with the Cutillas group at Barts and the London School of Medicine and Dentistry.

We again helped organise DREAM Challenges (Dialogues in Reverse Engineering Assessment of Methods). DREAM is a community effort that uses 'challenge' events to advance the inference of mathematical models of cellular networks. We finalised the analyses of two challenges, one predicting the toxic effects of compounds based on genetic and chemical data, and one inferring signalling networks from phospho-proteomic data in cancer cells. We also started to run a new challenge predicting drug combination therapies in cancer, jointly with AstraZeneca, Sage Bionetworks, and the Sanger Institute.

Future plans

Based at our new home at RWTH Aachen University, we will continue to develop methods and tools to understand signal transduction in human cells and its potential to yield insights of medical relevance. Our main focus will be modelling signalling networks using phospho-proteomics data, and integrating these networks with gene regulation and metabolism. An area of particular interest for us will be single-cell signalling data. We will also develop methods to identify drug targets by integrating genomic and transcriptomic data


An international study published in Nature Biotechnology presented the combined results of a 2013 DREAM Challenge: a crowdsourcing initiative to test how well the effects of a toxic compound can be predicted in different people. The study, which is relevant to public and occupational health, shows that computational methods can be used to predict some toxic effects on populations, although they are not yet sensitive enough to predict such effects in individuals. It also presents algorithms useful for environmental risk assessment.

Julio-Saez Rodriguez

PhD University of Magdeburg, 2007. Postdoctoral work at Harvard Medical School and M.I.T.

At EMBL-EBI since 2010.
Joint appointment at Genome Biology Unit (EMBL-HD).
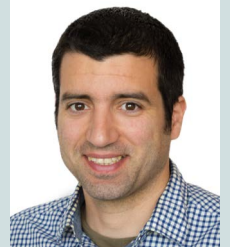
with information on signalling pathways. Using these novel methods we will address questions such as: What are the origins of the profound differences in signal transduction between healthy and diseased cells and, in the context of cancer, between normal and transformed cells? What are the differences in signal transduction among cancer types? Can we use these differences to predict disease progression? Do these differences reveal valuable targets for drug development? Can we study the side effects of drugs using these models?

Selected publications

Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium, et al. (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 528:84-87

Eduati F, Mangravite FM, Wang T, et al. (2015) Opportunities and limitations in the prediction of population responses to toxic compounds assessed through a collaborative competition. *Nature Biotechnol.* 33:933

Henriques D, Rocha M, Saez-Rodriguez J, Banga JR (2015) Reverse engineering of logic-based differential equation models using a mixed-integer dynamic optimisation approach. *Bioinformatics* 31:2999-3007

Terfve CD, Wilkes EH, Casado P, et al. (2015). Large scale models of signal propagation in human cells derived from discovery phosphoproteomic data, *Nature Commun.* 6:8033
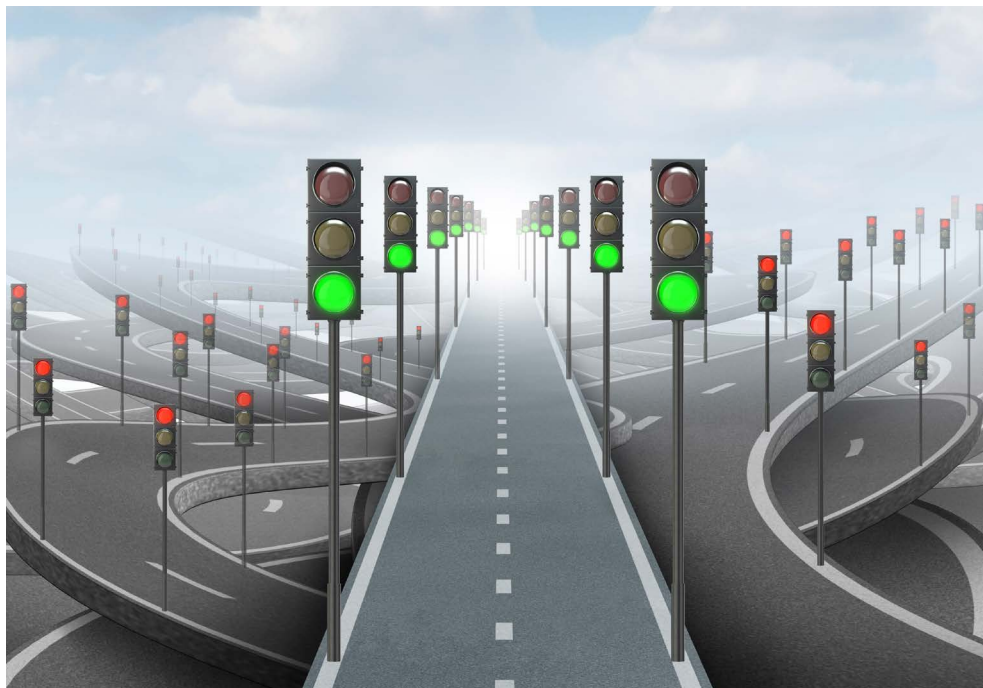
A method developed by the Saez-Rodriguez group takes the noisy data from MS experiment, which is a large network with many interconnected cascades of kinase activities, filters the noise, and integrates the data. This is done entirely in the context of what is known about kinases and their substrates, so that it shows activities are connected. This gives a new perspective on how a given drug is impacting the system under study.

Stegle Group

Statistical Genomics &
Systems Genetics

We use computational approaches to map genotype to phenotype on a genome-wide scale. Using statistics, we seek 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.

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 the different steps of a given 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. Our methodological work ties in with experimental collaborations, and we actively develop methods to fully exploit large-scale datasets that are obtained using the most recent technologies. In doing so, we derive computational methods to dissect phenotypic variability at the level of the transcriptome, epigenome and the proteome, and derive advanced statistical methods for the emerging field of single-cell biology.

Major achievements

In 2015 we developed and applied methods for linking genetic variation data and phenotype. We derived a new statistical model that allows studying genetic associations between sets of genetic variants and multiple correlated phenotypes (Casale et al. 2015). The model makes it possible to interrogate very large cohorts with hundreds of thousands of samples, increases statistical power and clarifies the genetic basis of phenotypic correlation between genetically diverse individuals.

In addition to deriving new statistical pools, we actively applied these methods to study the regulatory consequence of copy-number changes and other structural variants in the human genome on gene-expression levels. In a collaboration with Korbel team at EMBL Heidelberg, we surveyed the effect of structural variants on gene expression at a genome-wide scale using the data from the final release of the 1000 Genomes Project (Sudmant et al., 2015).

In parallel to our efforts in population genomics, we extended our methodological work to the field of single-cell genomics. In collaboration with the Marioni and Teichmann groups at EMBL-EBI we devised new ways to dissect transcriptional heterogeneity between single cells (Buettner et al., 2015). Our approach, for the first time, enables modelling both known and unknown factors that underlie single-cell transcriptome variation. This method has already helped identify new sub-clusters of cells in single-cell RNAseq studies of differentiating T-cells and will be an important building block for our future aims.

