

The European Bioinformatics Institute - Cambridge

Annual Scientific Report 2014

EMBL-EBI



On the cover

Protein Factory

Ribosomes (red), the protein factories of the cell, bind mRNA (green strands) which carries the instructions to make proteins. Amino acids, the building blocks of proteins, are carried to the ribosome via tRNA molecules (blue blobs). Protein (yellow chain & figures) pass from the Ribosome to carry out their roles within the cell.

Illustration by Spencer Phillips, EMBL-EBI.

Scientific consultation: Dr Matthew Conroy, EMBL-EBI.

© 2015 European Molecular Biology Laboratory

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

A digital version of the brochure can be found at www.ebi.ac.uk/about/brochures

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



Contents

Foreword	3
Major achievements	4
European coordination	8
Services	10
Genes, genomes and variation	12
Expression	16
Proteins and protein families	18
Molecular and cellular structure	20
Chemical biology	22
Molecular systems	24
Cross-domain tools and resources	26
Research	28
Support	36
Facts and figures	42
Funding & resource allocation	44
Growth of core resources	46
Collaborations	48
Our staff in 2014	50
Scientific Advisory Committees	52
Major database collaborations	56
Publications	58
Organisation of EMBL-EBI Leadership	68



Foreword

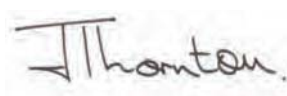
We are pleased to present EMBL-EBI's 2014 Annual Scientific Report, which provides our groups and teams with the opportunity to showcase the scope of EMBL-EBI's service and research activity, as well as the progress made in our core mission areas of industry collaboration and training.

2014 was a landmark year for EMBL-EBI and indeed for its parent organisation EMBL, who respectively celebrated the 20th and 40th anniversaries. EMBL held a range of events in Heidelberg, Hamburg and Monterotondo, and EMBL-EBI welcomed alumni and staff to an anniversary celebration at the Hinxton campus in June. This was a unique and memorable occasion to reflect on the huge strides made in bioinformatics in the two decades since our founding, and to look ahead in the company of alumni, friends and collaborators who have played a part in our history. The beautiful bioinformatics timeline produced as an historic record of EMBL-EBI's evolution features in part on the cover of this annual report.

The year was especially noteworthy for progress in the area of EMBL-EBI's relations with industry. Through our Industry Programme we have developed deep links with companies in the pharmaceutical and agricultural sectors, and in 2014, with the help of EMBLEM, we signed a public-private partnership agreement with GSK and the Wellcome Trust Sanger Institute to establish the Centre for Therapeutic Target Validation (CTTV). Housed in dedicated collaborative space in our South Building, scientists from the three organisations are working in an open, pre-competitive partnership, applying genomic techniques to the task of better defining targets in early-stage drug development. This is pioneering work and a fresh approach for industry, as the centre is committed to making the outputs from the collaboration available to all. We are excited about the prospect of other companies joining this partnership in 2015.

It is a testament to the wonderful support we continue to receive from our funders that EMBL-EBI is able to deliver an ever-growing range of data services and leading-edge bioinformatics research. The 'new normal' is EMBL-EBI's management of exponentially increasing data volumes against a challenging funding backdrop, and we extend our thanks to all those staff, collaborators and partners who make this possible.

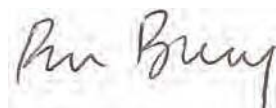
Sincerely,



Janet Thornton, Director



Rolf Apweiler, Joint Associate Director



Ewan Birney, Joint Associate Director

Major achievements 2014

Data-driven science has come of age as high-throughput technologies have matured, giving rise to a second wave of genomics with variation at the fore.

The major achievements we report here highlight successes in 2014, but the common thread uniting them all is our consistent, sustained delivery of data services that enable research throughout the world, combined with bioinformatics research to help to interpret new data.

As we celebrated our 20th year in operation, we welcomed a new kind of collaboration with industry: The Centre for Therapeutic Target Validation (CTTV), an initiative that brings all knowledge to bear on the economically, socially and scientifically important issue of identifying and validating drug targets. As part of our Innovation and Translation programme, we developed an open, pre-competitive public-private partnership with GSK and the Wellcome Trust Sanger Institute to reduce the time it takes to develop a new medicine by improving the definition of targets in early-stage drug development. Associate Director Ewan Birney played a key role as Interim Head in the first year of this project as it defined its scientific direction, and in 2015 will be succeeded by Jeffrey Barrett.

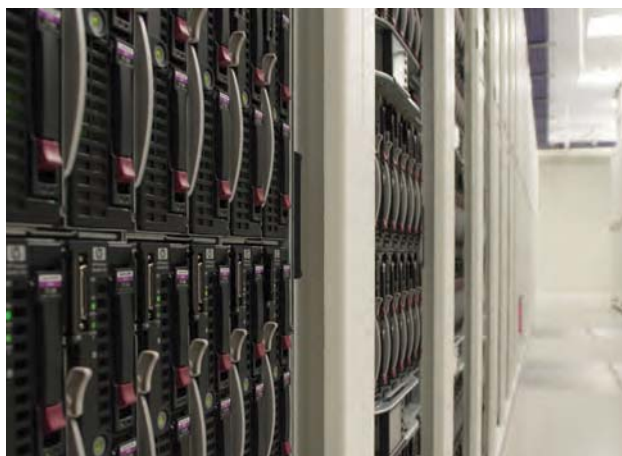
Another major achievement in 2014 was the migration of our public-facing data centres from London to Hemel Hempstead in the UK, a massive undertaking led by the Systems Infrastructure team and completed without any unplanned downtime to any EMBL-EBI service. In fact, during September, when a large part of the work was carried out, our Job Dispatcher service ran close to 12 million jobs for our users without any significant interruption. The success of this endeavour was thanks to the efficiency and commitment of our Systems and Networking team, and to the thorough planning initiated in 2009, when we originally organised the systems to run in two London-based data centres.

Services

The launch of our European Variation Archive (EVA) this year reflected the increasingly widespread availability of detailed, high-resolution genetic variation data from humans and many other species, including plants. The EVA is the first archival resource at EMBL-EBI to provide a single access point for submissions, archiving, and access to high-resolution variation data of all types. At its launch, the resource contained 1.7 billion submitted variants from large-scale efforts including the 1000 Genomes Project, Exome Variant Server, Genome of the Netherlands Project and UK10K. Agriculturally relevant species including sheep, cow, maize and tomato followed soon after, extending the EVA's utility beyond the biomedical domain.

As the world-wide Protein Data Bank welcomed its 100,000th structure deposition since its inception in 1971, our Protein Data Bank in Europe team prepared to handle the increasing size and complexity of submitted data. New hybrid methods are revolutionising structural biology and will help to determine the shapes of large biological complexes that were hitherto inaccessible. As part of the wwPDB, we laid the groundwork for supporting these developments and sharing the associated data.

Cryo-Electron Microscopy is part of this revolution, and as with all new technologies it will take time for a 'gold-standard' method for interpretation to emerge. In 2014 we launched EMPIAR: a new resource for raw, 2D electron microscopy images that is enabling this community to combine their efforts and reach consensus in this area. Built on the PDBe infrastructure, these extremely large datasets are helping researchers examine 2D images in extremely high resolution, offering opportunities to develop better methods to improve the quality of the 3D molecular structures they ultimately produce.



Genomes

Raw nucleotide sequences provide a strong foundation for our data resources, which add many layers of knowledge that are invaluable to researchers working in the biomedical, agricultural and ecological sciences.

In 2014 our Ensembl team incorporated a vast amount of knowledge into the newly refined and fully annotated reference human genome, GRCh38. As this reference human genome begins to be used in clinical or diagnostic settings, the accuracy of the assembly will be paramount and must be meticulously maintained and updated. Ensembl makes the new assembly easily accessible to global initiatives including the Blueprint project and the 1000 Genomes Project.

One of the first outputs of the Global Alliance for Genomics and Health was the new Genomics API: an open-source software that allows researchers to share anonymised genetic data seamlessly across platforms and collaborate on a global scale. This useful tool – the first of what we hope will be many – was the result of a large-scale, international, public–private technical collaboration.

The most complete version of the whole-genome dataset for bread wheat entered the public domain through Ensembl Plants in 2014. Generated under the auspices of the International Wheat Genome Sequencing Consortium (IWGSC), the data are providing insights into how wheat was domesticated and how traits relating to pest resistance and stress tolerance have developed – vital knowledge that will help plant breeders improve crops. Similarly the new reference sheep genome, published by the International Sheep Genomics Consortium (ISGC) and made freely available through the Ensembl genome explorer provides the most precise genomic data from ruminant species yet. This can be used to pinpoint which genes can express specific features, for example lanolin production and related research, could have a real impact for rural economies.

Our ten-year collaboration with researchers in Kenya bore fruit in 2014 with the publication of the tsetse fly (*Glossina morsitans*) genome, which is available through VectorBase. This genome adds to a substantial body of knowledge about this pest, which transmits sleeping sickness to humans and animals. Over 140 insect disease vector biology scientists – half of whom work in African research institutes – examined and manually curated the genome annotations, based on their specialist knowledge in all aspects of tsetse fly biology, from sense of smell to reproduction and immunity. The result is a foundational resource for research into new ways of controlling the spread of sleeping sickness.



Research

Our research groups use computational approaches to understand life at a fundamental level, usually in close collaboration with external wet-lab groups. Two of our research groups operate both ‘wet’ and ‘dry’ components and in 2014 produced some truly ground-breaking work.

The Bertone group, in collaboration with their colleagues in Germany, Japan and elsewhere in the UK, resolved a long-standing challenge in stem cell biology by successfully ‘resetting’ human pluripotent stem cells to a fully pristine state, at the point of their greatest developmental potential (Takashima et al., *Cell*, 2014). The discovery paves the way for the production of superior patient material for translational medicine. Reset cells mark a significant advance for human stem cell applications, such as drug screening of patient-specific cells, and are expected to provide reliable sources of specialised cell types for regenerative tissue grafts.

The Teichmann group uncovered a fundamental mechanism for regulating a protein’s shape (Perica et al., *Science*, 2014). To determine how a protein morphs from an active (RNA-binding) configuration to an inactive one in two very different environments, they looked at a family of bacterial RNA-binding proteins that control a basic process in metabolism. They found that this morphing process is controlled by mutations – not at the site where the binding happens, but at a distance, where they act indirectly to change the protein’s shape. Undertaken initially as a purely computational study, the work involved experiments in biophysics and structural biology as well as elastic network modelling. The findings have implications for the manipulation of proteins, with potential applications in biotechnology and drug development.

Major achievements 2014



In a purely computational setting, the Stegle group created a powerful new method that vastly improves the accuracy of Genome-Wide Association Study (GWAS) analyses (Fusi et al., *Nature Commun*, 2014). The new method, WarpedLMM, automates processes previously carried out manually by individuals, and makes it possible to attribute a greater proportion of phenotypic differences between individuals to genetics. WarpedLMM is a practical improvement on the widely used Linear Mixed Models, and can help create a more accurate picture of the genetics of quantitative traits.

Taking a truly interdisciplinary approach to a common problem in protein research, the Thornton group created software that makes it possible to quickly compare a potentially useful enzyme against thousands of well-known reactions (Rahman et al., *Nature Methods*, 2014). The new software, EC-BLAST, replaces an arduous manual process. The computer scientists, chemists, biologist and physicist who developed EC-BLAST spent five years working through roadblocks to quantifying the comparison of enzymes already well described by the Enzyme Commission (EC). The interdisciplinary approach was vital to ensuring the final product was fit for purpose.

The Brazma research group, as part of the International Cancer Genome Consortium, analysed data from large-scale DNA and RNA sequencing of renal cell carcinoma patients in Europe and achieved insights into the genetic architecture of clear-cell renal-cell carcinoma (Scelo et al., *Nature Commun*, 2014). The consortium found a clear association between cancer incidence and exposure to aristolochic acid – an ingredient in some herbal remedies. These findings have implications for public health, particularly in Romania.

European coordination

EMBL-EBI is a pivotal partner in large-scale collaborations with global impact, contributing broad scientific and technical expertise as well as essential data infrastructure for research.

Perhaps most importantly, we place a high value on community building, so that people working in different areas can identify shared challenges and work together to overcome them.

One significant development in this area in 2014 was the establishment of ELIXIR as an independent entity, with Niklas Blomberg as Founding Director. EMBL-EBI hosts the ELIXIR Hub in our South Building, and we maintain close links with the infrastructure in many ways, both formal and informal.

ELIXIR pilot: marine metagenomics

In 2014, EMBL-EBI participated in three new ELIXIR pilot actions, one of which focused on the emerging area of marine metagenomics. Large-scale sampling of the Earth's waters, for example by MicroB3's Ocean Sampling Day and the Tara Oceans expedition, have begun to reveal in detail what little we know of the species inhabiting the planet.

As this important area of science matures, it is critical that we make sure the data infrastructure is equipped to handle new flows of data and that communities can agree on best practice. Our role in building consensus in different scientific communities is very much at the fore in this regard, as marine metagenomics research is rapidly gaining traction in biotechnology and environmental research.



BioMedBridges

EMBL-EBI coordinates the BioMedBridges project on behalf of ELIXIR, with many service teams contributing their expertise to improve access, security and interoperability between 12 biomedical sciences research infrastructures. BioMedBridges contributes greatly to community building, and in 2014 the partners facilitated the agreement data standards, and provided a secure framework for sharing sensitive data ethically and legally.

ProteomeXchange

In 2014 the ProteomeXchange consortium launched a public portal for exchanging proteomics data and information, highlighting a growing trend in the sharing of data generated in mass-spectrometry experiments. ProteomeXchange integrates resources such as PRIDE, developed in our Proteomics Services team, and the PeptideAtlas, developed in the US, providing a single point of entry to a range of public proteomics resources and making these valuable datasets more discoverable and easier to reuse.

Integrating RNA data resources

EMBL-EBI brings communities together to agree on the best way to share data and ultimately enable scientific progress on a large scale. RNAcentral, the first unified resource for all types of non-coding RNA data, provides an invaluable resource by combining the strengths of over 17 member databases, with many more on the way. Its launch in 2014 was the culmination of many years of developing standards and data-sharing agreements in this pivotal community, and the result is a public resource of fundamental utility.

Discovery in context

In 2014 Europe PMC became the most highly accessed website at EMBL-EBI. This rapidly growing resource helps people explore the scientific literature in new ways, providing deep links to its underlying molecular data and enabling serendipitous discovery. The Literature Services team, which leads the development of Europe PMC, ensured that it delivered consistently to over 1 million unique IP addresses every month, with negligible down time. Developed jointly between EMBL-EBI, the University of Manchester and the British Library, Europe PMC has become the world's largest citation network and links over 3 million full-text articles with their underlying datasets.

Looking ahead

As this Annual Scientific Report goes to press, EMBL Council has approved the appointment of two new directors of the institute in mid-2015: Ewan Birney and Rolf Apweiler. They will be stepping into their new roles at a pivotal moment in biology, as we begin the transition from understanding molecules to characterising cells, organelles and systems quantitatively. New and advancing technologies are allowing us to pinpoint molecules in whole cells in real time using technologies such as 3D Electron Microscopy at extremely high (3Å) resolution. The integration of molecular and cellular imaging data is in sight. In parallel, the determination of personal genomes allows us to understand differences between individuals and makes these data relevant in the clinic. To say that this is an exciting time for bioinformatics would be akin to saying the moon landing was a bit of an adventure.

This is a time of great promise for life scientists of all walks and disciplines, as technologies have matured to a point where many practical distractions are removed, and we can concentrate firmly on asking creative questions and being confident in the quality of the answers. EMBL-EBI is at the middle of it all, privileged to collaborate with scientists and engineers the world over to turn data into knowledge that can truly benefit all species, including humankind.



European coordination

An important part of our mission is coordinating bioinformatics in Europe so that researchers can make the best possible use of public data and infrastructure for the life sciences.

EMBL-EBI is a pivotal partner in large-scale collaborations with global impact, as we contribute broad scientific and technical expertise as well as essential data infrastructure for research. Perhaps most importantly, we place a very high value on community building, so that people working in different areas can identify shared challenges and work together to overcome them.

Scientific literature: discovery in context

It is a testament to the profound utility of Europe PMC that the World Health Organization (WHO) and the European Research Council (ERC) have become major funders of Europe PMC, which in 2014 became the most highly accessed website at EMBL-EBI. This rapidly growing resource helps people explore the scientific literature in new ways, providing deep links to its underlying molecular data and enabling serendipitous discovery. The Literature Services team, which leads the development of Europe PMC, ensured that it delivered consistently to millions of unique IP addresses during the year, with negligible down time. Europe PMC has become the world's largest citation network, and links over 3 million full-text articles with their underlying datasets.

RNAcentral: at last

EMBL-EBI brings communities together to agree on the best way to share data and ultimately enable scientific progress on a large scale. RNAcentral, the first unified resource for all types of non-coding RNA data, provides an invaluable resource by combining the strengths of over 17 member databases, with many more on the way. Its launch in 2014 was the culmination of many years of developing standards and data-sharing agreements in this pivotal community, and the result is a public resource of fundamental utility.