Future plans

In 2016 we will continue to develop innovative statistical approaches to analyse data from high-throughput genetic and molecular profiling studies. Our on-going efforts are motivated by single-cell genomics data, in particular using new assays that allow to prolife multiple molecular layers in the same sets of cells in parallel. By linking these layers, we hope to gain new insights into gene regulation, the sources of transcriptome heterogeneity and, ultimately, cell-fate decisions in development and cell differentiation.
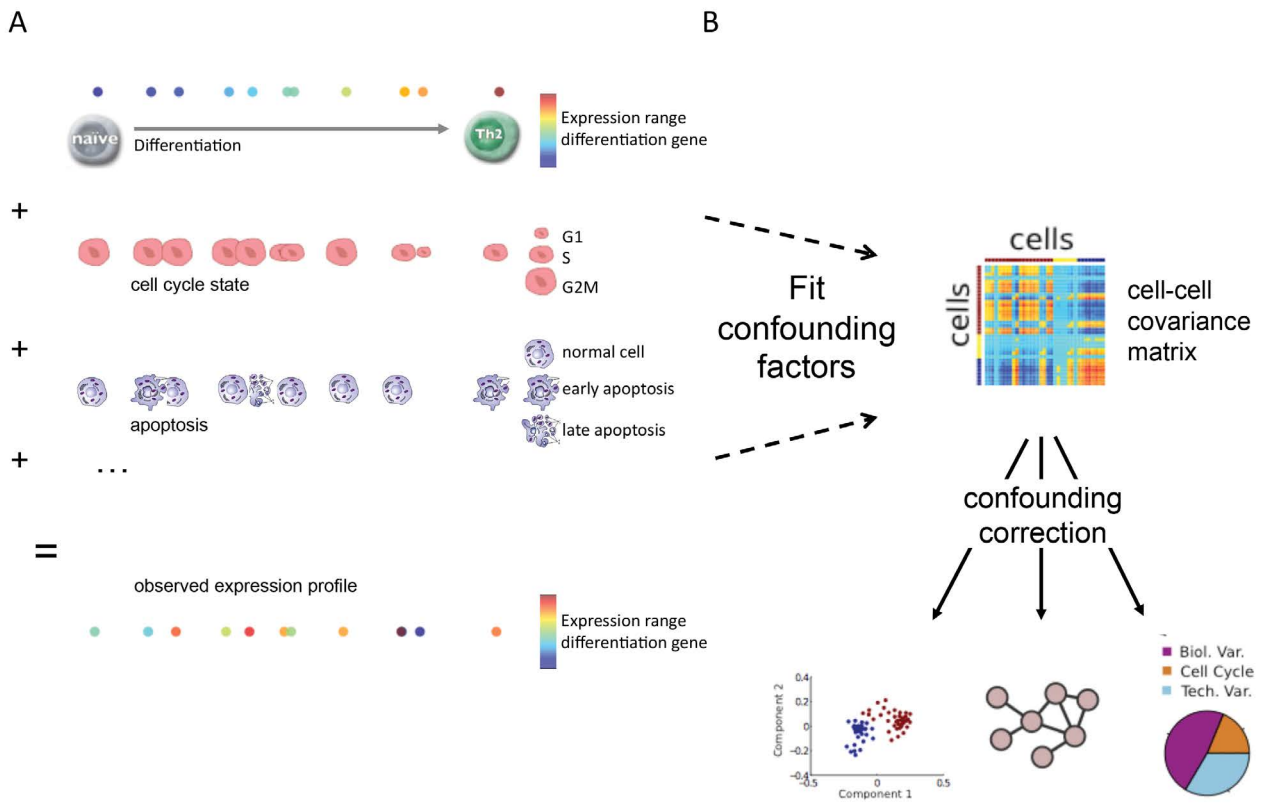
Selected publications

Buettner F, et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnol.* 33:155-160

Casale FP et al. (2015) Efficient set tests for the genetic analysis of correlated traits. *Nature Methods* 12: 755-758

Stegle O, Teichmann SA and Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nature Rev. Genet.* 16:133-145

Stephan J, Stegle O and Beyer A (2015) A random forest approach to capture genetic effects in the presence of population structure. *Nature Commun.* 6:7432

Sudmant PH, et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526: 75-81

Oliver Stegle

PhD in Physics, University of Cambridge, 2009.
Postdoctoral Fellow, Max Planck Institutes
Tübingen, 2009–2012.

At EMBL-EBI since November 2012

Statistical methodology to dissect transcriptional heterogenetiy in single-cell RNA-Seq datasets (adapted from Buettner et al., 2015). Left: underlying source of variation in single-cell transcriptome data. Right: Our scLVM approach to identify and account for such factors.

Teichmann Group

Gene Expression Regulation and Protein Complex Assembly

Our group seeks to elucidate general principles of gene expression and protein complex assembly. We study protein complexes in terms of their 3D structure, structural evolution and the principles underlying protein-complex formation and organisation.

Our group seeks to elucidate general principles of gene expression and protein complex assembly. We endeavour to understand how changes in cell state are regulated at the transcriptomic and epigenetic levels by studying the differentiation of mouse T-helper (Th) cells and embryonic stem cells (mESC) at the single-cell level. We develop novel computational and experimental approaches in single cell-genomics to address our questions. We also study the 3D structure and evolution of protein complexes, and the principles underlying protein-complex formation and organisation.

The recent development of computational and experimental tools in single cell genomics has led us to many exciting new discoveries in single-cell biology. We focus on the evolution and dynamics of regulatory and physical interaction networks, combining computational and mathematical approaches with genome-wide and gene/protein experiments.

Major achievements

In 2015 the field of single-cell genomics expanded rapidly, allowing our group to exploit scRNA-seq technology to make many exciting new discoveries in both stem cells and T cells. It was an exciting time of change as Sarah was awarded the European Molecular Biology Organisation (EMBO) Gold Award and our group prepared to move to the Wellcome Trust Sanger Institute.

Transcriptional regulation is critical for maintaining the pluripotency of mESCs, and culture conditions are also important for their self-renewal in vitro. Using single-cell RNA-seq (scRNA-seq) approach, we investigated the transcriptome profiles of mESCs in three different culture conditions (serum, 2i, and a2i), which represent slightly different pluripotent states. The past understanding of these distinct pluripotent states is limited to "bulk" analyses, which fail to capture the complexity of cellular states within mESCs. Our work (Kolodziejczyk et al., 2015) revealed additional pluripotency network genes, including Ptma and Zfp640, which demonstrate the value of this resource for future discovery.

We use single-cell technology to study immune-cell development, notably the T-cell receptor (TCR) – a diverse protein mediating the recognition of antigen fragments as peptides bound to major histocompatibility complex (MHC) molecules. To investigate TCR clonal relationships between cells alongside their transcriptional profiles and functional responses, we developed a novel computational method, TraCeR (Stubbington, Lonnberg et al., 2016). This new method enables us to link TCR sequence with transcriptional profiles in individual cells with high accuracy and sensitivity. By applying TraCeR to scRNA-seq data from a mouse Salmonella infection model, we showed that T-cell clonotypes span early-activated CD4+ T cells as well as mature effector and memory cells.

One of the key challenges in single-cell genomics is to ensure that only single, live cells are included in downstream analysis. With the aim of increasing data quality and bioinformatics reliability, we developed a computational pipeline (Ilicic et al., 2016) for processing scRNA-seq data and filtering out low-quality cells using a curated set of over 20 biological and technical features. Our pipeline ensures that only correctly annotated cells are included in analyses – a crucial tool for drawing more accurate biological conclusions.

We used mass spectrometry data together with a large-scale analysis of structures of protein complexes to examine the fundamental steps of protein assembly. In collaboration with colleagues at the Cavendish Laboratory at the University of Cambridge, we analysed the tens of thousands of protein complexes for which three-dimensional structures have been experimentally determined, and identified repeating patterns in the assembly transitions that occur (Ahnert et al., 2015). We used these patterns to create a new 'Periodic Table' of protein complexes, which provides a predictive framework for anticipating new, unobserved topologies of protein complexes. The core work for this study is in theoretical physics and computational biology, combined with mass spectrometry work by our colleagues at Oxford University.

Future plans

The Teichmann group will move to the Wellcome Trust Sanger Institute in 2016, where we will continue to explore structural bioinformatics of protein complex assembly and expand their work in genomics of gene expression. By integrating scRNA-seq data, epigenetic profiling, CRISPR/Cas9 screening and high-throughput high-content imaging data, we seek to gain a comprehensive overview of cellular circuitry and decision-making in switching from one cell type to another. We will pursue methods development in single-cell bioinformatics approaches, as there are many open questions that require new statistical and computational techniques. Together with the Marioni and Stegle groups at EMBL-EBI, we are keen to find new ways to dissect technical from biological cell-to-cell variation in gene expression, predict regulatory relationships, gene expression modules and cell states from the new flood of single cell RNA-sequencing data.

Sarah Teichmann

PhD 2000, University of Cambridge and MRC Laboratory of Molecular Biology. Trinity College Junior Research Fellow, 1999-2005. Beit Memorial Fellow for Biomedical Research, University College London, 2000-2001. MRC Laboratory of Molecular Biology, 2001-12. Fellow and Director of Studies, Trinity College, since 2005. Department of Physics/Cavendish Laboratory, University of Cambridge, 2013-2016.

At EMBL-EBI and Sanger Institute, 2013-2016.

Selected publications

Ahnert SE, et al. (2015) Principles of assembly reveal a periodic table of protein complexes. *Science* 350:aaa2245

Kolodziejczyk AA, K et al. (2015) Single cell RNA-sequencing of pluripotent states unlocksmodular transcriptional variation. *Cell Stem Cell* 17:471-485

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58:610-620

Stubbington MJ, Lönnberg T, et al. (2016) T-cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* 13:329-332

Ilicic T, et al. (2016) Classification of low-quality cells from single-cell RNA-seq data. *Genome Biol*. 17:29
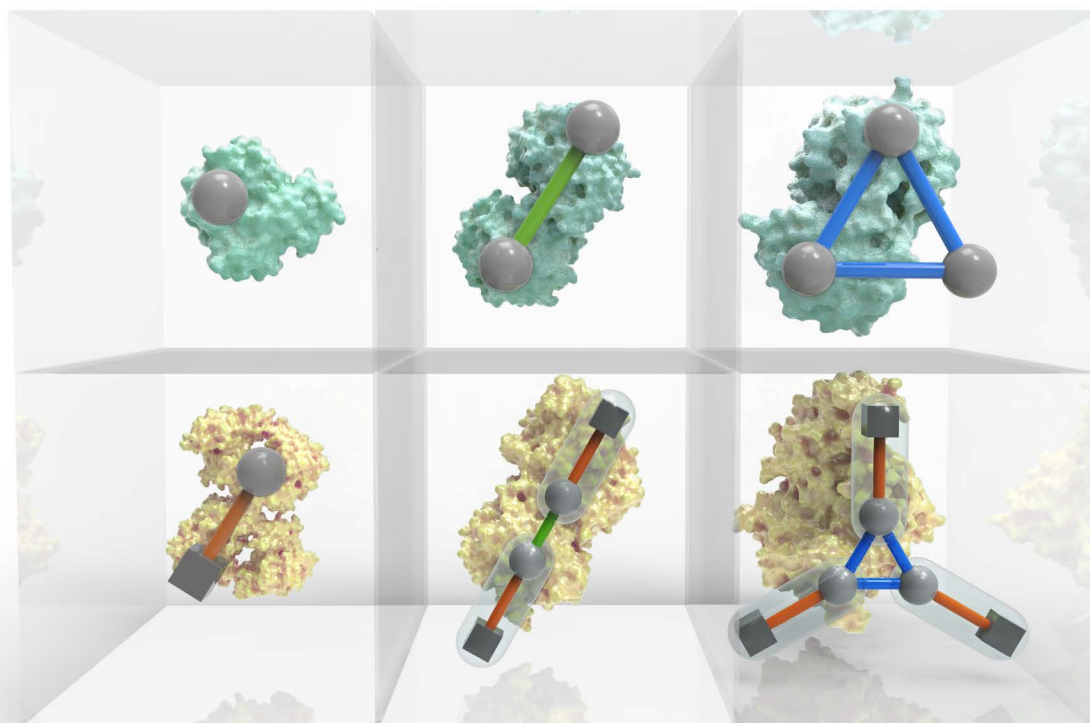
The Periodic Table of Protein Complexes, published in Science*, offers a new way of looking at the enormous variety of structures that proteins can build in nature, which ones might be discovered next, and predicting how entirely novel structures could be engineered. Created by an interdisciplinary team of researchers, the Table provides a valuable tool for research into evolution and protein engineering.*

Thornton Group

The goal of our research is to understand how biology works at the molecular level, with a particular focus on protein structure and evolution and ageing.

We explore how enzymes perform catalysis by gathering relevant data from the literature and 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 involves integrating heterogeneous data with phylogenetic relationships within protein families, which are based on protein-structure classification data derived by colleagues at University College London (UCL). The practical goal of this research is 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 diseases.

To understand more about the molecular basis of ageing in different organisms, we participate in a close collaboration with experimental biologists at UCL. Our role is to analyse functional genomics data from flies, worms and mice and, by developing new software tools, relate these observations to effects on lifespan.

Major achievements

Chemistry and evolution of isomerases

Biologists are challenged with the functional interpretation of vast amounts of sequencing data derived from genomics initiatives. Among all known proteins, the function of enzymes is probably the most investigated and best described at the molecular level. Together with enzymes changing the redox state of substrates and transferring chemical groups between molecules, isomerases catalyse interconversion of isomers, molecules sharing the same atomic composition but different arrangements of chemical groups. In this study (Martinez-Cuesta et al, 2016), we catalogued the isomerisation reactions known to occur in biology using a combination of manual and computational approaches. This method provides a robust basis for comparison and clustering of the reactions into classes. Comparing our results with the Enzyme Commission (EC) classification, the standard approach to represent enzyme function on the basis of the overall chemistry of the catalysed reaction, expands our understanding of the biochemistry of isomerization. The grouping of reactions involving stereoisomerism is straightforward with two distinct types (racemases/epimerases and cis-trans isomerases), but reactions

entailing structural isomerism are diverse and challenging to classify using a hierarchical approach. This study provided an overview of which isomerases occur in nature, how we should describe and classify them, and their diversity.

Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies

Enzymes, as biological catalysts, form the basis of all forms of life, but how these proteins have evolved their functions remains a fundamental question in biology. Using a range of computational tools and resources we compiled information on all experimentally annotated changes in enzyme function within 379 structurally defined protein domain superfamilies, linking the changes observed in functions during evolution, to changes in reaction chemistry (Furnham et al, 2015). Many superfamilies show changes in function at some level, although one function often dominates one superfamily. We used quantitative measures of changes in reaction chemistry to reveal the various types of chemical changes occurring during evolution and provided detailed examples. We used structural information of the enzymes' active sites to examine how different superfamilies have changed their catalytic machinery during evolution. Some superfamilies have changed the reactions they perform without changing catalytic machinery; in others, large changes of enzyme function, both in terms of overall chemistry and substrate specificity, have been brought about by significant changes in catalytic machinery. Interestingly, the relatives of some superfamilies perform similar functions using different catalytic machineries. Our analysis highlighted characteristics of functional evolution across a wide range of superfamilies, providing insights that are useful in predicting the function of uncharacterised sequences and in the design of new synthetic enzymes.

Longevity GWAS using the *Drosophila* Genetic Reference Panel

The results of our genome-wide association study (GWAS) for *Drosophila* lifespan, based on 197 Drosophila Genetic Reference Panel (DGRP) lines (Ivanov et al. 2015), suggested that the top associated genes provide good candidates for further investigation into their relationship with lifespan and ageing. While our

findings regarding association were intriguing, the direction of effect required experimental validation. In collaboration with the Institute of Healthy Ageing at UCL, we tested the effect of manipulating 10 of the genes most significantly associated with lifespan, using *Drosophila* strains available from public stock centres. We tested RNAi knock-down and UAS over-expression lines using a ubiquitous adult-onset driver system (da-GS). Preliminary results suggest that a proportion of the genes significantly reduced lifespan. Moreover, over-expression of one of the tested genes significantly increased lifespan as compared to control.

Allergy: the price of immunity?

Allergy is an increasingly widespread clinical problem that leads to various conditions such as allergic asthma and susceptibility to anaphylactic shock. These conditions arise from exposure to a range of environmental and food proteins ('allergens') that are recognised by a form of immune system antibody called IgE. This part of the immune system is thought to have evolved to provide mammals with additional rapid-response mechanisms to combat metazoan parasites. In this study, we addressed the pertinent question, 'what makes an allergen an allergen?' Although they constitute a very small percentage of known proteins, they appear to be diverse and unrelated.
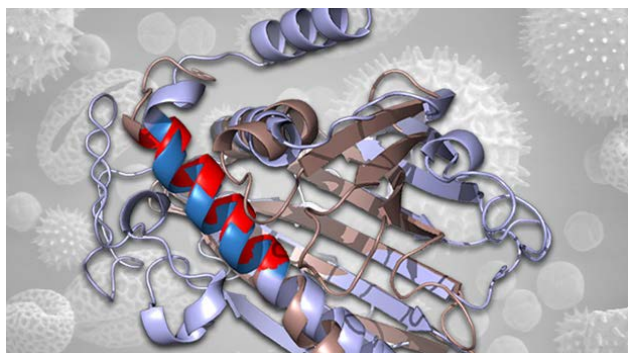
Thornton group research in 2015 showed that the off-target effects of the immune system in allergy are due to the significant molecular similarities they identified between environmental allergens and parasitic worm proteins. The findings demonstrate that allergy in humans is a trade-off for immunity to parasites.

Using computational studies (Tyagi et al., 2015), we established molecular similarity between parasite proteins and allergens that affect the nature of immune response and are able to predict the regions of parasite proteins that potentially share similarity with the IgE-binding region(s) of the allergens. Our experimental studies supported the computational predictions, and we presented the first confirmed example of a plant-pollen-like protein in a worm that is targeted by IgE. The results of this study will enable us to predict likely allergens in food and environmental organisms and help in the design of protein molecules to treat allergy in the future.

Janet Thornton

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 and their mechanisms using structural and chemical information, we will explore how enzymes, their families and pathways have evolved. We will continue our study of reactions and use this new knowledge to improve chemistry queries across our databases. We will study sequence variation in different individuals and explore how genetic variations impact on the structure and function of a protein, and sometimes cause disease. Using evolutionary approaches, we hope to improve our prediction of protein function from sequence and structure. We will continue our ageing studies, exploring longevity sub-phenotypes, identifying small molecules that might modulate lifespan in model organisms, and validating and characterising the effects of a small set of genes on lifespan in Drosophila.

Selected publications

Cuesta SM, Rahman SA, Thornton JM (2016) Exploring the chemistry and evolution of the isomerases. *Proc Nat Acad Sci* (in press); doi:10.1073/pnas.1509494113

Furnham N, et al. (2015) Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *J. Mol. Biol.* 428:253-267

Ivanov DK, et al. (2015) Longevity GWAS using the Drosophila genetic reference panel. *J. Gerontol Series A: Biol. Sci. Med. Sci.* glv047

Tyagi N, et al. (2015) Comparisons of allergenic and Metazoan parasite proteins: allergy the price of immunity. *PLoS Comput. Biol.* 11:e1004546

SERVICE TEAMS

TRAINING PROGRAMME

RESEARCH GROUPS

INDUSTRY PROGRAMME

EXTERNAL-FACING ACTIVITIES

CONTENTS

Industry Programme

Industry Programme

The EMBL-EBI Industry Programme has been an important and vibrant part of EMBL-EBI since 1996, providing regular contact with commercial users and informing them of the institute's future directions.