The rise of open proteomics

In 2014 the ProteomeXchange consortium launched a public portal for exchanging proteomics data and information, highlighting a growing trend in the sharing of data generated in mass-spectrometry experiments. ProteomeXchange integrates resources such as PRIDE, developed in our Proteomics Services team, and the PeptideAtlas, developed in the US, providing a single point of entry to a range of public proteomics resources and making these valuable datasets more discoverable and easier to reuse.

ELIXIR pilot actions

In 2014, ELIXIR was established as an independent entity, with Niklas Blomberg as Founding Director. EMBL-EBI hosts the ELIXIR Hub in our South Building, and we maintain close links with the infrastructure in many ways, both formal and informal.

We participated in three ELIXIR pilot actions in 2014: Interoperable, controlled-access big data transfer; Integration of raw data repositories using e-infrastructure services; and Marine metagenomics: putting users first.

One of the most dynamic areas of science in 2014 was marine metagenomics, in which large-scale sampling of the Earth's waters (e.g. MicroB3's Ocean Sampling Day and the Tara Oceans expedition) began to reveal in detail what little we know of the species inhabiting the marine world. Making sure the data infrastructure is equipped to handle new flows of data and that communities can agree on best practice are absolutely critical to ensuring the long-term impact of these studies. EMBL-EBI's role in building consensus in different



scientific communities is very much at the fore in this regard, as marine metagenomics research is rapidly gaining traction in biotechnology and environmental research.

In this collaboration with ELIXIR Norway and other partners, we began to build a dedicated data-management e-infrastructure and bioinformatics pipelines to support marine research. In 2014 we started to harmonise existing pipelines, develop new components and improve established tools in order to establish service platforms that are sustainable in the long term. Perhaps most importantly, this pilot action provides a platform for strengthening Europe's user community for marine metagenomics analysis.

The Interoperable, controlled-access big data transfer action builds on our collaboration with the Centre for Genomic Regulation (CRG) in Spain to develop the European Genome-phenome archive (EGA). In 2014 we resolved several important limitations to computing, network bandwidth and storage buffer areas that affect the delivery of EGA data. The result of this work is a suite of data transfer solutions that can be applied throughout ELIXIR.

As part of the ProteomeXchange consortium, our participation in the Integration of raw data repositories using e-infrastructure services action led to substantial progress in connecting national data-storage services and international repositories for raw mass-spectrometry proteomics data through ELIXIR. Led by Bioinformatics Services to Swedish Life Science (BILS) and using the EUDAT European infrastructure, this pilot action demonstrates how collaboration among research infrastructures and e-infrastructures can optimise the management of growing volumes of data.

BioMedBridges

EMBL-EBI coordinates the BioMedBridges project on behalf of ELIXIR, with many service teams contributing their expertise to improve access, security and interoperability between 12 biomedical sciences research infrastructures. BioMedBridges tackles interoperability challenges directly, removing technical stumbling blocks related to interoperability of data from a variety of disciplines and scales, from molecular through cells, organs, organisms and the environment. It also works to harmonise data on different species; between different technologies, from nanotechnology through spectroscopy to synchrotrons; and across different research communities that have not traditionally worked closely together.

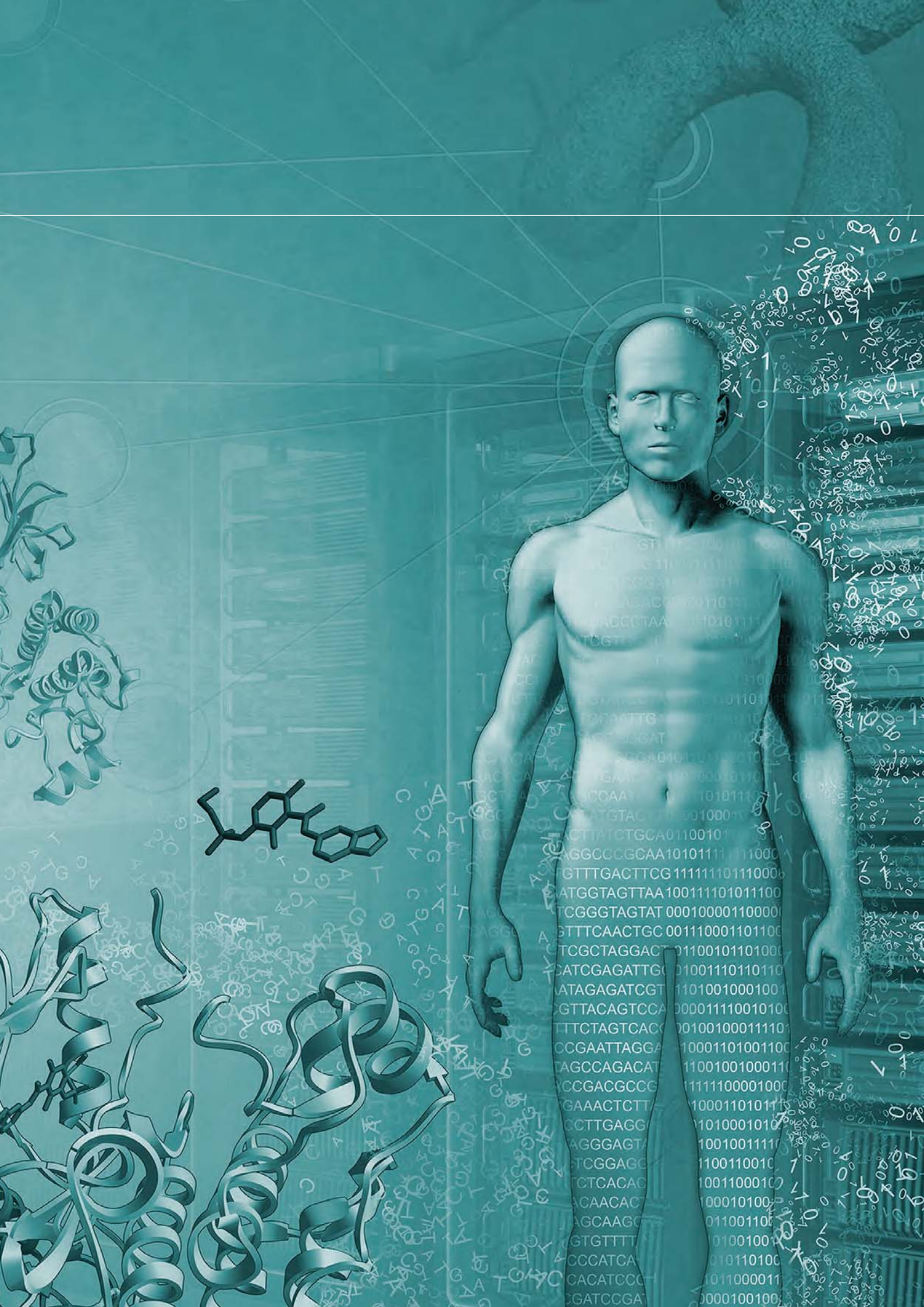
BioMedBridges contributes greatly to community building, and in 2014 the partners provided a secure framework for sharing sensitive data ethically and legally. In a series of workshops, BioMedBridges partners and experts from outside the project discussed the development of data standards to serve different scientific communities, opportunities for support from the European e-infrastructures, and tools to support researchers with questions on ethical and legal requirements around sensitive data. In the project's 2015 workshop, to be held in November at the newly refurbished Wellcome Genome Campus Conference Centre, we will look back on what the project has achieved since its inception, and look to the challenges ahead.





Services







Genes, genomes and variation

Our resources that focus on genes, genomes and variation data represent the largest service cluster at EMBL-EBI.

In 2014 we launched two major resources in this domain: The European Variation Archive (EVA) and RNACentral. The launch of the EVA marks the modern era in molecular biology, when we can look at genetic variation among individual animals, plants and other life forms at an unprecedented scale. RNACentral represents a major effort by many communities to create a framework for sharing data and knowledge about noncoding RNAs.

Other major developments include Ensembl's incorporation of a vast amount of knowledge into a fully annotated reference human genome, GRCh38. Their work builds on the release of a new assembly by the Genome Research Consortium, and provides a solid foundation for future genomics research. The European Nucleotide Archive, a foundational resource that underpins many EMBL-EBI endeavours, brought together diverse marine-research communities including Tara Oceans and Ocean Sampling Day to establish and implement data standards, coordination and publishing. This work is essential for the large-scale delivery of marine genomics and metagenomics studies.

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

Ensembl produces and maintains evidence-based automatic annotation and incorporates manually curated annotation on selected eukaryotic genomes. Our annotation is based on mRNA, protein and functional genomics information. Ensembl provides valuable insights into variation within and among species, and allows users to compare whole genomes to identify conserved elements. It is integrated with several other important molecular resources, including UniProt, and can be accessed programmatically. Ensembl collaborates with genome sequencing centres and annotation groups around the world.

www.ensembl.org

Ensembl Genomes

The falling costs of DNA sequencing have led to an explosion of reference genome sequences and genome-wide measurements and interpretations. Ensembl Genomes provides portals for bacteria, protists, fungi, plants and invertebrate metazoa, offering access to genome-scale data through a set of programmatic and interactive interfaces, exploiting developments originating in the vertebrate-focused Ensembl project. Collectively, the two projects span the taxonomic space.

www.ensemblgenomes.org

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, variant data from human and other species. All users can download data from accessioned studies, or submit their own data to the archive. It is also possible to query all variants in the EVA by study, gene, chromosomal location or dbSNP identifier using our VCF Browser.

www.ebi.ac.uk/eva

European Genome-phenome Archive (EGA)

The EGA 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

Metagenomics portal

Our Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data, and provides insights into both the taxonomic composition and functional/metabolic potential of a sample.

www.ebi.ac.uk/metagenomics

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>

RNAcentral

RNAcentral is a federated database of non-coding RNA sequences allowing a unified view of RNA sequence data. RNA sequence data is provided by the RNAcentral Expert Databases and by user submissions via the European Nucleotide Archive.

<http://rnacentral.org/>

1000 Genomes

The 1000 Genomes Project is a fully open resource consisting of nearly 2500 sequenced individuals from five major world populations (Europeans, East Asians, Americans, Africans and South Asians). It is designed to capture and provide all common (>1%) genetic variation. Data is made freely and openly available in advance of publication.

<http://1000genomes.org>



GRCh38



Genes, genomes and variation

European Nucleotide Archive

Team leader: Guy Cochrane

- *Provided a new assembly-submission system that reduces turnaround time for a large proportion of submissions from several weeks to a couple of days;*
- *Added further functions and streamlined the Webin submission system, supporting new workflows, platforms and data types;*
- *Developed a new cross-reference system supporting the integration of all ENA data types with external data resources;*
- *Collaborated with marine-research communities such as Tara Oceans and Ocean Sampling Day to establish and implement data standards, coordination and publishing;*
- *Provided on-going data management and content support for EBI Metagenomics, RNACentral and other EMBL-EBI data resources;*
- *Redesigned the ENA web site and added new functionalities for discovery, browse and retrieval services;*
- *Explored cloud delivery of ENA content for different user communities, including a read-level API in support of the Global Alliance for Genomics and Health;*
- *Advanced the CRAM sequence-data compression technology and released new incremental versions of the software toolkit.*

www.ebi.ac.uk/ena

www.ebi.ac.uk/ena/submit

www.ebi.ac.uk/ena/about/cram_toolkit

www.ebi.ac.uk/ena/submit/checklists

www.ebi.ac.uk/ena/data/view/PRJEB402

Vertebrate Genomics

Team leader: Paul Flicek

- *Issued four major releases of Ensembl, including one that annotated the new human genome assembly, GRCh38, and provided updates to other highly used resources;*
- *Launched an improved version of our Variant Effect Predictor (VEP), which now includes a new web interface providing additional information;*
- *Formally released the Ensembl REST server as a new programmatic interface for the project, and published WiggleTools for computing statistics across large, genome-wide datasets;*
- *Expanded Ensembl's collaboration with the Global Alliance for Genomics and Health (GA4GH) and started working with the Functional Annotation of Animal Genomes Project (FAANG);*
- *Issued two public releases of data for the BLUEPRINT project, which in 2014 published its first findings in three papers published in Science;*
- *Developed and published GWAVA, a new method for annotating non-coding variants in human genomes;*
- *Reported the first comprehensive results comparing the evolutionary changes between transcription factor binding and gene expression in mammals.*

www.ensembl.org

www.blueprint-epigenome.eu

www.hipsci.org

Non-vertebrate Genomics

Team leader: Paul Kersey

- Released the first version of RNAcentral, a new resource for information about non-coding RNAs;
- Issued three public releases of Ensembl Genomes, and contributed to the regular data releases of VectorBase, WormBase and PomBase;
- Added the genomes of four invertebrate metazoans, 12 plants, nine fungi and eight protists to the public release;
- Increased the number of bacterial genomes available through the Ensembl public interface from under 10,000 to over 20,000;
- Made major contributions to a number of genome analysis papers, including the first release of the genome sequence of tsetse fly, *Strigamia maritima* (a centipede), the Glanville fritillary butterfly, and release of 16 different species of *Anopheles* mosquito;
- Improved the representation of the polyploid bread wheat genome, linking evolutionary histories to supporting alignments, and adding substantial datasets concerning polymorphism and internal variation;
- Released the first version of WormBase ParaSite, a new resource providing rapid access to the latest sequence data from parasitic worms.

www.ensemblgenomes.org
<http://rnacentral.org>

Variation Archive

Team leader: Justin Pascall

- Launched the European Variation Archive (EVA) in full production mode, providing a single access point for submissions, archiving, and access to variation data of all types;
- Expanded the EVA taxonomic coverage to include 11 species;
- Progressed the European Genome-phenome Archive (EGA) collaboration with CRG Barcelona to the point where we jointly host over half a petabyte of data.

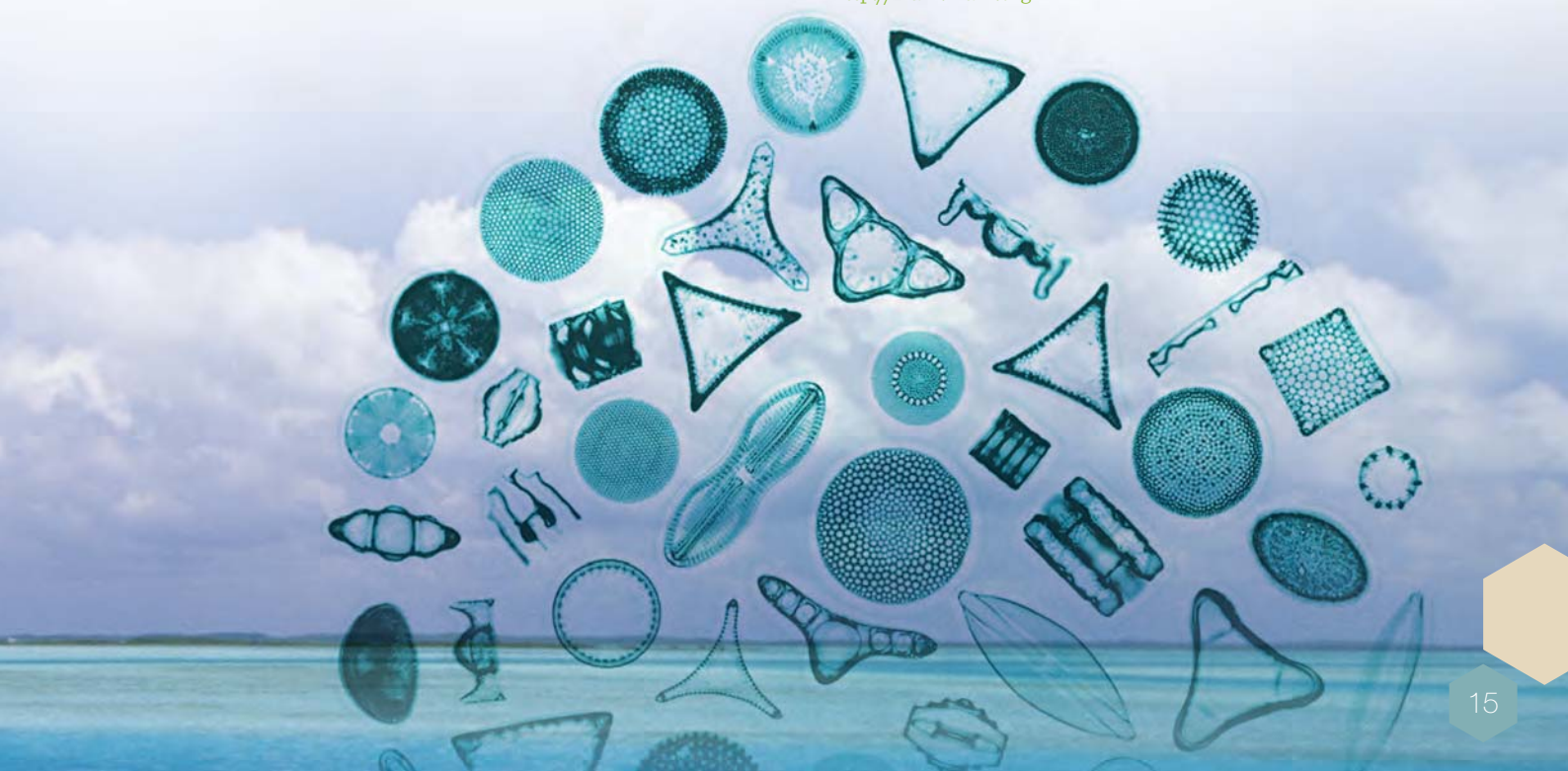
www.ebi.ac.uk/ega
www.ebi.ac.uk/dgva

Protein families

Team leader: Rob Finn

- Updated the Rfam database production pipelines and algorithms;
- Increased throughput of the EBI Metagenomics analysis pipeline, processing over 20 billion nucleotide sequences in 2014;
- Developed the EBI Metagenomics comparison tool, allowing users to compare GO terms for samples within metagenomics studies.

www.ebi.ac.uk/metagenomics
<http://rfam.xfam.org>





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 of biological systems 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 that supports several EMBL-EBI data resources. We are committed to the utility of the ENA platform and to achieving the broadest reach and utility of sequencing technology and data.