EMBL-EBI interacts with industry in many ways. We support pre-competitive collaboration through a subscription-based programme that welcomes larger companies making use of the data and resources provided by EMBL-EBI. Member companies represent primarily the pharmaceutical sector but also the agri-food, nutrition, healthcare and consumer goods industries. The programme provides quarterly strategy meetings, expert-level workshops on topics prioritised by the members and other forms of communication including webinars and face-to-face meetings.

A separate set of activities supports SMEs through outreach activities organised at EMBL-EBI in Hinxton and internationally in coordination with the EMBL International Relations dept. Our programme also serves as an interface between EMBL-EBI industry-focussed initiatives and organisations including the Innovative Medicines Initiative (IMI), the Pistoia Alliance, the Clinical Data Interchange Standards Consortium (CDISC) and others.

Major achievements

Our well-attended quarterly strategy meetings provided opportunities for members to learn first-hand about emerging developments at EMBL-EBI and to prioritise future activities, including knowledge-exchange workshops (see Table). In 2015 we ran nine of these workshops, which are a major benefit of membership: seven were held at EMBL-EBI, one in Cambridge, Massachusetts at the Genzyme Centre and one in Waltham, MA hosted by AstraZeneca. The US-held workshops had very high levels of participation, and served to extend the reach of the Industry Programme to member companies whose discovery activities are primarily in the US. The Sanofi (Genzyme)-hosted workshop covered Immunogenomics, and the AstraZeneca-hosted event focused on Translational NextGen Sequencing.

We were very pleased to welcome Astex Pharmaceuticals, part of Otsuka Pharmaceuticals, as a new member of the Industry Programme in December 2015.

In March 2015, we ran our annual event for small and medium-sized enterprises (SMEs): the SME Bioinformatics Forum. This was jointly organised with One Nucleus, a membership organisation for life-science and healthcare companies based in Cambridge and London, the heart of Europe's largest life science and healthcare cluster, and with the support of the Council of European Bioregions (CEBR) and UK Trade and Investment. We hosted, jointly with InnovateUK, a workshop on integrated 'omics, in advance of an InnovateUK funding round. Our third SME outreach event in 2015, also organised with One Nucleus, was the ON Helix Welcome Reception and BioNewsRound Award.

In collaboration with the Argentine Ministry of Science, Technology and Productive Innovation (MINCYT) and the Argentine Chamber of Biotechnology (CAB), we organised a two-day workshop in Argentina, EMBL's newest associate member state. The event focussed on applications of bioinformatics and genomics in healthcare, agriculture and livestock breeding.

Dominic Clark

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. 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. In addition to finding cost-effective paradigms to manage expanding volumes of data, we will collaborate to establish methods to ensure the appropriate integration of information and computational models so that the confidentiality of proprietary, licensed and personal information is protected in a manner that promotes innovation and translation into practical benefits.

Through efforts such as the Innovative Medicines Initiative and the Pistoia Alliance, we are keen to influence, support and encourage this integration through direct involvement in EU-funded projects and in the development of industry-driven data and information standards. Continuing these close co-operations will also be beneficial to the success of ELIXIR, the pan-European infrastructure for biological information.

Industry workshops

- *In silico ADMET prediction*
- *Immunogenomics*
- *Translational NGS*
- *Enabling the translational bioinformatician*
- *Quantitative systems pharmacology*
- *Data enhancement through scientific literature: integrating the literature with data to enable discovery*
- *The EMBL-EBI RDF Platform*
- *Electronic Medical Records for Drug Discovery: Connectivity Map and LINCS (organised in association with The Broad Institute)*

Innovative Medicines Initiative projects

Our programme helps its partners and EMBL-EBI researchers establish collaborations in the context of the Innovative Medicines Initiative. Examples include:

- *EBiSC: European Bank for induced-pluripotent Stem Cells*
- *eTOX: Developing innovative in silico strategies and novel software tools to better predict the toxicological profiles of small molecules in early stages of the drug development pipeline.*
- *EMTRAIN: A platform for education and training covering the whole life cycle of medicines research, from basic science through clinical development to pharmacovigilance.*
- *DDMoRe: The Drug Disease Model Resources consortium: Developing a public drug and disease model library.*
- *EHR4CR: Designing a scalable and cost-effective approach to interoperability between electronic health record systems and clinical research.*
- *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.*
- *Open PHACTS, the Open Pharmacological Concepts Triple Store: Reducing barriers to drug discovery in industry, academia and for small businesses.*

Selected publication

Hendrickx DM, Aerts HJ, Caiment F, et al. (2015) diXa: a data infrastructure for chemical safety assessment. Bioinformatics 31:1505-1507

External-Facing Activities
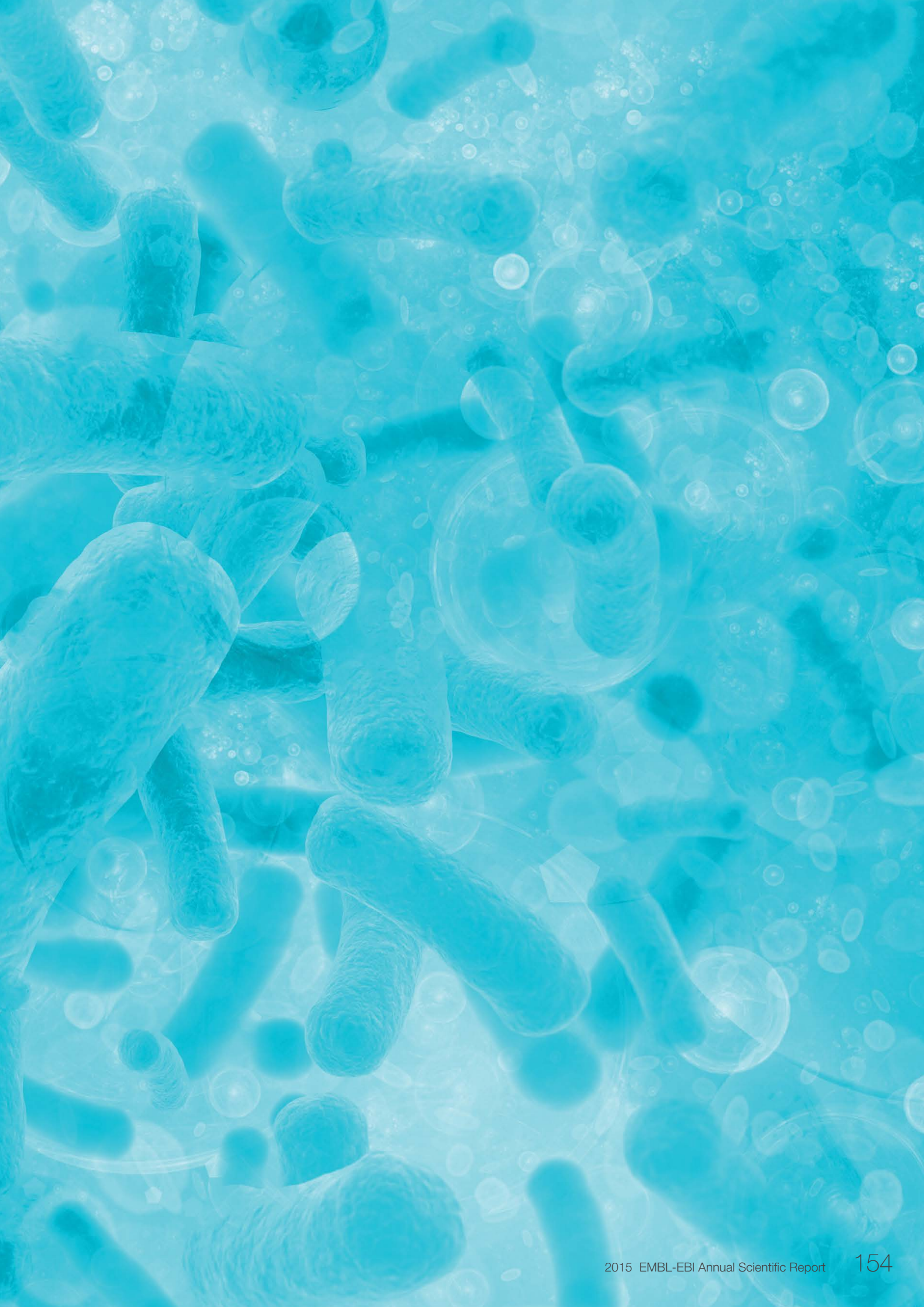
Web Production

The Web Production team manages the web infrastructure, which comprises more than 2000 web servers running on 700 virtual machines. These provide platforms for web service development and robust, secure frameworks for deploying public bioinformatics services.