Major achievements

One of our core activities in 2014 was the extension and adaptation of the Webin data-submission system. As a key component of the ENA platform, Webin is under a development programme in which we track and adapt to the changing needs of the user community. We also enhance usability for data submitters working in diverse domains and make every effort to support sustainable biocuration, in which all data flow from submitters into ENA through deep and consistent validation.

We completed a major new workflow for assembly data submissions. Effectively reducing the time taken to turn assembly submissions around (we have achieved reductions from several weeks to 2-3 days for half of the assemblies received), the service leads the submitter through a rapid, intuitive process to provide data in a choice of formats for multiple assemblies at a time, with support for functional annotation. Our work on functional annotation submission workflows allowed us to streamline the system, for example by removing a major legacy software component. This improved the consistency of the user experience and reduced human involvement in the submissions process. We added a number of sample checklists: data structures that define information

to be reported and validation to be applied. Examples include checklists for Ocean Sampling Day (MicroB3), Tara Oceans and Pathogen Surveillance (GMI:MDM) data. We also introduced support for new platforms, formats and data types. We now accept Oxford Nanopore data, offer extended CRAM submissions support and accept tabular taxonomic identification data from diversity studies.

A richly annotated sample record from the Tara Oceans dataset. Background: summary information and links to read data files. Inset: attribute list showing geospatial, oceanographic, biome descriptions, sampling information and further details.

Synchronising global endeavours

We continued with our data-coordination and -management activities, leveraging the services provided by the ENA platform. We focused in particular on the marine domain where, as contributors to the Ocean Sampling Day and Tara Oceans projects, we led data-standardisation activities, curated sample records and coordinated complex flows of data around multi-disciplinary consortia. We also intensified work on community reporting standards for sequence-based pathogen surveillance in the context of the COMPARE project, launched in December.



In 2014 we designed a new cross-referencing infrastructure for ENA and put it into production. This infrastructure allows batches of cross-references to be curated and integrated into appropriate browser views. The complete set of ENA data types for which an identifier exists can now be cross-referenced, with newly supported data types including 'taxon', 'non-coding' and 'assembly'. Because the new system allows cross-reference curation that is independent of other ENA data production cycles, cross-references can be curated as they arise, rather than saved up for bulk processing. Another new feature supports the annotation of cross-references with additional fields of information that summarise or report key characteristics of the remote record. This functionality will be useful for distributed projects such as RNACentral.

Improving discoverability

To make our data and services more discoverable, we redesigned the ENA website and introduced new functionality in the browser, its associated web services and data download. This includes richer indexing and direct user editing of queries in Advanced Search. We also integrated a number of new markers into our Marker Portal, which provides an entry point to data arising from the sequencing of marker loci across multiple species and strains. To improve the presentation of data, we introduced support for the OpenSearch protocol and added Geographical Markup Format 3 (GML3) endpoints. We also added a GlobusFTP end point for data download.

Embassy Cloud

As part of our ongoing efforts to support global scientific collaborations, we explored cloud technologies in the delivery of ENA. We operate a number of instances of the EMBL-EBI's Embassy cloud in which ENA datasets have been made available. Amongst these are: Embassy workspaces supporting analytical work on Tara Oceans data, a prototype 'Read API' for the Global Alliance for Genomics and Health (GA4GH), low-level access provision to all public CRAM-formatted data available from ENA and an early deployment of a compute environment for routine analytical workflows around pathogen surveillance data.

Future plans

Our focus areas in 2015 will be sustainable biocuration, data-submission workflows, reference-assembly services and support for pathogen-surveillance activities. Under the sustainable biocuration model, expert scientific knowledge impacts whole classes of data. Our activities will include the biocuration of sample and annotation checklists, structured vocabularies and validation rules, all of which are used to configure Webin for the capture and validation of data. We will develop the technologies necessary for these biocuration tasks, including an enhanced editing environment for sample

checklists that ensures consistency between checklists as their numbers grow, and support for external ontologies in ENA content.

We will introduce new submission workflows into Webin that will offer researchers a smoother and more integrated submission process. To do this, we will develop a workflow in the interactive web tool for the reporting of unclassified taxa, integrate all annotation-submission workflows and provide broader support for programmatic submissions of annotation data.

As reference assemblies continue to grow in importance and number, we will build reference-centric services around ENA content. For example, we will develop tools that allow reads aligned to a given reference to be discovered based on the reference and its characteristics.

As part of the COMPARE project, which aims to speed up the detection of and response to infectious disease outbreaks, we will create ENA-based services to support pathogen-surveillance activities. Shotgun-sequencing methods have become viable for routine and responsive detection and analysis of pathogenic organisms. These data must be captured and shared rapidly if they are to support outbreak detection and epidemiological investigation. Working with the appropriate communities, we will further develop standards and workflows and implement these in our Webin submission system. We will provide data coordination and management services and roll out analytical workflows using the Embassy platform.

Selected publications

Silvester N, Alako B, Amid C, et al. (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucl Acids Res* 43:d23-d29

The RNACentral Consortium (2015) RNACentral: an international database of ncRNA sequences. *Nucl Acids Res* 43:d123-d129

Hunter S, Corbett M, Denise H, et al. (2014) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucl Acids Res* 42:d600-d606

Federhen S, Clark K, Barrett T, et al. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand Genomic Sci* 9:1275-1277



Vertebrate Genomics

The Vertebrate Genomics team creates and maintains the genomic resources of the Ensembl project and is responsible for data management for a number of large-scale projects including BLUEPRINT. Other resources include the International Genome Sample Resource, which incorporates the data of the 1000 Genomes Project, and the curated GWAS catalog. All of these resources are publicly available and are widely used by the scientific community and by the team itself as part of our research into evolution, epigenetics and transcriptional regulation.

Our specific research projects focus on genome annotation as well as the evolution of transcriptional regulation and understanding tissue specificity. 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 included clarifying the role of CTCF, cohesin and other genomic structural proteins.

Major achievements

Ensembl

A major task for Ensembl in 2014 was the annotation and release of the new human genome assembly, GRCh38. The new assembly has new centromere representation, better segmental duplication accuracy, fewer assembly gaps and thousands of corrected rare bases – changes that will be particularly valuable for clinical genomics. We also issued four major releases of Ensembl, which include updates to the resources for the model species mouse, rat and zebrafish.

We now offer new visualisations of comparative genomics datasets on the Ensembl website: a new resource called 'Age of Base' estimates the evolutionary timing of the most recent mutation within the human genome, and secondary structures of non-coding RNAs can also now be displayed. In addition, we deployed a new BLAST search service. We improved the Variant Effect Predictor (VEP): a new web interface that performs an 'instant' calculation for single variants, improved pie charts and summaries of results and extra data types to output such as allele frequencies from the 1000 Genomes Project.

We transitioned our REST server, the highly accessed new programmatic interface for Ensembl, from a beta version to a fully supported service that is hosted within EMBL-EBI's Embassy cloud. Ensembl was also involved in several major genome-sequencing consortia that published results in 2014. These included the primates gibbon and marmoset in the trees, the cat, ferret, rabbit, sheep on land and blind cave fish and African cichlid fish in the water.

Data management and coordination

We issued two public releases of data as part of BLUEPRINT, an EU project to decipher the epigenomes of more than 100 different types of blood cells. In 2014, BLUEPRINT published its first findings and these provide valuable insights into how the immune system keeps infection at bay. BLUEPRINT has also established and shared reference epigenomes for some of the most abundant white cells in venous blood, cord blood and bone marrow.

We continued to lead data coordination in the HipSci project, which aims to create a catalogue of human induced pluripotent stem cells (iPSCs). We also became a central partner in FAANG, the Functional Annotation of ANimal Genomes Project, an international effort launched in 2014 to accelerate the production of comprehensive maps of functional elements in the genomes of domesticated animal species.

Research

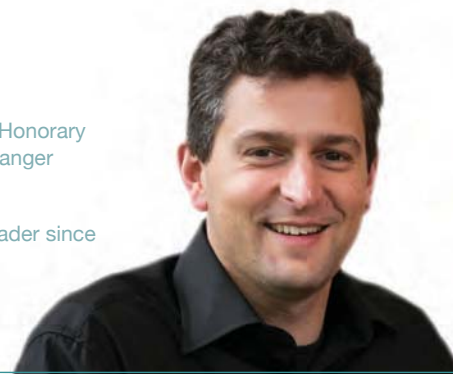
A number of major research projects came to fruition in 2014. EMBO Postdoctoral Fellow Emily Wong led work on the evolutionary coupling between transcription factor binding and gene expression in mammals. The results highlight a significant tolerance to evolutionary changes in transcription factor binding intensity in transcriptional networks in mammals; they also suggest that some transcription factor-dependent genes may be regulated mainly by a single transcription factor across evolution. Emily was awarded one of five EMBO Advanced Fellowships.

Graham Ritchie, an ESPOD postdoc jointly with Ele Zeggini's group at the Wellcome Trust Sanger Institute, published GWAVA, a tool integrating various genomic and epigenomic annotations to prioritise noncoding variants. He also participated in a number of published projects that aimed to find connections between genotype and phenotype.

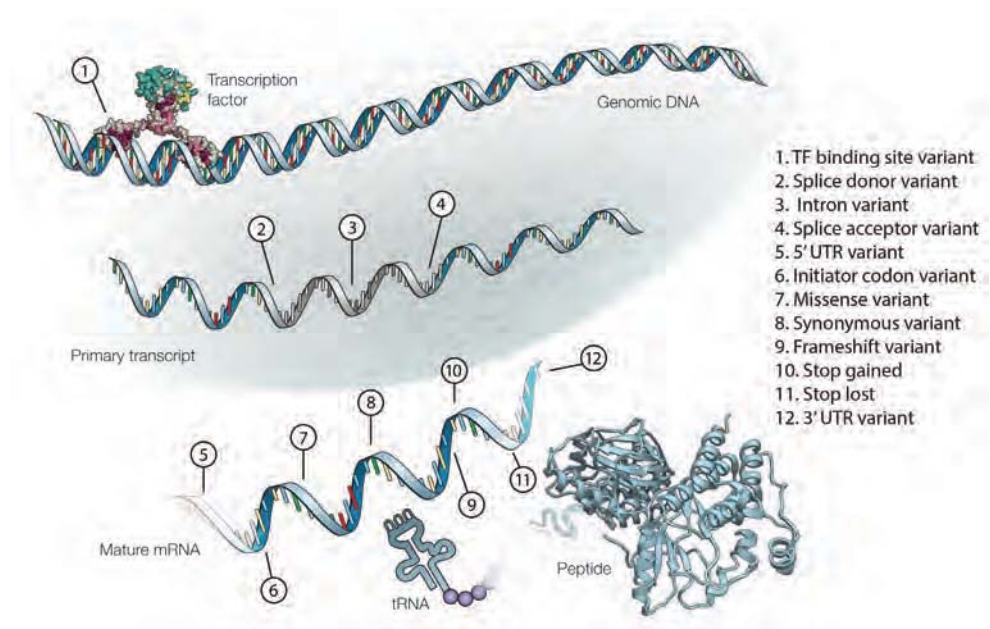
Paul Flicek

DSc Washington University, 2004. Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008.

At EMBL-EBI since 2005. Team Leader since 2007, Senior Scientist since 2011.



A set of annotation terms used to describe the potential effects of sequence variants according to the genic regions they fall in and their allele sequences. The terms are drawn from the Sequence Ontology and are depicted on the molecules they are predicted to affect. Variants categorized as any of the terms 2, 4, 9 and 10 are often collectively referred to as 'loss-of-function' variants, and are typically expected to severely affect gene function.



Our collaborative research with David Spector's group at Cold Spring Harbor Laboratory in New York and EMBL-EBI Research Group Leader John Marionni 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 2015 we will establish the International Genome Sample Resource (IGSR). The IGSR will maintain the data collections from the 1000 Genomes Project, and enable the addition of new samples and data types. Our research projects will continue to expand in the number of species, tissues and specific DNA-protein interactions explored. We will address these areas of research both in the context of our established collaborative projects with the Odom group at the University of Cambridge and as part of other collaborations. Our quest to understand the differentiation process and components of cell- and tissue-specific regulation will be enhanced by new collaborations with the GTEx Consortium, which aims to create a comprehensive public atlas of gene expression and regulation across multiple human tissues, and with the Centre for Therapeutic Target Validation.

Selected publications

Flicek P, Amode MR, Barrell D, Beal K., et al. (2014) Ensembl 2014. *Nucl Acids Res* 42: D749-D755

Ritchie GR, Dunham I, Zeggini E and Flicek P (2014) Functional annotation of noncoding sequence variants. *Nat Methods* 11:294-296

MacArthur JA, Morales J, Tully RE, et al. (2014) Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucl Acids Res* 42:D873-D878

Eckersley-Maslin MA, Thybert D, Bergmann JH, et al. (2014) Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev Cell* 28:351-365

Jiang Y, Xie M, Chen W, et al. (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344:1168-1173

Chen L, Kostadima M, Martens JH, et al. (2014) Transcriptional diversity during lineage commitment of human blood progenitors. *Science* 345:1251033



Non-vertebrate Genomics

High-throughput sequencing is transforming both understanding and application of the biology of many organisms. In our team we integrate, analyse and disseminate 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., 2014) for invertebrate vectors of human disease, WormBase (Harris et al., 2013) for nematode biology, PomBase (McDowall et al., 2012) for fission yeast *Schizosaccharomyces pombe*, and PhytoPath for plant pathogens. In the plant domain, we collaborate closely with Gramene in the US and with a range of European groups in the transPLANT project.

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

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.

Major achievements

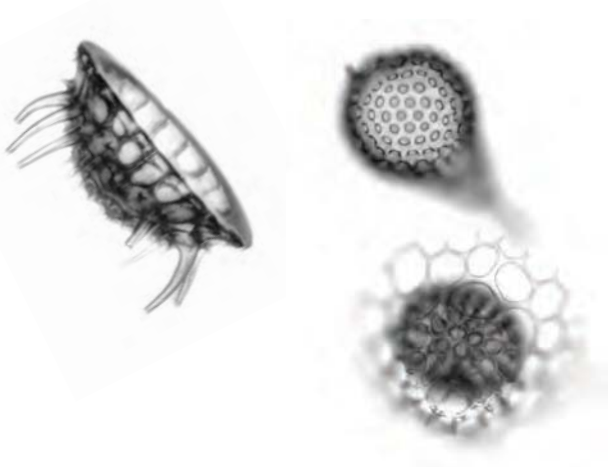
In 2014 we issued three public releases of Ensembl Genomes, and contributed to the regular data releases of VectorBase, WormBase and PomBase. Ensembl Bacteria now includes over 20 254 genomes (an increase from 9675 at the end of 2013) from distinct species. In addition, we added new genomes from four invertebrate metazoans, 12 plants, nine fungi and eight protists to the public release. There are now 21 grass genomes in Ensembl Plants, 20 mosquitoes in Ensembl Metazoa, and 46 fungal or protist plant pathogens included in the release.

Among the significant additions is the chromosome survey sequence from the hexaploid genome of bread wheat. We introduced this large dataset together with new visualisation and comparative analysis features, allowing users to navigate between different homoeologues and view sequence alignments and evolutionary history.