As part of the Technical Services Cluster, our team is responsible for the global EMBL-EBI search engine, which indexes the records from all EMBL-EBI data resources, and for the Job Dispatcher framework and corresponding SOAP/REST web services, which provide programmatic access. We also facilitate access to infrastructure services such as core project-management and user-support software.

Major achievements

In 2015 the EBI Search engine indexed over 1.5 billion data records from EMBL-EBI data resources. The Job Dispatcher framework, through which users run Blast, InterProScan and Clustal Omega services, handled more than 150 million job requests (compare to 110 million jobs in 2014). This represents an increase of 36% over a 12-month period. The EBI Search and Job Dispatcher tools are amongst EMBL-EBI's top 10 most used services.

The web visibility of EMBL-EBI data resources and tools depends largely on work done by our team. Approximately 700 virtual hosts and more than 2000 distinct service endpoints exist, many consuming RESTful APIs. In 2015 we phased out the SOAP API and ensured that the 20 EMBL-EBI data resources that used it were switched over to the RESTful interface.

The data centre move to Hemel Hempstead, completed in early 2015, presented an opportunity to introduce virtual frameworks more widely. As part of the institute-wide Tools Committee, we phased out several legacy services (e.g. MaxSprout, DaliLite, ClustalW2, CENSOR, WU-BLAST) and deployed new ones (e.g. NCBI-Blast+, MegaBLAST, HMMer3 search utilities).

Uninterrupted service

During the data centre move, which involved more than 600 virtual machines and the physical consolidation of EMBL-EBI's external-facing IT infrastructure, service was uninterrupted and usage increased considerably – our team maintained a steady level of response to service requests. In our Hinxton facility, we deployed live services for high-usage services such as the EBI Search, Job Dispatcher, UniProt and Europe PMC, and enabled a lightweight fail-over infrastructure for other services.

Increased usage

Approximately 10 million unique addresses made more than 8 billion requests to EMBL-EBI during 2015. This figure includes everything on the ebi.ac.uk domain, Europe PMC, UniProt and EnsemblGenomes, but excludes Ensembl.org (as it was run from the Wellcome Trust Sanger Institute). Our FTP and Aspera traffic totalled 148.1 petabytes.

The most heavily used services via the Job Dispatcher framework were InterProScan, Clustal Omega, pfamscan and NCBI Blast+. Projects that consume these tools include uniprot.org, ensemblgenomes.org, pfam.xfam.org, PDBe.org and InterPro. External users integrating these tools into their workflows and pipelines include Blast2GO, Galaxy, .Net Bio, BlastStation, BioServies, KEIO Bioinformatics Web Services, Yabi, BlastStation, STRAP, T-Coffee, CCP4, Geneious and GMU-metagenomics.

Together with the service teams, we continue to provide up-to-date sequence libraries. As of December 2015 EMBL-EBI hosted some 13 000 distinct sequence libraries, primarily in the ENA and UniProt but also in resources such as EnsemblGenomes, which comprises approximately 121 000 nucleotide and 30 500 protein sequence libraries.

EBI Search API

The RESTful API of the centralised EBI Search was heavily used in 2015, with more than 20 data resources using it as their core search system. These include RNA central, the Enzyme Portal, MetaboLights, Proteomics Data Discovery Index (ddi), InterPro and the Expression Atlas, among others. Our team puts considerable development efforts into ensuring performance remains high despite fast-paced data growth.

We fully virtualised all of the EBI Search engine components in 2014, which led to a 20% reduction in its memory footprint. We also improved the functionality of its facets, which help our users filter results (e.g. using ontologies, controlled vocabularies, taxonomy, keywords, dates, molecule type). We overhauled this functionality in the search system's core Lucene libraries, which resulted in faster results filtering and more consistent term binding and quality assurance (e.g. spotting potential annotation errors).

Rodrigo Lopez

Web Production
MSc Veterinary Medicine, Oslo
Veterinaerhoyskole, 1984. MSc Molecular
Biology and Toxicology and Informatics,
University of Oslo, 1987.

At EMBL-EBI since 1995.

Reporting on usage

We further developed our log analytical capabilities to accommodate the broad adoption of Elastic Search, Logstash and Kibana (ELKs) for the interactive visualisation of live servers and for general improvements in usage. We use this technology to visualise web statistics, EBI Search usage, Job Dispatcher and live web server behaviour, and to generate the graphical summaries of usage provided in this Report.

Outreach, training and support

In 2015 the Technical Services Cluster consolidated helpdesk activities, optimising our feedback and development pipelines and empowering other teams to handle their own support requests more efficiently. Our team handled approximately 700 tickets, the bulk of which were concerned with programmatic access to tools, search and data acquisition, best practices and training.

We participated in 14 outreach and training events, both in the context of the EMBL-EBI Training Programme and independently at external conferences and workshops. Our trainers provided overviews of how best to access EMBL-EBI resources to a range of audiences, and took part in more specialised training such as tools and techniques for sequence manipulation and searching, multiple sequence alignment and using Web Service APIs.

Future plans

During During 2016 we will focus on two major strands of work: improving the EBI Search interface, and improving the exposure of mappings between different data resources. To improve usability and the discoverability of data, we will analyse the usage patterns associated with query types and develop a RESTful client to replace the existing web user interface. This work will involve close collaboration with user experience designers in the Web Development team.

We will further improve interoperability between EMBL-EBI services and the EBI Search, and enhance data logistics to keep up with data growth. We will do this in the context of projects led by Pfam, InterPro, UniProt, ENA and EnsemblGenomes teams. We will continue to phase out obsolete services (e.g. FingerPRINTScan, PROSITE Scan) and launch more scientifically relevant ones.

Security is paramount to all our operations. Our web administrators will deploy the most up-to-date intrusion-detection tools and, based on industry best practices, enhance the security of EMBL-EBI web services. We will also adopt hybrid Cloud technologies to enable flexibility and provide a broader range of services to EMBL-EBI service teams and research groups.

We will contribute to the deliverables of the ELIXIR Excelerate programme, for example by establishing standard service metrics and refining the ELIXIR tool and database registry.

Selected publications

Li W, Cowley A, Uludag M, et al. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 43:w580-w584

Squizzato S, Park YM, Buso N, et al. (2015) The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI. Nucleic Acids Res 43:w585-w588
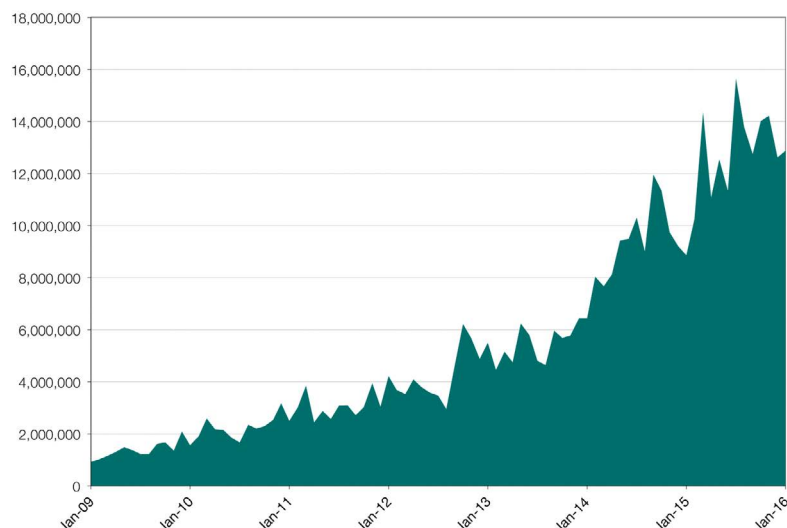
Jobs run on the Dispatcher framework, 2009 through 2015. In 2015 we ran more than 151 million jobs, helping users establish analytical pipelines to query data and combine it with their own. The average number of jobs per month was 12.6 million (compare to 9.2 million in 2014).

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. We are 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 BioMedBridges, 1000genomes, INSDC, ProteomeXchange and HGNC. We also provide front-end user experience (UX), web design and development for the CTTV Target Validation 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

Events and projects

EMBL-EBI offers a wide variety of events, seminars and talks that are held both onsite and at host institutes throughout the year. In 2015 our team worked with several teams to integrate the various diverse event-management systems into a single, centralised portal for announcing events on our own website and potentially those of partner organisations. The new system avoids duplication of information and effort, makes events more discoverable and presents them in a number of attractive ways. The Events Portal project was a collaborative effort to harmonise and update event information creation, presentation and dissemination, and on the technical side involved the merging of three Drupal 6 sites into a single Drupal 7 instance. Information pertaining to events, people, locations and funders is now centralised in the EMBL-EBI content database and integrated with the EDAM Topics ontology, making it easier to tag events consistently.