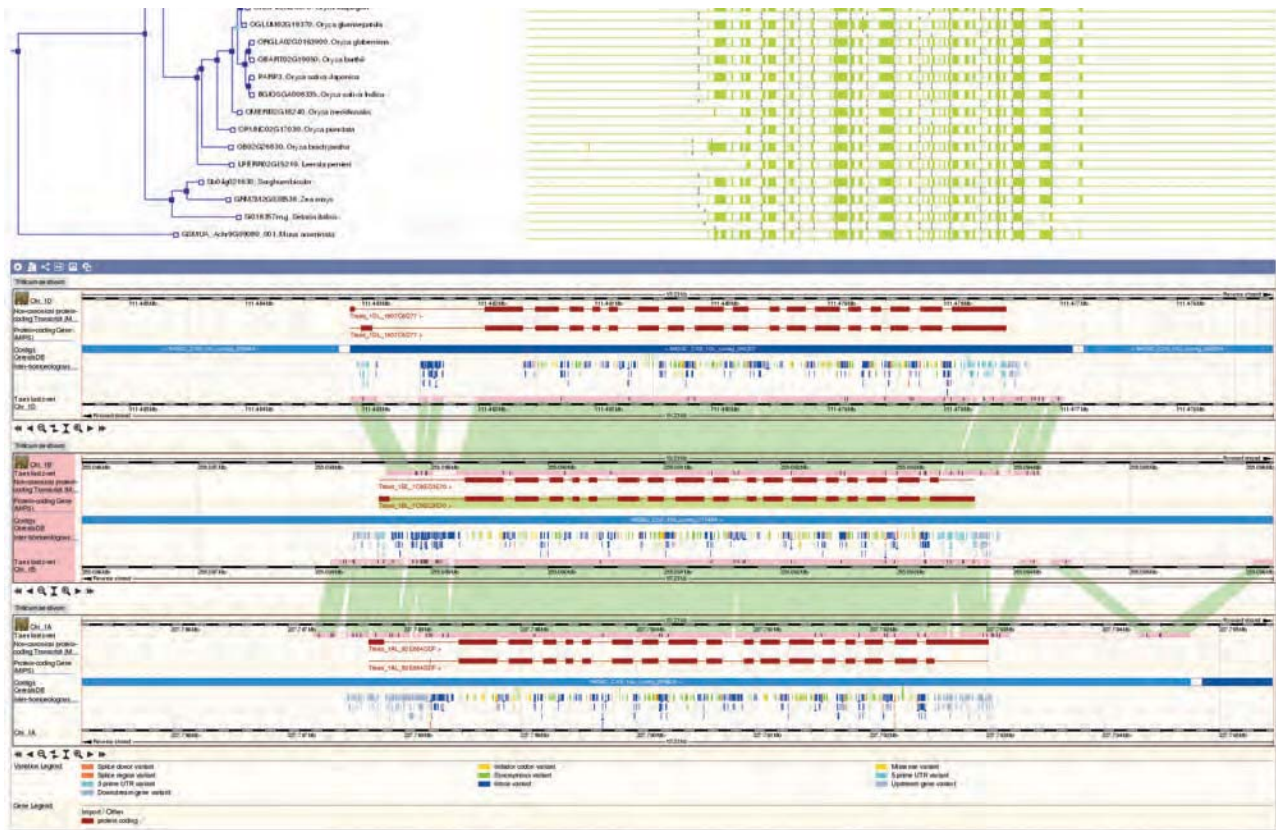
A number of major annotation projects with which the team has been involved were published in 2014, including the first reference genome sequences of the tsetse fly, *Strigamia maritima* (a centipede), the Glanville fritillary butterfly and 16 different species of *Anopheles* mosquito.

In collaboration with the Bateman and Cochrane teams, we developed RNACentral: a new resource providing information about non-coding RNAs. The site provides access to functional annotation, sequence similarities, and genomic locations, and new features linking sequence to structure are currently under development. RNA biology is a rapidly growing area and the new site fills a gap in EMBL-EBI's service portfolio.

Parasitic worms are responsible for more than a billion human infections globally and have a devastating impact on livestock and agriculture. In collaboration with the Wellcome Trust Sanger Institute, we released a new resource focused on the large numbers of parasitic roundworms and flatworms that have recently been sequenced. The first release of WormBase ParaSite offers access to sequence and annotation from 82 species and new features for data mining are under development.



At EMBL-EBI since 1999.



Future plans

We will continue to develop more powerful tools for data mining and data extraction, particularly across multiple species. We expect to release new solutions for exploring plant pathogens, parasitic worms and bacteria in 2015.

Ahola V, Lehtonen R, Somervuo P, et al. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* 5:4737

Neafsey DE, Waterhouse RM, Abai MR, et al. (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347:1258522; published online 27 November 2014

International Glossina Genome Initiative (2014) Genome Sequence of the Tsetse Fly (*Glossina morsitans*): Vector of African Trypanosomiasis.. *Science* 344:380-386

2014 EMBL-EBI Annual Scientific Report



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.

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 European Variation Archive (EVA), the global reference catalogue of genetic variation, provides a basis for interpreting each new genome and variant observed in research and clinical studies.

The EVA, which 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 (ENA and EGA), supporting value-added analysis and visualisation resources. 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.

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 European Genome-phenome Archive (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.

globally collaborative database has been positive. EMBL-EBI served as the hub for these efforts, recruiting a number of European datasets that can now be harmonised and accessed more easily.

The EVA has focused on both data content and technology development. To scale up the data to more than 5 billion variant records, we moved from traditional relational database models to a NoSQL data model, which now provides infrastructure that can scale horizontally to arbitrarily large numbers of datasets across hundreds of species.

The EGA project made advances in user-facing download services in 2014, improving the quality and reliability of service through a re-engineered access-control and download system and automated performance monitoring. The EGA completed a major transition successfully during the move of EMBL-EBI’s data centres, with minimal impact on user services. Infrastructural work to streamline the EGA system for easier maintenance will allow the EGA to operate more efficiently. A move to a REST API-based website laid the groundwork for our team to move the resource to a richer website and programmatic web services in the coming year.

The EGA collaboration with CRG Barcelona, coordinated by ELIXIR, reached the milestone of transferring a half of a petabyte of data to be hosted in both locations. The

Major achievements

In its third year, the Variation team moved the EVA project into full production, expanding the range of data and services to become a primary archive of genetic variation at EMBL-EBI. We expanded the web interface to include interactive search and filter functionality at the study, file and variant level. These data are deeply linked with the supporting VCF and sequencing BAM/CRAM files, providing a chain of provenance and study level granularity – a unique resource for the global research community.

The EVA is a collaborative effort and its primary partner database is dbSNP, developed at the National Center for Biotechnology Information in the US. In 2014 we solidified this collaboration and agreed on data-exchange standard operating procedures. The scientific community’s reaction to this



Justin Paschall

MA 2008, Washington University St. Louis.

Variation Team Leader at EMBL-EBI since 2012.



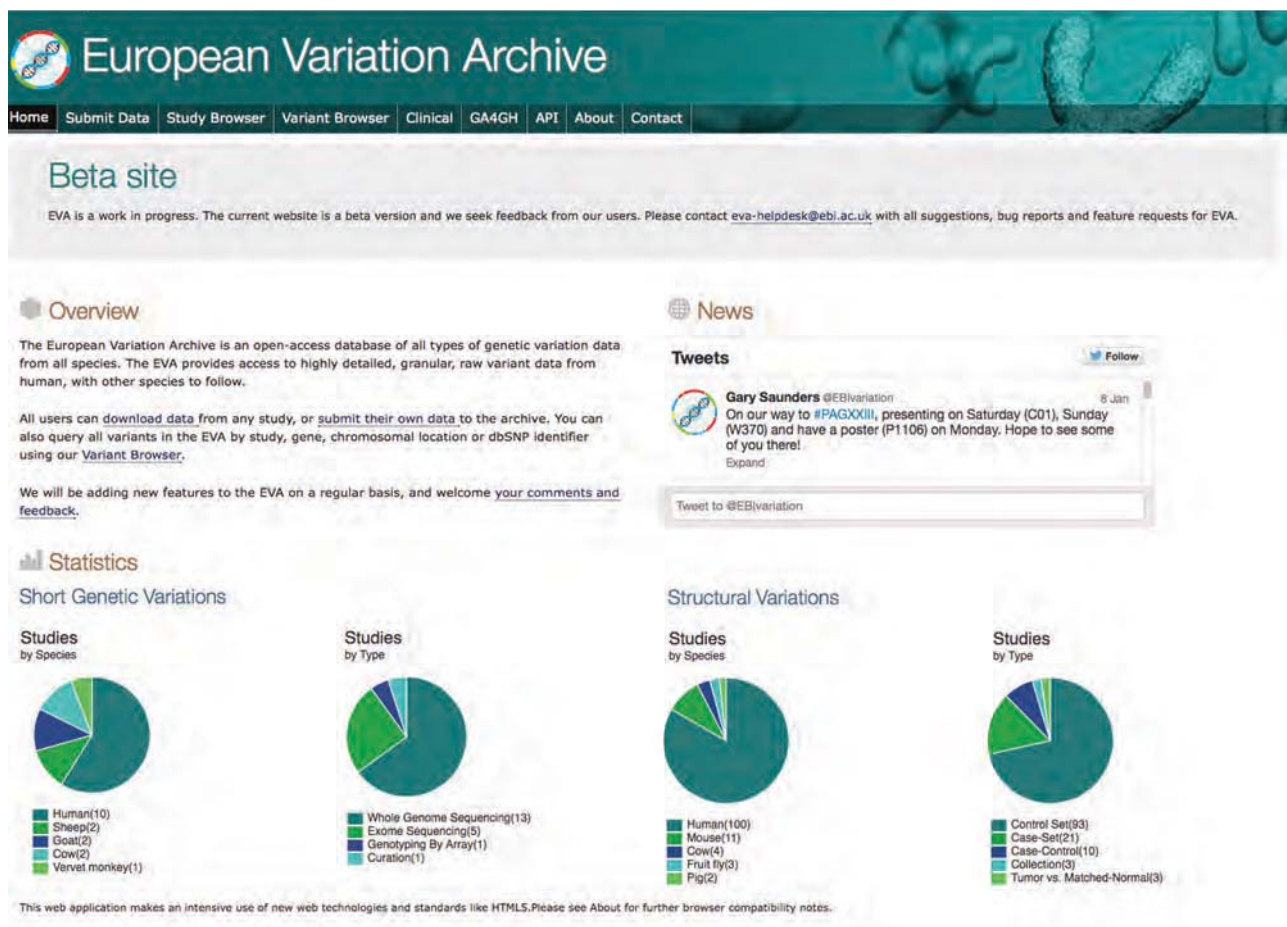
infrastructural work that enabled this automated, large-scale data transfer opened up possibilities to work with additional partners, and new opportunities for a range of EMBL-EBI and ELIXIR projects. EGA user service benefitted from distributed efforts in submission and helpdesk coordination between both groups. It also benefitted from technical collaborations to make portable and install the EGA's secure download system at the CRG, which enables data access at both locations and increases the overall bandwidth and reliability of access to the EGA data.

Future plans

In 2015 we will focus on consolidating progress in the EVA project, streamlining the submission process to remove bottlenecks so that EVA can scale to projected submissions of hundreds of studies per month.

EGA work will focus on bringing the distributed system of data submissions and data access into full production between EMBL-EBI and CRG Barcelona. We will also work to provide improved data-download and data-discovery systems on the EGA website. We will explore interactive and cloud-based data-access options, which may satisfy many user needs without requiring terabyte-sized dataset downloads, and maintain data security.

Both the EVA and EGA will continue working with the Global Alliance for Genomics and Health (GA4GH) to shape and implement future standards for genomic variation data models and access within the context of a global, collaborative informatics ecosystem.



The European Variation Archive (EVA) launched in 2014: www.ebi.ac.uk/eva



Expression

Our resource teams that focus on RNA and protein expression data are working to create a comprehensive, integrated and scalable atlas of expression. Our efforts in this area will make it easier for researchers to achieve a systems-based understanding of the human body and the many species with which we interact.

In December 2014 we welcomed Robert Petryszak as a new group leader for Functional Genomics. With significant expertise in coordinating functional genomics resources, he will take on leadership of the RNA Expression Atlas. Our Expression Atlases now offer data from over 16 000 assays.

We devoted major efforts to redeveloping the ArrayExpress submission system, which we released in mid-2014. We developed a new tool, Annotare, to strike the balance between ease of use for the person submitting data, the need to collect sufficient information for others to reuse the data and support for reproducible science.

The functionality behind our gene-expression data resources is flexible and can be applied to many domains, and in 2014 the Sarkans team released the BioStudies database: a repository for supplementary data files from published life-science experiments.

Expression Atlas

The Expression Atlas is an added-value database that uses gene expression data from EMBL-EBI resources such as ArrayExpress, and makes these data accessible to the entire research community via standardised reanalysis and curation. It shows which genes are expressed under different conditions, and how expression differs among conditions. The Expression Atlas currently holds RNA expression data from microarray and RNAseq experiments, and future development will include protein and metabolite expression data.

www.ebi.ac.uk/gxa

ArrayExpress

ArrayExpress is one of the repositories recommended by major scientific journals to archive functional genomics data to support reproducible research. To serve this mission, we facilitate Minimum Information About a Microarray Experiment (MIAME) and Minimum Information About a High-Throughput Sequencing Experiment (MINSEQE) guidelines. While raw data from high-throughput sequencing-based experiments is handled by the European Nucleotide Archive, experiment descriptions and processed data are archived in ArrayExpress.

www.ebi.ac.uk/arrayexpress

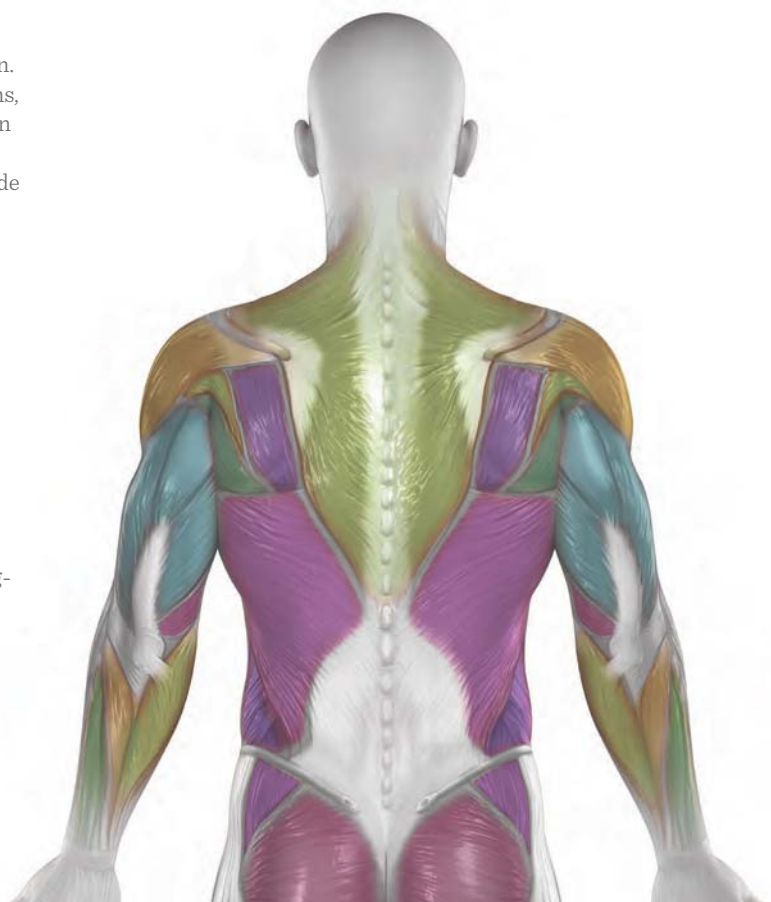
PRIDE

The PRIDE PRoteomics IDentifications database is an open, standards-compliant, public resource for mass-spectrometry-based proteomics data. The largest node in the international ProteomeXchange consortium, PRIDE includes protein and peptide identifications, post-translational modifications and supporting spectral evidence.

www.ebi.ac.uk/pride

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. BioStudies 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.





Functional Genomics

Team leader: Alvis Brazma

- Developed the concept of the BioStudies Database for storing descriptions of biological studies and unstructured data at EMBL-EBI;
- Led the effort to incorporate data from large genomics and proteomics studies in the Baseline Expression Atlas;
- In the context of the CAGEKID consortium, uncovered novel pathways and genes affected by recurrent mutations and abnormal transcriptome patterns.

Proteomics Services

Team leader: Henning Hermjakob

- As part of the ProteomeXchange consortium, processed almost 1000 submissions in 2014, 80% more than in 2013;
- To integrate data across studies, further developed our high-performance spectral-clustering approach, now clustering 60 million spectra within three days on a Hadoop cluster;
- Co-developed the qcML and mzTab community formats for quality assessment and summation of mass spectrometry-based proteomics experiments.

Functional Genomics Development

Team leader: Ugis Sarkans

- Replaced the ArrayExpress submission tools with an entirely new system: Annotare;
- Released the first version of the BioStudies database in beta;
- Completed the diXa toxicogenomics data warehouse.

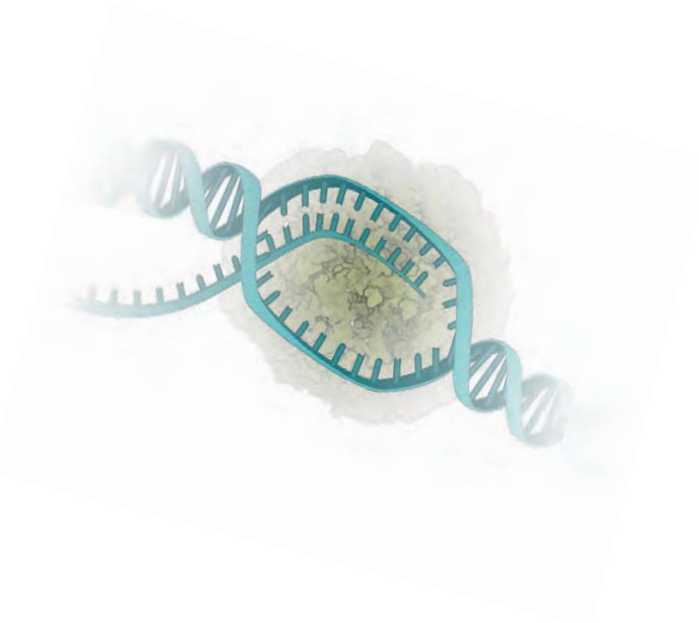


Functional Genomics

The Functional Genomics team provides bioinformatics services and conducts research in functional genomics data analysis, particularly concentrating on high-throughput sequencing-based gene expression and related proteomics data.

We are responsible for a number of 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 complements the Functional Genomics service team, and focuses on developing new methods and algorithms and integrating new types of data across multiple platforms. We are particularly interested in cancer genomics and relationships between transcriptomics and proteomics. We collaborate closely with several groups at EMBL-EBI, including the Marioni, Stegle and Saez-Rodriguez groups.



Major achievements

ArrayExpress and the Expression Atlas

In 2014 we released a new submission tool for functional genomics data, built using the community-supported data annotation software Annotare. Our tool simplified both microarray and sequencing-based data submission to ArrayExpress significantly, and provides a direct means for user feedback. This gave us a clearer picture of our users' experiences, which were quite positive, and provides a simple means to learn about their needs. The resource grew substantially during the year: ArrayExpress became the third largest sequencing data broker to the European Nucleotide Archive in terms of the number of studies.

We increased the content of the RNA-sequencing-based Baseline Expression Atlas significantly. In 2014, the resource grew from four datasets to 24, comprising almost 2000 sequencing assays by the end of the year. Taken together, the Baseline and Differential Expression Atlases now offer data from over 16 000 assays. In terms of interface improvements, we implemented new features that make the resource easier to use.

Research

As part of the EU-funded GEUVADIS project, we led the analysis of transcript isoform use and fusion gene discovery from lymphoblastoid cell lines of 465 individuals who participated in the 1000 Genomes Project. The human transcriptome contains in excess of 100 000 different transcripts. We analysed transcript composition in 16 human tissues and five cell lines to show that, in a given condition, most protein-coding genes have one major transcript expressed at a significantly higher level than others, and that in human tissues the major transcripts contribute almost 85% to the total mRNA.

With our collaborators from Canada, France, Latvia, the UK and other countries we co-led a European renal cancer project, CAGEKID, which is part of the International Cancer Genome Consortium (ICGC). Our findings supported previous reports on frequent aberrations in the epigenetic machinery and PI3K/mTOR signaling, and uncovered novel pathways and genes affected by recurrent mutations and abnormal transcriptome patterns including focal adhesion, components of extracellular matrix (ECM) and genes encoding FAT cadherins.

Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. MSc in mathematics, University of Latvia, Riga.

At EMBL-EBI since 1997.



Future plans

Integration of baseline RNA sequencing gene expression and proteomics data will be at the core of our efforts in developing the biology-centric Expression Atlas. Our 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.

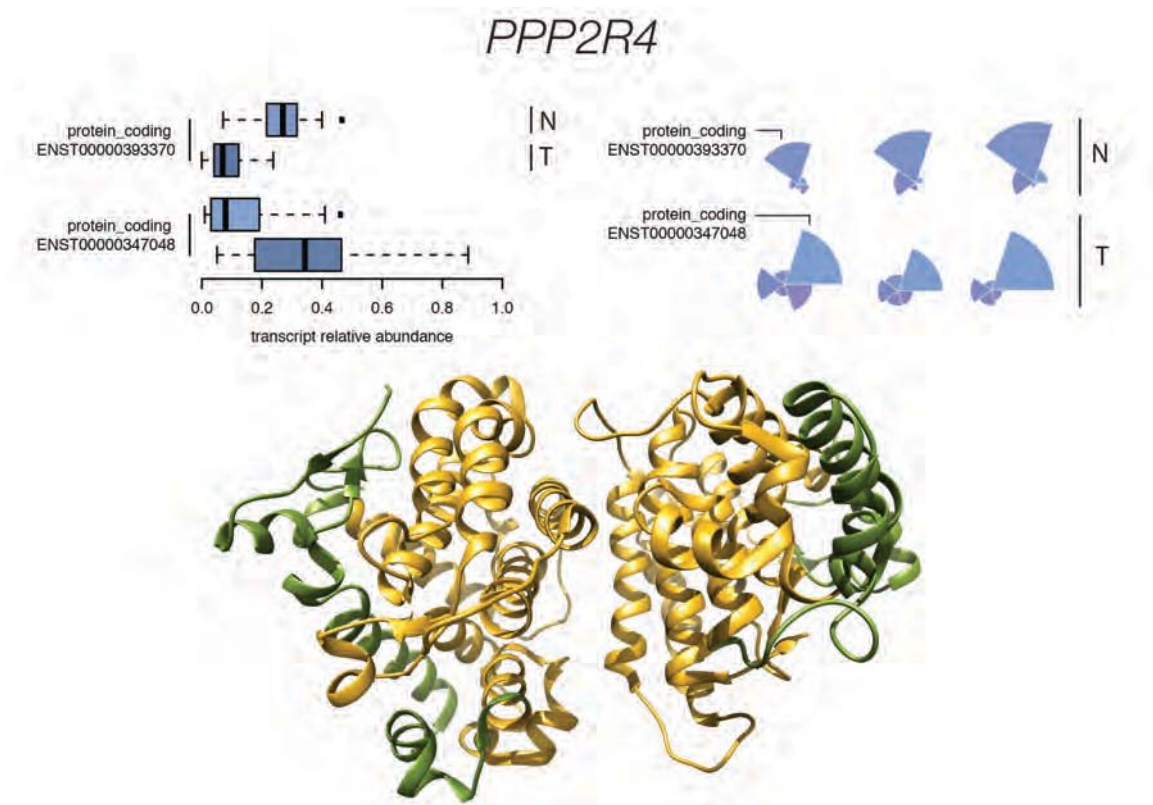
Selected publications

Fonseca NA, Marioni J, Brazma A. (2014) RNA-Seq gene profiling--a systematic empirical comparison. *PLoS One* 9:e107026

Marguerat S, Lawler K, Brazma A, Bähler J. (2014) Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress. *RNA Biol.* 11:702-14

Petrzszak R, Burdett T, Fiorelli B, et al. (2014) Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucl Acids Res* 42:d926-d932

Scelo G, Riazalhosseini Y, Greger L, et al. (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* 5:5135



Loss of function through alternative splicing in renal cancer, using the example of the protein phosphatase 2A activator (PPP2R4). Ensembl data shows a switch between two PC transcripts. The APPRIS database, which houses protein structure, function and conservation information for splice variants (developed at CNIO and INB in Spain), shows the principal transcript in N, but not in T. Using the EMBOSS, Needle and UniPDB tools, we see there is less than 35% protein overlap.



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 EMBL-EBI services.

We also contribute to the development of the BioSamples database, 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

We devoted major efforts to redeveloping our submission system, which we released in mid-2014. With the new tool, Annotare, we struck a balance between providing wizard-style functionality for ease of use, and a spreadsheet-like interface to enable larger submissions.

ArrayExpress

Our team recorded and analysed the usage of ArrayExpress data in detail. We developed an interface that allows us to see access statistics for each individual study easily, and will release certain views on this information publicly in 2015. One insight we gained was that data from 39 studies have been downloaded more than 1000 times, while 90% of the studies in ArrayExpress have been downloaded at least 10 times.

We worked on the externally accessible release-date management system to enable data submitters to enter publication information. One outcome of this development is that the ArrayExpress content team can concentrate on value-added data curation. We ceased to actively develop ArrayExpress in June 2014, but continue to maintain the resource.

BioStudies

The BioStudies database holds descriptions of biological studies and links to data from these studies in other databases at EMBL-EBI or outside, and is a repository for supplementary data files from published life-science experiments that do not fit in the structured archives at EMBL-EBI. In 2014 we intensified our development of the this database significantly, undertaking work on fundamental building blocks: a flexible

format for describing a study, a validation/storage/indexing engine and a data-access interface for searching and browsing studies. In summer of 2014, we started work on a simple data-submission tool.

BioSamples

We continued to collaborate with the Samples, Phenotypes and Ontologies team on the basic infrastructure components of the BioSamples database, with a particular emphasis on scalability and robustness of the user interface, database backend and data transformation tools. We also collaborated with the Expression Atlas team on building the Expression Atlas user interface.

Toxicogenomics and biomedical informatics

Our work on the diXa toxicogenomics data warehouse resulted in a generic data-management solution for biological studies that we will repurpose for the BioStudies database.

We participate in a number of medical informatics projects. We provide data management solutions for the EU-AIMS project on autism spectrum disorder; integrate our R cloud scientific computation infrastructure into medical data exploration systems, for example for the European Medical Information Framework (EMIF) project for the reuse of patient health records in clinical research; and work towards a generic data security infrastructure for the BioMedBridges project, which facilitates data sharing in the life sciences.

Ugis Sarkans

PhD in Computer Science, University of Latvia, 1998. Postdoctoral research at the University of Wales, Aberystwyth, 2000.

At EMBL-EBI since 2000.



Future plans

We plan to further develop the Annotare tool for data submissions so that it is useful for ArrayExpress, BioSamples and BioStudies. We will automate the ArrayExpress data flow so that our curators can concentrate on biological curation of data for the Expression Atlas.

To make BioStudies an appealing destination for supplementary data files in the life sciences, we will work on content presentation. We will further develop the data submission tool so that integrating 'publication' and 'data' information becomes easier. For example, we will provide a mechanism for logging into the submission tool using ORCID identifiers. We will also implement automated checks that direct data to the appropriate public data archive (e.g. sequence, proteomics or molecular structure data), where possible.

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. The BioStudies database will tie various modalities together, and provide study metadata and robust links to the BioSamples database. This will all support 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, the former containing overall study descriptions and grouping assays, and the latter grouping samples and providing annotation.

Our continued participation in medical informatics projects will provide us with a better understanding of the data types and data management patterns across a range of life-science communities, which is essential for the development of the BioStudies database.

References

Ruggeri B, Sarkans U, Schumann G and Persico AM (2014) Biomarkers in autism spectrum disorder: the old and the new. *Psychopharmacology* (Berl) 231:1201-1216

Faulconbridge A, Burdett T, Brandizi M, et al. (2014) Updates to BioSamples database at European Bioinformatics Institute. *Nucl Acids Res* 42:d50-d52

Kurbatova N, Brandizi M, Burdett T, et al. (2014) ArrayExpress update-simplifying data submissions. *Nucl Acids Res* 43:D1113-D1116

The diXa toxicogenomics data warehouse: a precursor to BioStudies

diXa Home / Studies			
26 Studies found			
ID	Project	Title	Description
DIXA-011	PredTox	Study FP001RO: Evaluation of the Acute Toxicity, Gene Expression, Protein Expression, Metabolite Production, Clinical Chemistry and Pathology Profile Following an Oral Administration of Compound R2717 to Rats	Repeated dose 14-day toxicity study in adult male rats (<i>Rattus norvegicus</i>), Wistar strain, using chemical compound FP001RO, administered daily orally, sponsored by Hoffmann-La Roche AG, funded by European Union Framework Program 6 - Innovative Medicines Initiatives - Integrated Project Predictive Toxicology (EU FP6 InnoMed PredTox: LSHB-CT-2005-518170). The study will examine the gene,
DIXA-012	PredTox	Study FP002BI: Evaluation of the Acute Toxicity, Gene Expression, Protein Expression, Metabolite Production, Clinical Chemistry and Pathology Profile Following an Oral Administration of Compound BI-2 to Rats	Repeated dose 14-day toxicity study in adult male rats (<i>Rattus norvegicus</i>), Wistar strain, using chemical compound FP002BI, administered daily orally, sponsored by Boehringer Ingelheim Pharma GmbH, funded by European Union Framework Program 6 - Innovative Medicines Initiatives - Integrated Project Predictive Toxicology (EU FP6 InnoMed PredTox: LSHB-CT-2005-518170). The study will examine the
DIXA-013	PredTox	Study FP003SE: Evaluation of the Acute Toxicity, Gene Expression, Protein Expression, Metabolite Production, Clinical Chemistry and Pathology Profile Following an Oral Administration of Compound AS605134 to Rats	Repeated dose 14-day toxicity study in adult male rats (<i>Rattus norvegicus</i>), Wistar strain, using chemical compound FP003SE, administered daily orally, sponsored by Serono, funded by European Union Framework Program 6 - Innovative Medicines Initiatives - Integrated Project Predictive Toxicology (EU FP6 InnoMed PredTox: LSHB-CT-2005-518170). The study will examine the gene, protein and metabolite profiles,
DIXA-014	PredTox	Study FP004BA: Evaluation of the Acute Toxicity, Gene Expression, Protein Expression, Metabolite Production, Clinical Chemistry and Pathology Profile Following an Oral Administration of Compound Bay 16-4749 to Rats	Repeated dose 14-day toxicity study in adult male rats (<i>Rattus norvegicus</i>), Wistar strain, using chemical compound FP004BA, administered daily orally, sponsored by Bayer Schering Pharma AG, funded by European Union Framework Program 6 - Innovative Medicines Initiatives - Integrated Project Predictive Toxicology (EU FP6 InnoMed PredTox: LSHB-CT-2005-518170). The study will examine the gene,



Proteins and protein families

EMBL-EBI provides world-leading resources for researchers working with protein sequences and protein families, including UniProt, InterPro and Pfam, among others.

We provide the well-known UniProt resource in collaboration with the SIB Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR). We also provide two of the most important protein families resources: InterPro and Pfam. For each of our resources, we combine expert biocuration of the literature with computational tools to provide deep and comprehensive annotations of protein sequences.

In 2014 we welcomed Rob Finn as Team Leader for Protein Families, steering the development of InterPro and Pfam. Another important development was the release of an all-new UniProt website, which was based on extensive usability testing. UniProt's reference proteome set expanded by a factor of two, and will become the basis of curation for both Pfam and InterPro.

The number of sequences within UniProt has grown substantially, though this is due in part to closely related genomes. During the coming year we will begin to remove near-identical proteomes from UniProt, which will reduce the size of the resource and make it easier for scientists to identify key information about proteins

UniProt

UniProt is a collaboration among EMBL-EBI, SIB and the PIR group in the United States. Its purpose is to provide the scientific community with a single, centralised, authoritative resource for protein sequences and functional annotation. The consortium supports biological research by maintaining a freely accessible, high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross references and querying interfaces.

The work of our teams spans several major resources under the umbrella of UniProt, each of which is optimised for a different purpose:

- *The UniProt Knowledgebase (UniProtKB) is the central database of protein sequences and provides accurate, consistent and rich annotation about sequence and function;*
- *The UniProt Archive (UniParc) is a stable, comprehensive, non-redundant collection representing the complete body of publicly available protein sequence data;*

- *UniProt Reference Clusters (UniRef) are 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: Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH-Gene3D, PANTHER and HAMAP. During the integration process, InterPro rationalises instances where more than one protein signature describes the same protein family or domain, uniting these into single InterPro entries and noting relationships between them where applicable.

InterPro adds biological annotation and links to external databases such as GO, PDB, SCOP and CATH. It precomputes all matches of its signatures to UniProt Archive (UniParc) proteins using the InterProScan software, and displays the matches to the UniProt KnowledgeBase (UniProtKB) in various formats, including XML files and web-based graphical interfaces.

InterPro has a number of important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects. InterPro is used by Ensembl and in the GOA project to provide large-scale mapping of proteins to GO terms.

www.ebi.ac.uk/interpro

Pfam

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



MEROPS

The MEROPS database comprises proteolytic enzymes (also termed proteases, proteinases and peptidases), their substrates and inhibitors. MEROPS uses a hierarchical, structure-based classification of proteolytic enzymes and protein inhibitors. Each peptidase or inhibitor is assigned to a family on the basis of statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a clan.

<http://merops.sanger.ac.uk>

Enzyme Portal

The Enzyme Portal provides integrated enzyme-related data for all EMBL-EBI enzyme resources as well as the underlying functional and genomic data.

www.ebi.ac.uk/enzymeportal

Protein Families

Team leader: Rob Finn

- Developed the InterPro Domain Architecture (IDA) tool, which allows the database to be searched with a set of domains;
- Refactored the InterPro production pipeline to cope with the growth of UniProtKB;
- Increased InterPro coverage of UniProtKB proteins by performing significant data curation and integration;
- Migrated the Pfam infrastructure from the Wellcome Trust Sanger Institute to EMBL-EBI.

www.ebi.ac.uk/interpro

<http://pfam.xfam.org>

UniProt Development

Team leader: Maria Martin

- Released a new UniProt website and developed new interfaces and tools for the Enzyme Portal and QuickGO, with a focus on optimising user interaction with these websites;
- Developed a method for identification of highly redundant proteomes in preparation for their removal from UniProtKB in 2015;
- In collaboration with Ensembl and COSMIC, incorporated variation data from human, mouse and zebrafish into UniProt KB;
- In collaboration with PRIDE and ProteomeXchange, incorporated proteomics data in UniProtKB;
- Extended Protein2GO, the common annotation tool for the GO Consortium, to include GO annotation of protein complexes and RNAs;
- Developed data services for the delivery of UniProt disease-related information and web interfaces to the Centre for Therapeutic Target Validation.