Working with the Training team, we completely redesigned and modernised the training programme webpages, presenting users with clear information and dynamic tools to find events of interest easily. The Training and Events projects together resulted in a vastly improved presentation of events, seminars and courses, reduced technical debt and enriched our central content database with ontology terms, which supports the further integration of related content types.

Global website

In 2015 we reengineered the global EMBL-EBI website to be responsive on mobile devices. Although the majority of our web traffic is from desktop machines, mobile device usage has increased considerably and the need for optimised display on various devices undeniable. We retrofitted the existing site to a responsive design using UX principles, modern design approaches and analytics. This retrofit was a stop-gap measure; the,larger project to improve the backend will be initiated in 2016. This will involve redeveloping the site to use a responsive framework that can also be extended to other groups and teams at EMBL-EBI. In collaboration with External Relations and as part of the newly formed Web Content Committee, we intend this redesign process to also enhance the user experience when navigating between the EMBL-EBI website and its services, ELIXIR and EMBL.

CTTV

Members of our team working on Open Targets (formerly the CTTV) in 2015 went from outlining a database schema and roughly designing a website on paper to launching a fully functional Target Validation Platform, which had 8000 visits in its first month of operation. The team has taken a UX-driven, agile, iterative approach to engage with stakeholders and potential users to design, implement and test functionality, constantly refining and improving the platform. They carried out extensive user research to identify the needs of potential platform users working in both academia and the pharmaceutical industry, then designed and implemented the platform to be as intuitive, informative and responsive as possible with clear use cases and goals. Following a series of test releases, the platform was fully launched to the public in December 2015. In addition, our team moved the platform, Open Targets website and intranet to an Embassy Cloud workspace.

Technical Services Cluster Portal

The Technical Services Cluster (TSC) supports all EMBL-EBI services and resources as well as internal staff and their projects. We built a dedicated information portal in 2015 to make it easier for our internal users

to find the support they need. The new knowledge base provides access to documentation in a number of ways, including quick searching, related article suggestion, subscribing to articles and clear routes of further support. Working with our colleagues in the Systems Applications, Systems Infrastructure and Web Production teams, we improved our central ticketing system by designing an online 'help wizard' that helps users direct their questions to the relevant team and locate the appropriate documentation.

Reducing technical debt

In 2015 we upgraded or moved 14 Drupal 6 websites and portals in anticipation of support ending in February 2016. We worked with various stakeholders to identify the best solution for transferring and maintaining information from these sites.

User Experience Consultations

EMBL-EBI resources are developed according to our users' needs, and every year we carry out a detailed User Survey. In 2015 analysis of over 1400 responses to our survey provided valuable insights that we shared with resource teams and consulted on change requests as appropriate. Our UX experts collaborated with teams and groups across the institute on projects as diverse as facilitating workshops, planning survey planning design and analysis, designing services and establishing a user research workflow.

The events portal, launched in 2015, is a database-driven platform that pulls together all information for an event from a central source. This saves time, ensures content is up-to-date throughout the website and allows for consistent, appealing presentation.

Brendan Vaughan

Web Development
BSc Industrial Biochemistry, University of Limerick, 1995. MSc Bioinformatics, University of Limerick, 1997. Human Genome Mapping Project Resource Centre, 1998-1999. Lion Bioscience, 1999 - 2004.

At EMBL-EBI since 2004.
Team Leader since 2013.

Future plans

We are excited about working on the new responsive framework for the website which will not only make updating and maintaining the site in a mobile world easier, but will also allow us to tailor content to devices and react to any new interfaces more quickly. As part of the process a thorough content review in conjunction with analytics will also afford us the opportunity to ensure the content we serve is relevant and accessible for users.

As part of the ELIXIR Excelerate project the team is tasked with helping to create a mechanism to update the ELIXIR Registry with information about tools and resources housed at the EBI. This work is related to the efforts in the Tools Committee to characterise and store tool information and will most likely involve enhancing the content database.

While the release of Drupal 8 was anticipated earlier in 2015, it wasn't November that it finally appeared. The team has already undergone research and training in Drupal 8 and looks forward to using Drupal 8 for the next generation of our websites where appropriate.

The team is also part of several internal projects at EMBL-EBI to improve integration of services to our users in areas such as data submission and tool provision.

Selected publications

1000 Genomes Project Consortium, Auton A, Brooks LD, et al. (2015) A global reference for human genetic variation. Nature 526:68-74

Budd A, Dinkel H, Corpas M, et al. (2015) Ten simple rules for organizing an unconference. PLoS Comput Biol 11:e1003905

Lappalainen I, Almeida-King J, Kumanduri V, et al. (2015) The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet 47:692-695

Squizzato S, Park YM, Buso N, et al. (2015) The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI. Nucleic Acids Res. 43:w585-w588

Systems and Networking

The Systems Infrastructure Team manages EMBL-EBI's compute servers, storage, virtualisation, data centres and networking, including managing the campus 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.

Major achievements

In In 2015 our team maintained EMBL-EBI's growing compute infrastructure, which is now based on approximately 30 000 Central Processing Unit (CPU) cores. As our user base is very broad, with wide-ranging technical requirements, we organised two training days for approximately 70 members of staff to get up to speed with EMBL-EBI clusters. To support internal users working in a distributed computing environment, we built a new, 36-node Apache Hadoop cluster.

We maintained EMBL-EBI's secure, high-performance GridFTP data-transfer protocol in 2015, and as part of this work we completely rebuilt and automated the installation process. This makes it much simpler for users to share large datasets on our networks. In addition, we installed MongoDB machines to provide an alternative database platform for our internal users.

Security is paramount to the integrity of the controlled-access European Genome-phenome Archive (EGA), which EMBL-EBI jointly develops with the CRG in Barcelona. Our team re-implemented the vault firewall for the EGA in 2015, using a software approach. This improves the performance of the system substantially.

Our team put considerable efforts into maximising flexibility and optimising compute performance in 2015. We rebuilt our compute clusters' General Parallel File System completely, improving high-speed file access for multiple applications executing on different nodes within the clusters at the same time. We also continued to develop Elastic clusters, and plan to complete this work in 2016.

Given the scale of the storage challenge for EMBL-EBI, our team's focus is always on identifying and implementing the right technologies to maximise efficiency. For example, Docker containers create a wrapper for an application, with all of its dependencies, which provides a more standardised and software-development-friendly unit. Importantly, they also

Detail from the Genome Campus data centre, with Dawn Johnson, data centre engineer..

guarantee that the application always runs the same, regardless of its environment. In 2015 we enabled internal users to run these containers in our clusters, which resulted in improved efficiency and more consistent performance.

Our team is responsible for maintaining the raw data storage capacity of the institute, which exceeded 60 Petabytes at the end of 2015. A substantial portion of this represents raw sequence archives. Although

Petteri Jokinen

Systems & Networking
MSc in Computer Science 1990,
Helsinki University.

At EMBL-EBI since 1996.

we have moved away from using tape storage as a full backup system, favouring a distributed, object-based redundancy model, in 2015 we developed a new, cost-effective Object Tape Archive (comprising an IBM tape library and two tape robots) to ensure we had a classic backup for the multi-Petabyte sequence archives. For internal users, we added more high-performance storage systems, for example deploying a Lustre (Linux Cluster) for the Functional Annotation of Animal Genomes project.

We assisted EMBL-EBI teams in planning and implementing the deployment of multiple individual devices and larger projects. In addition to the Hadoop and GPFS projects described above, these included the OpenStack cloud infrastructure and increased capacity for virtual machine (VM) ware. We also improved and updated monitoring of all switch-related infrastructure components.
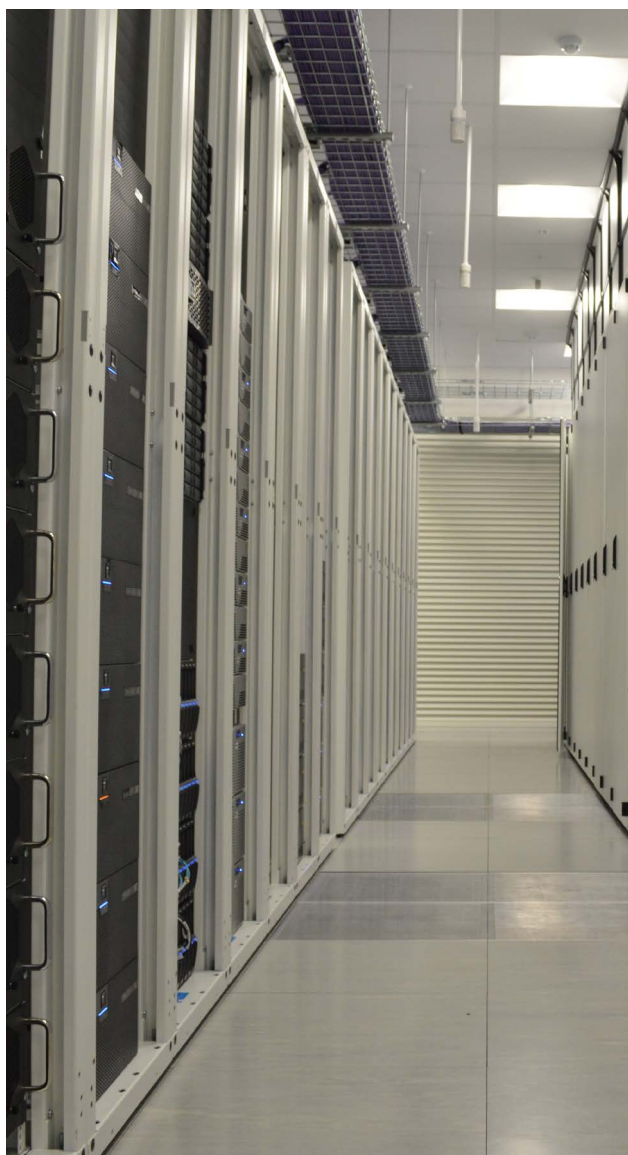
Detail from our Tier 3+ data centre in Hemel Hempstead.