UniProt Content

Team leader: Claire O'Donovan

- As part of the Consensus CDS (CCDS) project, worked on an authoritative complete proteome set for Homo sapiens, in part by ensuring a curated, complete synchronisation with the HUGO Gene Nomenclature Committee (HGNC);
- Thanks to funding from the Centre for Therapeutic Target Validation (CTTV), prioritised the annotation of proteins and their variants involved in human disease;
- Played 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;
- Expanded our automatic annotation effort and prioritised the annotation of enzymes and pathways to enable proteome annotation, in cooperation with our collaborators in the genome annotation community;
- Provided high-quality annotations for human proteins as part of the GO Consortium Reference Genomes Initiative, and reached out to the computational community through our involvement in the Critical Assessment of Function Annotation experiment (CAFA);
- Contributed to the UniProt Proteomes effort with the selection of proteomes and provision of their descriptions.



Protein Families

The Protein Families team is responsible for the InterPro, Pfam and Rfam data resources, and coordinates the EBI Metagenomics project. InterPro integrates protein family data from 11 major sources, hierarchically classifying the different protein family, domain and functional site definitions to provide a unified view of the diverse data.

A tool called InterProScan (available as either a web service or a downloadable software package), allows the identification of InterPro entries on protein sequences. Pfam, a member database of InterPro, generates new protein family entries and has the largest sequence coverage of any of the InterPro member databases. Both InterPro and Pfam have a number of important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects.

Rfam classifies non-coding RNA sequences into families, using probabilistic models that take into account both sequence and secondary structure information, termed covariance models (CMs). Rfam is uniquely placed to annotate non-coding RNAs in genome projects and is a major contributing database to RNACentral, a sequence resource launched in 2014.

Metagenomics is the study of the sum of genetic material found in an environmental sample or host species, typically using next-generation sequencing (NGS) technology. EBI Metagenomics enables researchers to submit sequence data and associated descriptive metadata to public nucleotide archives. Once deposited, our team helps ensure the data is functionally analysed using an InterPro-based pipeline, taxonomically analysed using the QIIME software package, and that the results can be visualised and downloaded via a web interface.

Major achievements

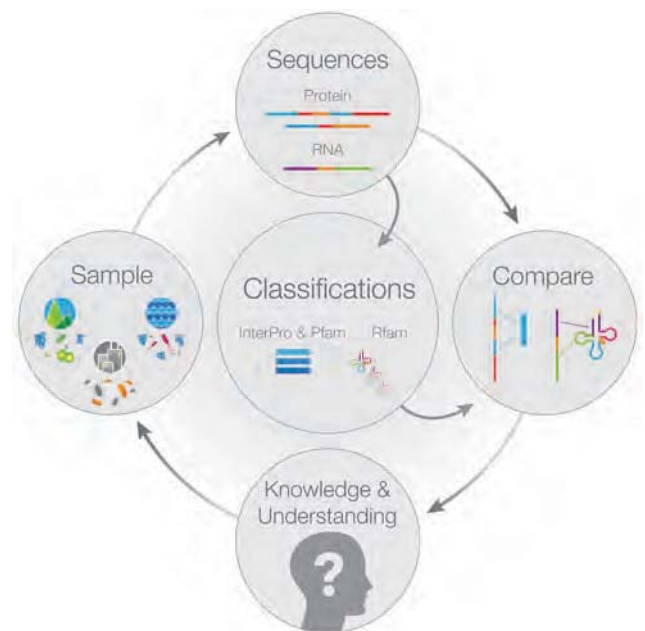
InterPro

The InterPro Domain Architecture (IDA) tool was released in 2014. The tool allows users to search the InterPro database with a particular set of domains, and returns all of the domain architectures and associated proteins that match the query. This makes it easy to rapidly identify all of the different domain combinations in which one type of domain co-occurs with another, or if a particular domain is followed by another (e.g. an SH3 domain is found C-terminal to a protein kinase domain, or vice versa), and to list the proteins that match each domain architecture.

We reconfigured the InterPro production pipeline to cope with the growth of UniProtKB, which doubled in size during 2014. Unlike the old system, the new pipeline is built entirely on InterProScan, which streamlines the production process, removes a number of bottlenecks and helps ensure consistency

of results across all the different InterPro output sources (database, website-based sequence searches or sequence searches using the stand-alone version of InterProScan).

The InterPro database now offers improved coverage of UniProtKB proteins, increasing to 83.4% in the latest release (v. 49.0), up from 81.9% in 2013. This is partly thanks to significant data curation and integration efforts, which led to an additional 1779 signatures being incorporated into the database in 2014. Our focussed curation of InterPro2GO term associations led to 1363 entries being assigned new or updated Gene Ontology (GO) terms; 45% of InterPro entries now have at least one term associated. The total number of GO mappings increased by 2069 overall.



Billions of new protein sequences are brought to light through the study of DNA sequences, both on the level of individual genomes and metagenomes. These are compared to collections of protein families (e.g. in Pfam and InterPro) and of RNA families (e.g. in Rfam) to identify the presence of similar families on them. Over time, these new sequences enter reference databases (e.g. ENA, UniProtKB) and are incorporated into family classifications. These annotations help scientists understand the organisms in their samples, which ultimately leads to the generation of hypotheses and further exploration.

Rob Finn

PhD in Biochemistry, Imperial College London;
Wellcome Trust Sanger Institute, 2001-2010; Janelia
Research Campus, 2010-2013.

At EMBL-EBI since 2014.



Pfam

The Pfam and Rfam teams moved to EMBL-EBI from the Wellcome Trust Sanger Institute in late 2012. In 2014, the infrastructure for these databases and their respective websites were migrated to EMBL-EBI. We also decommissioned the Pfam mirror sites, as EMBL-EBI provided increased infrastructural redundancy.

We migrated over 90% of the 16 152 Pfam seed alignments (i.e. for each family, the alignment of the few representative sequences of that family, which is used to train the profile-hidden Markov model) to use only the UniProtKB reference proteome sequence set. As the reference proteomes are significantly more stable and typically contain the best functional annotations, this will reduce the amount of manual curation required going forward. We modified the Pfam quality-control methods to calculate quality metrics on just the reference proteome sequences. These changes, which will be available from Pfam release 28.0, will enable the team to release Pfam far more frequently than in the past.

Rfam

We completely rewrote the Rfam production pipelines to ensure future scalability. At the core of the update was the adoption of Infernal 1.1 software, which we used to generate CMs and to search sequence databases to identify members of an Rfam entry. Infernal is over 100 times faster than previously employed software, as it implements heuristic filters. We have stopped producing alignments of all family members, as many families now run into millions of sequences and gigabytes of storage. Such alignments, although possible to produce, are very difficult to present to users via the web and are incomprehensible by humans without computational aid. Infernal has also allowed the addition of longer models, most notably the full-length ribosomal RNAs. Its only drawback is that it required us to manually redefine every entry's curated threshold. These new developments were reflected in Rfam 12.0, which contained 2450 entries (up 242 since the last release) and annotation of 19 million sequence regions in the underlying sequence database (a three-fold increase over 2013).

EBI Metagenomics

By December 2014 the EBI Metagenomics hosted 90 public metagenomics projects, comprising 2302 separate samples and a significant number of privately held studies. The total number of raw nucleotide reads processed by the resource passed 45 billion.

We developed an EBI Metagenomics comparison tool that allows users to compare the predicted GO terms for different samples within a metagenomic study, and to visualise the results as bar charts, stacked columns and heatmaps, and to perform principal component analysis.

We held several national and international training workshops in 2014, highlighting the utility of EBI Metagenomics to the scientific community. These included workshops at UNSW in Sydney and Monash University in Melbourne in February, in collaboration with BioPlatforms Australia and CSIRO, and at UCB in Brasília in November, funded under a BBSRC Brazil Partnership award, in collaboration with CNPq.

Future plans

In 2015 we will begin to combine the InterPro and Pfam websites, with the ultimate aim of providing a single entry point for protein-family data at EMBL-EBI. The single site will reduce overheads and provide a consistent representation of family data, combining the strengths of the complementary resources. In addition to unifying the website, we will identify and remove overlapping redundancies between the two production pipelines.

We will transfer the web-based HMMER homology search, developed at Janelia Research Campus, to EMBL-EBI. HMMER provides fast, accurate searching of protein sequence databases such as UniProtKB. Because it uses the same technology as Pfam, we plan to integrate the HMMER user interface with its internal curation system.

We will release Pfam 28.0 early in 2015, followed by another release as improvements come into force. We anticipate adding over 500 new entries to Pfam, based on existing approaches for identifying new families and based on methods that utilise the wealth of InterPro data. We will also move the Rfam database to a more genome-centric view of non-coding RNA. We will also improve the data flow between Rfam and RNACentral, enhancing the content of both resources.

Our team will extend the EBI Metagenomics comparison tool to taxonomic analyses, and provide more powerful and precise metadata searching. We also plan to extend the pipeline by adding new analysis algorithms and visualisations so that it offers enhanced taxonomic and functional analysis of samples.

Selected publications

Finn RD, Bateman A, Clements J, et al. (2014) Pfam: the protein families database. *Nucl Acids Res* 42 (database issue), d222-d230.

Finn RD, Miller BL, Clements J, and Bateman A (2014) iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucl Acids Res* 42 (database issue), d364-d373.

Mitchell A, Chang HY, Daugherty L, et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucl Acids Res* 43:d213-d221.



UniProt Development

The work of our team spans several major resources under the umbrella of UniProt, a comprehensive resource of protein sequences and functional annotation: the UniProt Knowledgebase, the UniProt Archive and the UniProt Reference Clusters.

Our team develops the software and services for protein information in the UniProt, Gene Ontology (GO) annotation and enzyme data resources. 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 2014 our team released new web interfaces and functionalities built in response to user feedback gathered in a number of user workshops, usability interviews/sessions, helpdesk reviews and surveys. We developed the new UniProt website in collaboration with our colleagues at SIB, and launched it in September 2014. We also launched a new blog (insideuniprot.blogspot.co.uk), which highlights important developments, helpful features and interesting datasets in UniProt.

We worked with our user community, NCBI RefSeq, Ensembl and Ensembl Genomes to provide a collection of non-redundant reference proteomes and to maintain well-annotated organisms for biomedical and biotechnological research. We developed new pages in the UniProt website for each of these datasets, including information about their species and corresponding genome assembly. New species released in 2014 include *Ovis aries* (Sheep), *Zea mays* (Maize) and *Triticum aestivum* (Wheat), among others.

The team developed a new method to identify identical and/or very similar species and strains that can make it difficult to explore protein data. We made plans to incorporate this method in our release pipelines by removing redundant proteomes from UniProtKB, making them available only through the sequence archive UniParc.

In collaboration with genomic resources such as 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 began distributing variants with consequences at the protein level from the human 1000 Genomes Project, mouse and zebrafish species.

In 2014 we extended the UniRule and the Statistical Automatic Annotation System (SAAS), two systems for automatic annotation of large volumes of uncharacterised proteins.

UniRule provides new functionality for prioritising rules for annotation, and allows annotation of GO terms in rules. SAAS now annotates protein names and GO terms. These systems annotate over 33 million sequences in UniProtKB/TrEMBL. The team worked to develop new automatic annotation methods based on domain architectures and association rule selection by dominance relationships, which will increase the accuracy and coverage of annotations. We extended our collaborations with the automatic annotation communities, hosting students who analysed published methodologies such as GoPred, a collaboration with Prof. Rengul Cetil-Atalay from the Middle East Technical University in Turkey.

In 2014 we continued to re-engineer QuickGO, the UniProt GO browser, using Apache Solr for indexing and filtering over 400 million GO annotations. We began to extend the scope of GO annotation to support annotations to entities other than proteins. The first new entity type (protein complexes) is identified using IntAct Complex Portal identifiers. We further developed Protein2GO, the web-based GO curation tool used by UniProt and GO Consortium curators to contribute annotations to the GOA project, to include new features requested by users, and to reflect the evolving annotation rules agreed by the GO Consortium.

Our team maintains the Enzyme portal, a resource that integrates enzyme-related data for all EMBL-EBI enzyme resources as well as the underlying functional and genomic data. We integrated the database within the UniProt infrastructure and developed new web interfaces in response to user feedback. In collaboration with the Web Production team, we refined the enzyme search within the EBI-Search.

Future plans

In 2015 we plan to provide a protein sequence feature viewer summarizing functional sites, and new UniRule web pages in the UniProt web site. We will 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 continue to focus on usability issues and to engage with our users to ensure we maintain a global genome/proteome- and gene-product-centric view of the sequence space. We also aim to collaborate with the ProteomeXchange resources in the integration of post-translational modifications in UniProtKB, and the provision of experimental unique peptide mappings for reference species. We will continue to co-operate with variation projects (e.g.,

Maria-Jesus Martin

BSc In Veterinary Medicine, University Autonoma in Madrid. PhD in Molecular Biology (Bioinformatics), 2003.

At EMBL-EBI since 1996.

Team Leader since 2009.



ESP) to integrate relevant genome and proteome information, and will release COSMIC variation information relevant for proteins.

In 2015 we also plan to extend the scope of GO annotation to encompass entities other than proteins, in particular RNA molecules. We will explore drug-target predictions in collaboration with ChEMBL and PDBe. We will develop data services and interfaces for target validation within the CTTV project. We will work in the development of new web interfaces and functionalities for the Enzyme Portal.

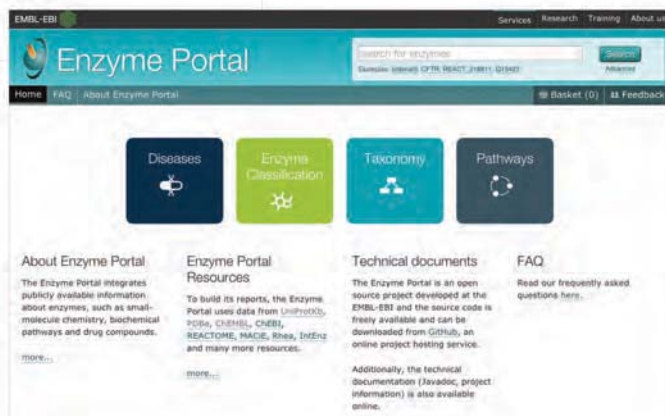
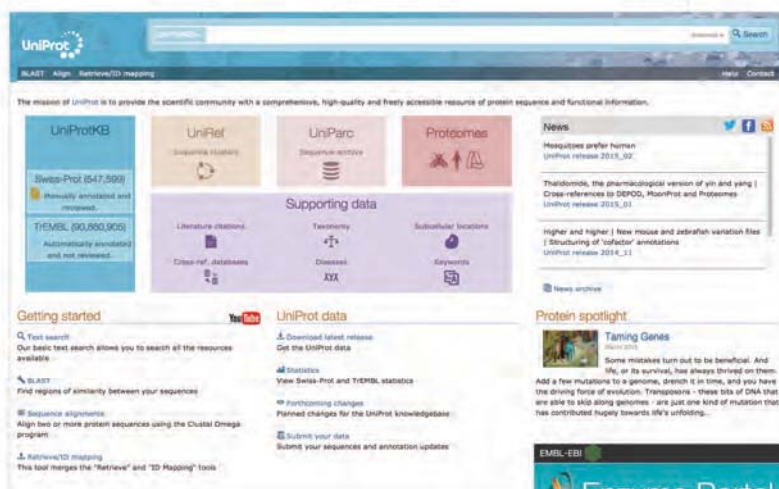
Famiglietti ML, Estreicher A, Gos A, et al. (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum Mutat* 35:927-935

Jupp S, Malone J, Bolleman J, et al. (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 30:1338-1339

Garcia L, Yachdav G, Martin MJ (2014) FeatureViewer, a BioJS component for visualization of position-based annotations in protein sequences. *F1000Res* 3:47

Selected publications

Huntley RP, Sawford T, Martin MJ, O'Donovan C (2014) Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* 3:4



UniProt's new web interface, built in response to user feedback.



UniProt Content

One of the central activities of the UniProt 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 2014. 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

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 2015 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 2015 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



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, has a mission 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).

Protein Data Bank

The PDB is the single international archive for 3D biomacromolecular structure data. PDBe is the European partner in the Worldwide Protein Data Bank organisation (wwPDB), which maintains the PDB archive. The other wwPDB partners are the Research Collaboratory for Structural Bioinformatics (RSCB) and Biological Magnetic Resonance Bank (BMRB) in the US and the Protein Data Bank Japan.

<http://pdbe.org>

Electron Microscopy Data Bank

EMDB is a public repository for 3D electron microscopy (EM) density maps of biomacromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography and electron (2D) crystallography. Founded at EMBL-EBI in 2002, EMDb is now maintained by the EMDataBank partners: PDBe, RCSB PDB and the National Center for Macromolecular Imaging in Houston, Texas.

<http://pdbe.org/emdb>

Electron Microscopy Pilot Image Archive

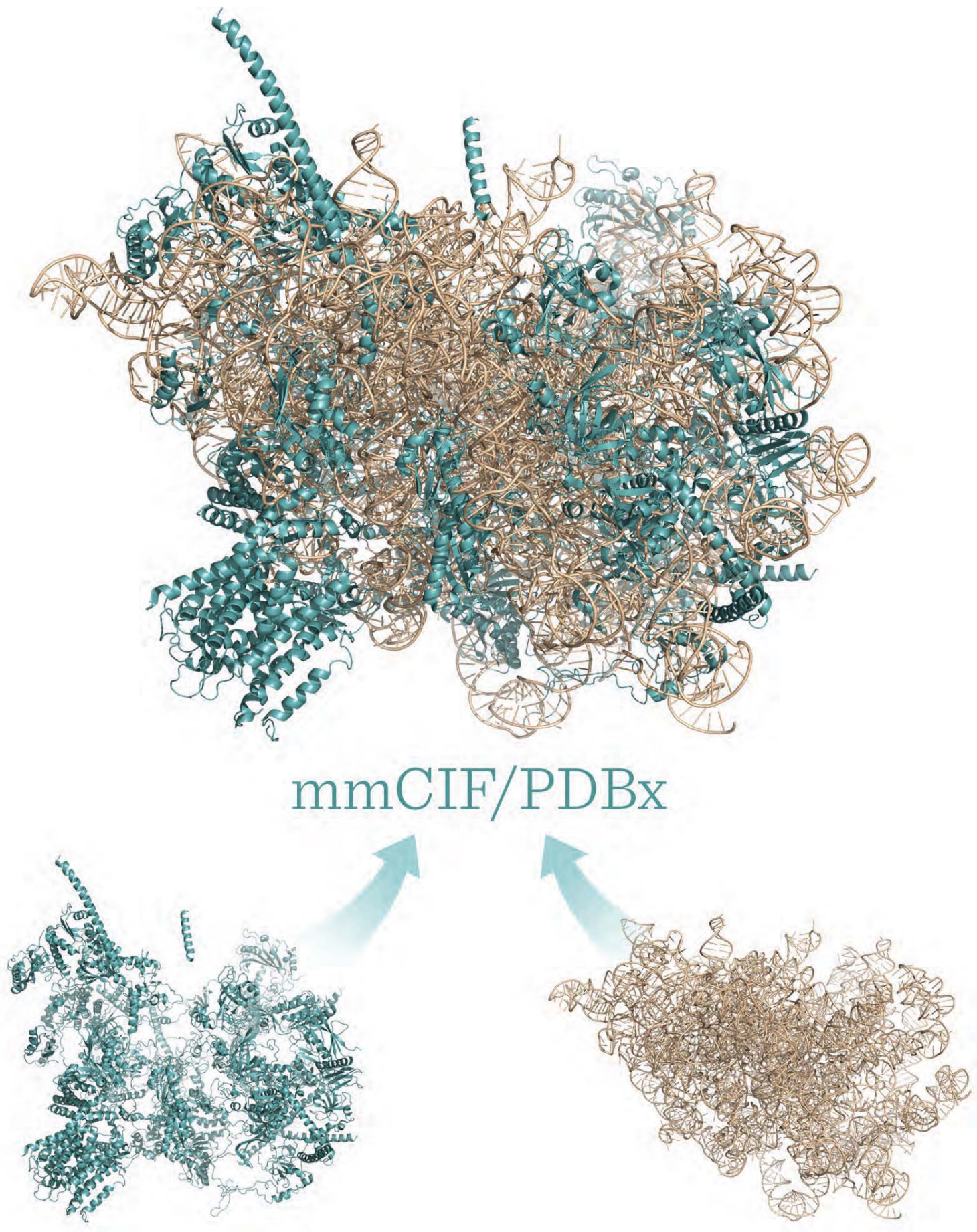
EMPIAR is a public archive for raw, 2D EM images, and complements EMDb. It provides state-of-the-art data to facilitate methods and software development, training and validation, which will lead to better 3D EM structures. EMPIAR is a PDBe project based on input from the EM community.

<http://pdbe.org/empiar>

Protein Data Bank in Europe teams

Gerard Kleywegt & Sameer Valenkar

- *Curated a record 2712 PDB entries;*
- *Curated 298 EMDb entries;*
- *In collaboration with partners in the US and Japan:*
 - *released a new common tool for curation of 3D structural data on biomacromolecules;*
 - *released structure-quality reports for all X-ray crystal structures in the PDB;*
 - *celebrated the 100,000th PDB entry;*
 - *adopted mmCIF as the official archive-distribution format;*
 - *released 300 large structures as single files that were previously split across multiple PDB entries;*
 - *consulted with community experts about the archiving needs in the era of integrative structural biology and hybrid methods for structure determination.*
- *Extended the wwPDB deposition and annotation software, as well as the wwPDB structure-validation pipelines, to handle 3DEM and NMR data;*
- *Established EMPIAR, a pilot archive for raw 3DEM and tomography image data;*
- *Released several tools for validating 3DEM data;*
- *Further developed the new PDBe website (including the PDB and EMDb entry pages), a powerful new search system, and an API, all scheduled for release in 2015;*
- *Improved and extended the internal production process and data and database infrastructure to support the new website, search system and API;*
- *Hosted several workshops and external meetings;*
- *Intensified PDBe outreach efforts in social media.*



Since December 2014, the mmCIF/PDBx format has been the primary archive-distribution format for the PDB. This makes it possible to represent even very large biomacromolecular structures as single entries and files. In the past, users would sometimes need to download several individual files and put them together to gain a complete picture of large and complex structures. Shown here is the structure of the idle mammalian ribosome-Sec61 complex (Voorhees et al., *Cell* 2014), which can be accessed as entry 3j7q. This single entry contains all the information formerly held as two entries (4w24 and 4w25).



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 and with the needs of biologists.

PDBe handles the deposition and annotation of 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

Milestones

In 2014 we celebrated several milestones in the history of the PDB and wwPDB, such as the release of the 100 000th entry, which generated widespread attention including an editorial in *Nature*.

The X-ray crystallography module of the new joint wwPDB Deposition and Annotation software system (D&A) went into production at the start of 2014. PDBe annotators used the system to process all structures from Europe deposited through it. PDBe's NMR and EM experts worked with their international colleagues to extend D&A so that it is capable of handling NMR and 3D EM structures as well. The extended D&A system went into internal testing at the end of the year.

In March, wwPDB validation reports were released for all X-ray crystal structures in the PDB archive, marking the first time that a comprehensive set of quality information was distributed with the PDB. Our team developed the validation pipeline that produces the reports, and began testing a version that also handles NMR and 3DEM structures.

We hosted the first meeting of the wwPDB Hybrid Methods Task Force, which provides advice on how structural data and models that do not fit the traditional PDB mould should be archived and validated. A news article about the meeting, published in *Nature*, asserted that structural biology “is in the grip of a revolution”.

At the end of 2014, the four-decade-old ‘PDB format’ was replaced by the more versatile mmCIF/PDBx format, which became the primary format for archive distribution. One immediate benefit of this transition was that large structures, which previously had to be split over multiple PDB entries, could be distributed as a single entry and file.

More data, roomier formats

In 2014 we curated a record 2712 PDB entries (up from 1788 in 2013, an increase of 52%). The total number of PDB depositions reached 10 291; PDBe's share of global depositions thus increased from 17% to 26%. In addition, users deposited 298 EMDB entries with PDBe (up from 268 in 2013, an increase of 11%), representing 44% of worldwide depositions.

A report of a PDBe-led workshop held in 2012 was published (Patwardhan et al., 2014), highlighting challenges and opportunities in archiving, validating, integrating, visualising and disseminating 3D cellular imaging data. The Medical Research Council (MRC) and Biotechnology and Biological Sciences Research Council (BBSRC) co-funded a grant based on the outcomes of this workshop, and one component of this work, a pilot archive for raw image data, was launched in October. Called the Electron Microscopy Pilot Image Archive (EMPIAR), its purpose is to provide easy access to state-of-the-art raw data to facilitate methods development, teaching and validation, which will lead to better 3D EM structures.

We worked with our EMDataBank partners and the EM community to improve and extend the EMDB data model and to handle 3D EM data and models (and their validation) in the new D&A system. We also developed a Fourier-Shell-Correlation server and a tilt-pair-validation server, and added new types of images (e.g. rotationally averaged power spectra and central slices) to the visual-analysis pages for EMDB entries.

In collaboration with our wwPDB partners, we implemented the handling of NMR data in the new D&A system. We also extended the wwPDB validation pipeline to handle more NMR data and tested it on all NMR structures in the PDB archive. We continued to work with the major developers of NMR structure-determination software to specify a unified format for NMR restraints, and expect this format to be adopted quickly as the standard.

Outreach

Our team organised seven PDBe roadshows in 2014 (Cambridge, Helsinki, Vienna, Brussels, Warwick, York and Oxford). PDBe website usage almost doubled between 2012 and 2014, and our social media content became more engaging for our growing following (over 2 000 followers on Facebook and Twitter). We hosted dozens of visitors and participants in workshops and meetings: the EMBO course on computational structural biology, the PDBe API workshop, the wwPDB Hybrids Methods Task Force meeting and the wwPDB Format

Gerard Kleywegt

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.



Selection of PDBe social-media highlights in 2014. **Centre:** As of 14 May 2014, the PDB had over 100 000 entries. **Top left:** On Virus Appreciation Day, PDBe released a 'fold your own virus' activity, which involved constructing a 3D paper model of a rhinovirus. **Top centre:** The football World Cup inspired this image of ryegrass mottle virus (PDB entry 2izw). **Top right:** Even very large structures, such as this rotavirus particle (PDB entry 4v7q), can now be represented as single entries using the versatile mmCIF format. **Bottom right:** New detector technology makes it possible to determine 3D EM structures with an unprecedented level of detail, rivalling that attainable with X-ray crystallographic methods (shown here, the tobacco mosaic virus: EMDB entry 2842). **Bottom centre:** To mark WHO Blood Donor Day, PDBe highlighted the enzyme 'histo-blood group ABO system transferase', which synthesises the ABO blood group antigen. **Bottom left:** A composite image of a poppy, comprising three PDB structures (entries 3zc9, 3bas and 4bqm), featured on Armistice Day.

Working Group meeting, in addition to Scientific Advisory Committee meetings for PDB and wwPDB. Our team members published six papers and served as ambassadors for EMBL-EBI and PDB, presenting posters and talks at many international conferences.

Future plans

To transform the structural archives into a truly useful resource for biomedical and related disciplines, we will continue to provide advanced services such as PDBePISA, PDBeFold and PDBeMotif, and develop powerful search and browse facilities. We will devote considerable efforts to: the annotation, validation and visualisation of ligand data; integration with other data resources; validation and presentation of information about the quality and reliability of structural data; and exposing experimental data in ways that clarify the extent to which experimental data support structural models and inferences.

Our work on the redesign of the PDBe website, internal database and weekly update process over the past few years

will come to fruition in 2015. The new website's design and functionality have been guided by user-experience research, and our new processes to update and improve the content and consistency of the structural and structure-related data have enabled us to make these data available to external developers through the new PDBe API (see Molecular structure and PDBe development).

Version 2.0 of the joint wwPDB D&A system, scheduled to go into production in 2015, will support X-ray, NMR and 3DEM data. Our next priority will be to implement the entire system at all wwPDB sites. Following a transition period in which the old and new systems operate in parallel, all structure depositions will be handled with the new system.

In 2015, wwPDB will organise a workshop on improving the validation of small-molecule ligands in the PDB, and PDBe will host a meeting of the wwPDB X-ray Validation Task Force as well as a community workshop to inform the development of EMPIAR.

Over the next few years, we expect the field of cellular structural biology to expand rapidly and have an increasing impact on biology and related fields. Hybrid approaches to structure determination will become much more common, and we are collaborating actively in this area so we can meet the challenges presented and embrace new opportunities as they arise.

Selected publications

Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2014) The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des* 28:1009-1014

Dutta S, Dimitropoulos D, Feng Z, et al. (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101:659-668

Gutmanas A, Alhroub Y, Battle GM, et al. (2014) PDBe: Protein Data Bank in Europe. *Nucl Acids Res* 42:D285-D291

Patwardhan A, Ashton A, Brandt R, et al. (2014) A 3D cellular context for the macromolecular world. *Nat Struct Mol Biol* 21:841-845



PDBe Content and Integration

The aim of PDBe is to serve the biomedical community by providing easy access to macromolecular and cellular structure data. Our team is responsible for curation of macromolecular structure data, and for ensuring that the PDBe web interface serves users well. We design new tools to facilitate access to integrated, high-quality structural data.

Major achievements

When the wwPDB partners put the X-ray crystallography module of the new common wwPDB deposition and annotation (D&A) system into production, our annotation staff started using it to annotate all European depositions. We used it on a record number of PDB entries (2712 entries, 26% of worldwide depositions). We also annotated 298 EMDB entries (43% of global depositions) in 2014.

In December 2014, the extensible mmCIF format replaced the old PDB format as the official distribution format of the archive entries. This enabled the wwPDB partners to 'merge' over 300 structures of large molecules, such as ribosomes, which had previously been saved as multiple entries due to the limitations of the old format, into single entries.

By making public the wwPDB validation reports for all X-ray crystallographic structures in the PDB archive, we enabled users to access quality information for all structures and compare them to select the entry with the highest quality data.

The infrastructure to integrate information about small molecule crystal structures from the Cambridge Structural Database (CSD) into the PDB Chemical Component Dictionary (CCD), which we developed in collaboration with the Cambridge Crystallographic Data Centre (CCDC), was made available to all the wwPDB partners in 2014. We continue to work closely with them to develop tools to integrate the CSD data in the annotation process.

One of our goals was to complete the necessary work to support the new PDBe API, query system and PDB and EMDB entry pages, and to that end we developed a process to address some data-consistency issues in the PDB archive. We helped update the PDBe database infrastructure to incorporate additional data from the PDB archive files and value-added data on structure quality (based on the wwPDB validation reports), assemblies, and binding sites. We worked with the EuropePMC team to enhance the citation information available for PDB entries. We also collaborated with the Systems and Networking team to develop a more robust infrastructure for the weekly release of the database.

PDBe's improved infrastructure now supports a new REST API, a new query system and newly formatted PDB entry pages, all of which will be released in 2015. The new query system provides rich archive-browsing functionality and presents facets that allow users to drill down to highly detailed data. The query engine integrates the structure-quality information so that the search results can be ranked accordingly. The search now offers different views of the result set based on macromolecules, small molecules, and Pfam sequence families.

We developed the PDBe REST API, which provides a uniform way of accessing PDBe including the improved archive data, value-added information on binding sites, assemblies and structure quality, and up-to-date cross-reference information from the SIFTS resource. The new pages for PDB and EMDB entries now use the API to access rich data content that is displayed on a user-friendly web interface. We developed many new features including interactive topology diagrams and a sequence-feature viewer, which make it easier for non-experts to understand structure information. We also integrated sets of newly developed images depicting a variety of structure-related information, which also improve the usefulness and usability of these pages.

Together with the Samples, Phenotypes and Ontologies team, we initiated a close collaboration with Flax, a Cambridge-based company that specialises in search technology. This collaboration will help us improve PDBe and extend the newly developed query system.

Sameer Velankar

PhD, Indian Institute of Science, 1997. Postdoctoral researcher, Oxford University, United Kingdom, 1997-2000.

At EMBL-EBI since 2000.

Team leader since 2011.



Future plans

In collaboration with our wwPDB partners, we will release a major update (version 2.0) of the D&A system that will handle NMR and 3D EM structures. This system will be implemented at all the wwPDB partner sites, and we expect an increase in the number of depositions handled by PDBe. The updated database infrastructure, newly designed query system, API and entry pages will provide a sound basis for future developments. We will update other major services such as PDBeMotif and integrate it with the new infrastructure. We will continue to improve the search mechanism, provide additional value-added information via the REST API, and enrich the information available on the PDB and EMDB entry pages. We plan to make our user communities aware of all the new developments through sustained engagement on social media and scholarly publications, and through face-to-face interactions at roadshows, workshops and at international life-science conferences.

Selected publications

Dutta S, Dimitropoulos D, Feng Z, et al. (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101:659-668

Gutmanas A, Alhroub Y, Battle GM, et al. (2014) PDBe: Protein Data Bank in Europe. *Nucl Acids Res* 42:D285-D291

Lewis TE, Sillitoe I, Andreeva A, et al. (2014) Genome3D: exploiting structure to help users understand their sequences. *Nucl Acids Res* 43:D382-D386

Meldal BHM, Forner-Martinez O, Costanzo MC, et al. (2014) The complex portal - an encyclopaedia of macromolecular complexes. *Nucl Acids Res* 43:D479-D484

Sen S, Young J, Berrisford JM, et al. (2014) Small molecule annotation for the Protein Data Bank. *Database*:1-11



Responsive design for a more flexible user experience: PDBe's new pages display information about entries according to screen size. On a larger screen, the newly designed interactive topology and sequence-feature viewer are displayed alongside the 3D molecular viewer. Figures from relevant publications and their captions are shown in context, along with other value-added data.



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 2014 our ChEMBL database of bioactive entities grew to over 1.5 million compounds, and fully incorporated the SureChEMBL database of patented chemical structures, originally developed by Digital Science. This resource has become an invaluable addition to the drug-discovery toolkit, and offers a substantial amount of information relevant to malaria and tuberculosis research.