Data Centres

In 2014 our team played a major role in the move of our external-facing data centres from London to Hemel Hempstead, and in 2015 we consolidated and rationalised the hardware in both Hinxton and Hemel Hempstead, for example eliminating many unsupported elements and services from the Hinxton data centre network. We also upgraded the core network links in Hinxton from 10 to 40 Gigabytes per second, further enhancing service performance across the board.

Physical consolidation was the major challenge of 2015. Across all three data-centre locations, we retired and disposed of around 450 devices (a mixture of storage and compute assets) and installed approximately 230 devices. Furthermore, as priorities shifted and the roles of devices changed, we physically moved a considerable number of devices between sites, for example moving 28 racks of equipment between two quadrants in the Hinxton data centre.

Future plans

Our goal in 2016 is to implement a new, two-tiered storage strategy for EMBL-EBI: one for performance, and another for capacity. We also plan to build a new, high-performance computing cluster for internal research. We will continue to provide leading-edge technical infrastructure for our external users, offering them the opportunity to run Docker containers in our clusters in a controllable way and, through our Elastic project, to dynamically resize logical clusters without interrupting running jobs.

Our team will replace our ageing Domain Name System (DNS) infrastructure, and tackle application load balancing problems. We will introduce network access control to the EMBL-EBI desktop access network, making it simpler for staff to carry out internal operations remotely. During 2016 we will also improve network monitoring, log analysis and network automation. In addition, we will further explore the emerging open network-switching market, which offers potential cost savings and improved flexibility through the application of open-source management software on commodity network switch hardware.

Systems Applications

The Systems Applications team manages EMBL-EBI's core IT applications, which include virtualisation, cloud technologies including Embassy Cloud, e-mail, authentication and desktop systems. We also provide database administration, for example for Oracle and MySQL, and provide basic services to EMBL-EBI staff as well as external collaborations.

Major achievements

Virtualisation and cloud

The virtual infrastructure that supports more than 2000 virtual machines (VMs) continues to provide a solidly reliable service, with 100% uptime. We completed virtualisation of existing services and shifted focus to supporting new services. EMBL-EBI's technical infrastructure remained within acceptable capacity figures despite the linear growth in usage. We retired several old hypervisors and storage systems, migrating their virtual machines to other systems without downtime.

Embassy Cloud work focused on commissioning the new OpenStack infrastructure and maintaining it through its first six months of usage, during which it provided robust, high-performance service.

We moved EMBL-EBI's contribution to the PanCancer Analysis of Whole Genomes (PCAWG) initiative to the OpenStack infrastructure, supporting the scientists by providing more high-performance compute to the project (n.b. CPU utilisation peaked at upwards of 80% over 1000 cores during processing). We transferred the Cloud resources to the internal EMBL teams working on the next phase of the PCAWG initiative, for which the Embassy Cloud continued to provide a high-performance compute environment.

Open Targets (formerly CTTV), which is hosted on the Embassy Cloud, continued to grow in resource usage, which we accommodated through maintenance and expansion of the Cloud. Working with eMedLab partners our team enabled the EMBL-EBI contribution to the eMedLab project, functioning as an active part of their Operations team. We will further support eMedLab by managing the transfer and maintenance of selected EMBL-EBI datasets on the eMedLab infrastructure for access by their tenant users.

In 2015 we began to join The Embassy Cloud to the European Grid Infrastructure, putting significant efforts into an initial proof-of-concept implementation of their Federated Cloud software. This will progress to a trial in 2016 to support ELIXIR.

Best Data Centre Project of the Year' award from the Computer Weekly European User Awards 2015, for work on Delphix architecture planning for public-facing technical infrastructure, a project led by Systems Applications team member Manuela Menchi.

Databases

We completed the consolidation of all database instances from hundreds of individual physical servers with dedicated SAN storage systems to VMs and a shared infrastructure. This consolidation reduced the data-centre capacity required for databases, provides additional flexibility in updating the underlying hardware without impacting services, and reduces EMBL-EBI's exposure to licence costs that are calculated on physical core count. This work included the retirement of obsolete hardware, database infrastructure activities such as backup, data protection and software upgrades to the database management system (i.e. MySQL, MongoDB, Vertica, Oracle, MS SQL Server), a review of monitoring systems, testing and support – all in conjunction with the migration of the EMBL-EBI internal authentication system from Network Information Service (NIS) to Lightweight Directory Access Protocol (LDAP).

We contributed to wider projects involving the IT infrastructure and data centres, reviewing the Database

Andy Cafferkey

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.

Standby Facility and Database Disaster Recovery infrastructure in Flint Cross and aligning it with the wider EBI Disaster Recovery plan. We also deployed a new storage system that supports the Standby database backend at the Hemmel Hempstead data centre.

In terms of new services, in 2015 we deployed a new Delphix appliance on Silver storage, providing developers with an on-site environment comparable to the production services offered in Hemel Hempstead.

Our work on Delphix architecture planning for public-facing technical infrastructure was recognized with a 'Best Data Centre Project of the Year' award at the Computer Weekly European User Awards 2015 (Figure).

Desktop support

In 2015 we were pleased to expand our Desktop team, which now includes a Desktop Team Manager to coordinate the team's work and an AV support position to provide AV support for meeting rooms, the media room and the Kendrew lecture theatre.

We updated documentation and made it more widely available and accessible to staff, and reviewed our procedures to ensure we provide a consistent response to user queries and to optimise OS configurations, software deployment and the provision of telecommunications for staff. In 2015 we supported the 950 client devices in use by 625 users across OSX, Windows & Linux.

Future plans

In 2016 we will participate in the migration of Ensembl from the Wellcome Trust Sanger Institute to EMBL-EBI. These services will additionally require virtualisation as they are currently run on physical servers and will be run on the EMBL-EBI's Virtual Infrastructure. The complex migration will be a collaboration between Ensembl and the Technical Services Cluster.

We will evaluate, qualify and potentially integrate commercial cloud infrastructures into the EMBL-EBI's infrastructure to provide capacity for running internal workloads. Addressing the need to update the institute's email system, we will evaluate and plan the implementation of a replacement email system or a full collaboration system in 2016.

We plan to complete a hardware refresh for the core virtual infrastructure and vCloud Embassy Cloud. We will be replacing the majority of the hypervisors with new servers of at least double the capacity of the existing servers. This will provide a saving in software licences, as each licence will provide double the resources to the infrastructure.

We will also procure hardware to increase the capacity of the OpenStack Embassy Cloud and the vCloud Embassy Cloud by approximately 100%. This will enable us to accommodate growing usage by EMBL-EBI staff and their external collaborators.

External Relations

As the role of bioinformatics in improving health and economic benefit becomes more prominent, so the task of engaging EMBL-EBI's diverse stakeholders takes on greater significance. Our team handles public relations and communications for the institute as a whole, engaging with diverse stakeholders on multiple platforms and in person. Our goal is to convey the value of EMBL-EBI to different audiences in a clear, comprehensible and professional manner.

We support the work of EMBL-EBI's many ambassadors, in particular the institute's Directors and team leaders, in fostering good relations with policymakers, funders, potential collaborators and service users throughout the world. We work with leadership to refine and deliver key messages, and host visiting delegations of scientists, politicians and industry representatives (12 delegations in 2015). We endeavour to raise the profile of the EMBL-EBI brand by generating high-quality content and disseminating it through the press, our top-level website, social media channels, newsletters and printed publications. We also provide editorial and graphic design support to individuals throughout the organisation, helping them raise awareness of EMBL-EBI data services.

Major achievements

On the occasion of Janet Thornton stepping down as Director of EMBL-EBI, all staff were invited to mark


In 2015 we welcomed major funders and policymakers to an event on board the research schooner Tara in London, where scientists presented the major achievements of the EMBL-led Tara expeditions and underscored the importance of sustained exploration of marine environments.

the occasion in a day of celebrations on the Wellcome Genome Campus. Our team planned and produced the event, collaborated with staff to deliver an engaging, visually compelling team challenge game and liaised with an array of vendors.

We organised, publicised and created media for a large-scale event for entrepreneurs in genomics called BioBeat15: Translating Genomics to Biobusiness. Supported by the Sex in Science Programme and in collaboration with BioBeat, the programme included over a dozen leaders in biobusiness and a keynote presentation by Chief Medical Officer for England, Professor Dame Sally Davies. We also organised an event for life-science funders and policymakers aboard the *Tara* research vessel in London, on its way to climate-change talks in Paris. This provided an opportunity to showcase EMBL's key role in the life sciences to an influential audience, bringing home the central role of EMBL-EBI in managing the data produced in large-scale, global scientific endeavours.

In 2015 we wrote and distributed over 30 press releases and other news stories, covering breakthrough research, new service launches and major developments. We supported writers throughout the organisation in a number of ways. We hosted 15 visiting journalists and film crews, distributed press releases directly to journalists, and signposted to news content on our website and the EMBL news site using social media platforms and e-mailed newsletters.

Our team is responsible for content on the main EMBL-EBI website, and is an active participant in the Web Content Committee. In this context we worked with other teams on a range of projects including the new events portal, and helped prepare for the next development phase of the institute's main website.

We are a hub for information about the institute, and work closely with the Scientific Services and the Strategic Project Management Office to ensure we gather and report the correct facts and figures about EMBL-EBI. In 2015 we created and distributed print and digital versions of the EMBL-EBI Annual Scientific Report, and contributed to the EMBL-wide Indicative Scheme in terms of both content and design.

Lindsey Crosswell

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.
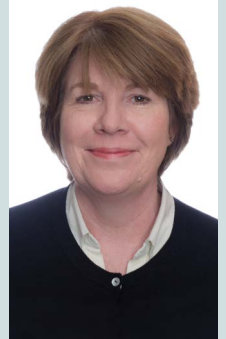
External Relations leads on brand and visual identity, and in 2015 we provided communications and branding support for a number of teams and projects in the form of web content consultation, key message development, graphics and slide presentations. We commissioned photographers to record key activities and reused the materials in various works. We designed printed materials for conference exhibitions, for example Plant and Animal Genomes, Intelligent Systems for Molecular Biology and BioData World Congress. We also contributed to the development of a new Wellcome Genome Campus brand, in collaboration with colleagues at the Sanger Institute and the Wellcome Trust.

When new staff arrive at EMBL-EBI, we take the opportunity to make them aware of our institute's mission and core values, familiarise them with our internal communications channels and point them to different sources of EMBL news and updates. In 2015 we inducted and photographed approximately 210 members of staff and visitors.

Our team participates in public engagement activities led by the Campus Public Engagement team, for example the Cambridge Science Festival, and in 2015 organised an event for secondary students in Stamford, UK called Big Biology Day.

Future plans

In 2016 our team will grow, enabling us to reach a broader audience and deliver different types of content through diverse channels, and helping the organisation analyse and act on issues relating to gender balance. We will continue to work with the Strategic Project Management Office to deliver documentation that is useful to our leadership and funders. We will work in the context of the Web Content Committee to improve the top-level EMBL-EBI website, and collaborate with the Web Development team and our colleagues in Heidelberg to define a digital content strategy that helps us all convey the excitement and impact of the organisation as a whole. As part of this work, we will help launch a new media asset management system to make the materials we create more readily accessible. External Relations will also take the lead on gender and diversity at EMBL-EBI, developing a strategic plan to bring improvements to gender balance..

A selection of imagery and graphic design, including a cover design for Genome Research, created in 2015 by Spencer Phillips, EMBL-EBI External Relations team.

Photos from BioBeat 2015: Translating Genomics to Biobusiness. Left: Chief Medical Officer for England, Professor Dame Sally Davies; centre: delegates during a Q&A session; right: Ruth McKernan, CEO of Innovate UK.s

Administration

The EMBL-EBI Administration Team facilitates the work of the institute by contributing to the EMBL-wide implementation of efficient administrative processes, enabling the effective deployment and development of resources within a complex regulatory environment.

Major Achievements

Strategic Project Management Office

Our newly formed Strategic Project Management Office supports on-going infrastructure projects, tracks and analyses the benefits and impact of EMBL-EBI activities on behalf of its funders, maintains and develops out project management framework, and supports best practices in procurement. In 2015 the team contracted a large-scale, independent analysis of the economic impact of the institute, and worked with the consultancy to carry out a diverse range of studies involving the input of thousands of users of our services, external collaborators and members of staff. One aspect of this work was to investigate the direct impact of EMBL-EBI on the worldwide pharmaceutical industry (see the EvaRIO Report, link), and another focused on quantifiable benefits to the global economy (see the Beagrie Report, link). The reports were finalised in 2015 and will be distributed in early 2016. These projects were the main focus of the team's activities during the year.

The Strategic Project Management Office also began to build a strong presence internally amongst project managers, who are distributed in diverse teams and welcomed efforts to build cohesion. An informal project management network now benefits from talks, informal meetings to share best practices, and a new, central source for project management tools, analysis and templates.

The team explored public procurement options available to both EMBL-EBI and other life-science organisations based in the UK for the ongoing Large Facilities Capital Fund project, with emphasis on meeting legal and value-for-money requirements. Thanks to these efforts, we are now in a position to adopt public frameworks rather than being required to renew and maintain bespoke frameworks.

Administration at EMBL-EBI in 2015. From left to right, back row: Muriel Cadilhac, Mark Green, James Whybrow, Pamela Swan, Menna Raafat, Brian Nsonga, Sheila Savill, James Smith, Frank O'Donnell, Christian Scherf (Administrative Director, EMBL), Charlotte Pearton, Julie Mace, Charles Shannon, Rebecca Sherry, Emma Sinha, Tracy Mumford. From left to right, front row: Agnieszka Egan, Rosa Heirman, Sara Troedsson, Julia Lant, Phyllida Hallidie (ELIXIR), Mary Barlow, Louise Morden, Sue Lee, Maria Bacadare Goitia, Christine Pettit, Hilary Little, Tracey Andrew, Debbie Howe, Lynn French.

Mark Green

EMBL-EBI Administration
Fellow of the Chartered Institute of Internal Auditors. At EMBL since 1997; joint appointment with EMBL-EBI.

At EMBL-EBI since 2003.

Finance

In 2015 we welcomed a member of the EMBL Budget Office to EMBL-EBI. This addition to our team enhanced the services we can provide to Group and Team Leaders, and improved the coordination of Finance activities with the EMBL Grants Office. The EMBL-EBI Grants Office put considerable efforts into entering historical grant data into the Converis reporting system, making the resource more useful, accessible and standardised. Thanks to this work we can now easily carry out meaningful queries about grants to EMBL-EBI staff over time.

The Finance and Purchasing Team has seen a steady increase in productivity, up approximately 10% over 2014 in terms of the number of orders and invoices processed. One of the team's larger projects in 2015 was the continued implementation of an "X-Flow" process, which has been adopted by EMBL for the automated processing of invoices.

Human Resources

We introduced flexible working arrangements for EMBL-EBI staff in 2015, and began monitoring its implementation and take-up. We also provided training as necessary for managers who carry out annual performance appraisals for their staff, a requirement introduced in 2014. We implemented the online absence recording system, precursor to the self-service portals to be introduced in 2016.

The Human Resources team managed a nearly 20% increase in recruitment during 2014. 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.

Cross-team collaboration

All of the Administration sub-teams examined processes and procedures in their immediate areas and in activities that contribute to the overall development of EMBL-EBI and EMBL. This led to improvements that are deeply appreciated by those whom administration serves.

Future plans

We will continue to develop benefit and impact analysis, and will release the results of work undertaken to our major stakeholders, including the general public. This will help shape the development of business cases related to EMBL-EBI infrastructure. We will also implement self-service portals for Human Resources data, introduce an online travel and expense system and improve our working processes and procedures.

European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus
Hinxton, Cambridge
CB10 1SD
United Kingdom

www.ebi.ac.uk
+44 (0)1223 494 444
comms@ebi.ac.uk

@emblebi
/EMBLEBI
EMBLmedia

EMBL member states:
Austria, Belgium, Croatia, Czeck Republic, Denmark, Finland, France, Germany,
Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Netherlands, Norway,
Portugal, Spain, Sweden, Switzerland, United Kingdom, Associate member states:
Argentina and Australia

EMBL-EBI is part of the European Molecular Biology Laboratory