We secured a second round of three-year funding to continue supporting metabolomics researchers with our MetaboLights resource, which represents the fastest-growing data platform at EMBL-EBI. This is an important development for those working in this emerging area of science, as we can extend our work to more sophisticated analysis and online visualisation and support community efforts to establish standards.

ChEMBL

ChEMBL, a quantitative database of bioactive compounds, provides curated bioactivity data linking compounds to molecular targets, phenotypic effects, exposure and toxicity end-points. ChEMBL focuses on interactions relevant to medicinal chemistry and clinical development of therapeutics. Pharmaceutically important gene families in ChEMBL can be viewed in the ADME, GPCR and Kinase SARfari web portals.

www.ebi.ac.uk/chembl

ChEBI

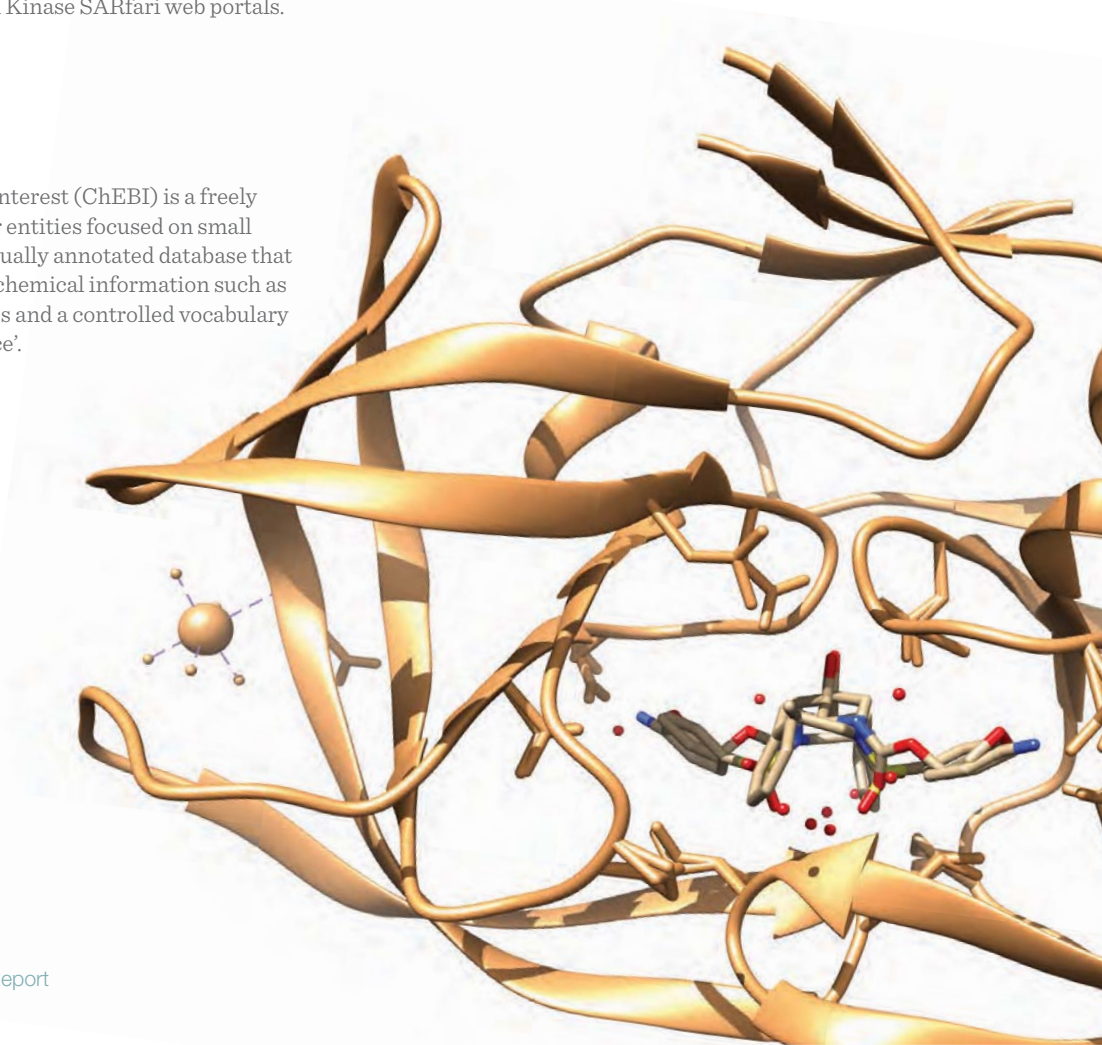
Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on small chemical compounds. It is a manually annotated database that provides a wide range of related chemical information such as formulae, links to other databases and a controlled vocabulary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

www.ebi.ac.uk/chebi

MetaboLights

MetaboLights is a resource for Metabolomics experiments and derived information. It is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations.

www.ebi.ac.uk/metabolights





ChEMBL

Team leader: John Overington

- Operated the SureChEMBL patent database of chemical structures donated by Digital Science to EMBL-EBI, integrated with ChEMBL and extended search functionality;
- Grew the ChEMBL database to over 1.5 million compounds and over 13 million bioactivity values, largely thanks to extraction of data from the scientific literature and deposition of additional bioactivity data for compounds tested for malaria and tuberculosis;
- Added coverage of agrochemical SAR data from the primary literature via a collaboration with Syngenta;
- Expanded the content of the UniChem structure cross-referencing resource to provide cross references to more than 87 million chemical structures from 27 source databases;
- Prototyped the application of the Pistoia Alliance Hierarchical Editing Language for Macromolecules (HELM) representation for peptides in ChEMBL content;
- Contributed bioactivity data and services to the RDF-based OpenPHACTS, an Innovative Medicines Initiative project;
- Extended UniChem search functionality to include relaxed stereochemistry/salt matching;
- Further developed the myChEMBL virtual machine, improving the user interface, web services and database performance;
- Secured long-term funding for ChEMBL and SureChEMBL via a five year grant from the Wellcome Trust.

www.ebi.ac.uk/chembl

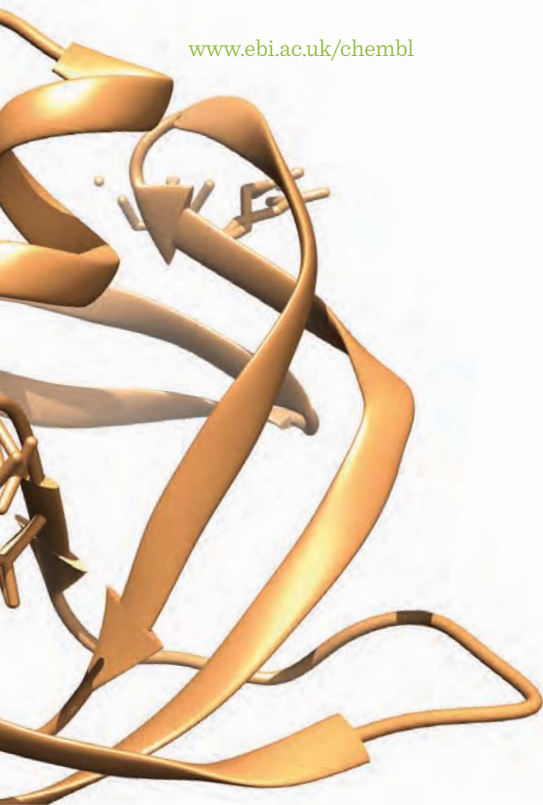
Cheminformatics & Metabolism

Team leader: 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;
- Successfully completed the implementation of the Enzyme Portal and helped move the project to the UniProt Development team;
- Secured three-year funding for MetaboLights, supporting development for online analysis, visualisations and online study curation;
- Congratulated two of our PhD students, Stephan Beisken and John May, for successfully completing their studies and receiving their PhD degrees from the University of Cambridge.

www.ebi.ac.uk/chebi

www.ebi.ac.uk/research/steinbeck/software





ChEMBL

Drug discovery is more costly than ever, and achieving innovation in efficacy and safety remains a significant challenge. Changes in the structure of the pharmaceutical industry over the past decade have led to an increase in drug-discovery activities in organisations that typically do not have access to large databases of legacy bioactivity data and experienced staff. Our team develops and manages ChEMBL, EMBL-EBI's database of quantitative small-molecule bioactivity data focused in the area of drug discovery.

ChEMBL contains data on curated chemical structures, bioactivity values and their relationship to biological targets and phenotypic assays. In 2014 we assumed responsibility for a new database, SureChEMBL, following a donation of the technology from Digital Science. SureChEMBL contains full patent text combined with 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 and drug safety.

Major achievements

In 2014 we received further long-term funding for ChEMBL through a Strategic Award from the Wellcome Trust.

The biggest change in 2014 was the transfer of full operation of the SureChEMBL patent database of chemical structures from Digital Science to EMBL-EBI. This world-leading, free resource is now available to the drug discovery community in a wide variety of forms and downloads. We ensured the complementarity and distinctiveness of ChEMBL and SureChEMBL were maintained, with very rapid cross linkage via the enhanced UniChem resource.

During 2014 the content of the ChEMBL database increased again by approximately 100 000 compounds and 2 million bioactivity values. There were substantial increases in the extraction of data from the scientific literature and in the deposition of bioactivity data for compounds tested for neglected diseases such as malaria and tuberculosis. UniChem, originally developed to support the EU-OPENSREEN chemical biology infrastructure, is our cross-EMBL-EBI chemical structure cross-referencing service, and now

provides database cross-references to more than 87 million chemical structures from 27 source databases, both internal and external to EMBL-EBI. UniChem is also a core part of chemical structure integration for the BioMedBridges project.

The SureChEMBL system is now widely used by researchers worldwide, and is our first experience of a large-scale cloud application running on Amazon Web Services (AWS). We started to extend the added-value annotation to include targets, diseases and disease models.

We extended changes to the ChEMBL data model to allow for better representation of protein families and complexes. We also added BioAssay Ontology (BAO) indexing of many assays, and classified cell-line assays with standard ontology terms; this enhanced search-and-retrieval functionality. We also incorporated some algorithms from our research group into ChEMBL, allowing the identification of the ligand-binding domain for a small molecule; this greatly simplifies searching and results processing.

myChEMBL, the fully open-source version of ChEMBL on a virtual machine, became popular with users as it allows the running of our systems in entirely private settings – an important feature for drug-discovery research, where novelty and intellectual property are essential.

Optimising drug metabolism and distribution in body tissues in preclinical species and understanding how this will translate to humans can be a bottleneck to clinical development. ADME SARfari, published in 2014, allows users to compare species differences in these protein sequences, their variants and tissue distribution. We also had further contributions of ADME data from AstraZeneca.

We continued to participate in three EU-funded projects: eTOX, diXa, and HeCaTos, all of which aim to better curate toxicity data and use it to predict toxicity. We also continued our work on OpenPHACTS, an IMI project that integrates pharmacological data across diverse resources, and contribute to the BioMedBridges and EU-OpenScreen infrastructure projects.

John Overington

BSc Chemistry, Bath. PhD in Crystallography, Birkbeck College, London, 1991. Postdoctoral research, ICRF, 1990-1992. Pfizer 1992-2000. Inpharmatica 2000-2008.

At EMBL-EBI since 2008.



Research

We studied the differential expression of all known, pharmacologically relevant and ADMET genes in mouse development, from pre-birth to natural old age. Building on this work, we undertook analysis of changes in the expression patterns of this gene set that could point towards differential efficacy and safety in paediatric and geriatric human populations.

We apply computer science methodologies to address drug and indication discovery, extracting knowledge from biomedical articles and drawing on ontologies and description logics to formalise biological information. As part of this work, we developed the Functional Therapeutic Chemical Classification System resource, which specifically addresses drug repositioning.

We extended methods to predict molecular targets from phenotypic screening data, an essential step in lead optimisation. Working on public-private partnership data, we validated two cases of target prediction for potential tuberculosis drug leads, including enzymology and structural biology studies. We also developed a general theoretical model of drug resistance and drug combinations that offer novel applications in improving drug safety and countering drug resistance in areas such as oncology and antibiotics.

Future projects and goals

In 2015 we will continue to broaden the utility and content of ChEMBL and SureChEMBL by adding additional annotation, for example on diseases and targets. 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 also scale UniChem to billion compound scale storage and integration capabilities.

Our research will focus on translational and safety biology, with extension to areas including the study of the impact of human genetic and physiological variation on drug efficacy, target validation and safety.

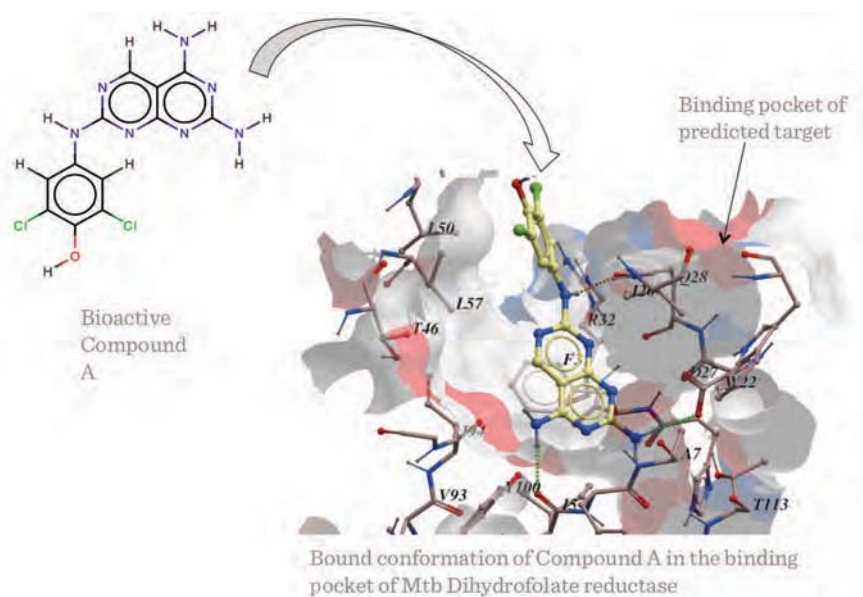
Selected publications

Bento AP, Gaulton A, Hersey A, et al. (2014) The ChEMBL bioactivity database: an update. *Nucl Acids Res* 42:d1083-d1090

van Westen GJ, Gaulton A and Overington JP (2014) Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput Biol* 10:e1003559

Costello JC, Heiser LM, Georgii E, et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 32:1202-1212

Bansal M, Yang J, Karan C, et al. (2014) A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol* 32:1213-1222





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 also 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 chemicals 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

The ChEBI database continued to grow, with 5400 entries added during 2014, bringing the total number of fully curated entries to over 42 000. Direct data submissions from ChEBI users now account for around 30% of new entries, with the remainder arising from user requests via SourceForge and other channels. We also added cross-references from ChEBI to several important metabolism- and chemistry-related resources.

We automated the release procedures for ChEBI, which enabled us to replace the monthly releases with live updates from December 2014. New entries and updates to existing ones are now visible on the website as soon as they have been created. This is a major benefit to users who need to include a ChEBI identifier in their own resource, as they can add it immediately. ChEBI developers also benefit, as they can devote more time to developing new tools that improve the efficiency of the curation process. This is of particular importance given the rapid growth of MetaboLights, which uses ChEBI for storing reference data on metabolites.

We added a 'species table' to the ChEBI schema to handle information on species, strain, organism part and literature references for any naturally occurring entity. This information, which is fully searchable, is now included in ChEBI as part of the curation process.

We expect to see a significant increase in the amount of data that is 'bulk loaded' into ChEBI, and accordingly have introduced an automatic classifier so that such data is immediately classified within the ChEBI ontology. This enhances the utility of bulk-loaded data to ChEBI users during the period the bulk-loaded entries await manual updating.

MetaboLights

MetaboLights reference compounds and studies have associated organism information. To help researchers explore this information more easily, we developed species search functionality based on model organisms; free-text search; and a tree of life that users can browse, with automatic classification that is compatible with any taxonomic identifier present in 89 different taxonomy sources.

By December 2014 there were over 15 000 compounds linked to ChEBI. MetaboLights enriched compounds with 5355 spectra from MassBank and 406 from BMRB (NMR). We began using a new JavaScript MS / NMR spectra viewer, which was supported by the BBSRC Tools and Resources Development fund (SpeckTackle grant).

We made substantial progress in creating online analysis tools for MetaboLights; our first approach is adapting community-accepted toolsets, for example MetaboAnalyst (David Wishart group, University of Alberta). We also began incorporating 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

There is a need for standards and recommended procedures in the emerging discipline of metabolomics data management, and in 2014 we secured funding for the COSMOS consortium, an EU-funded endeavour for the coordination of standards for metabolomics. The consortium, led by EMBL-EBI, combines the efforts of 14 leading European labs in Metabolomics, and

Christoph Steinbeck

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.



successfully delivered new standards including the NMR-ML open standard for Nuclear Magnetic Resonance data. It also delivered agreed procedures for the management and dissemination of data in metabolomics.

The success of COSMOS led to the formation of the MetabolomeXchange consortium, which continues the EMBL/GenBank/DDBJ tradition of global data exchange in biology.

Outreach

We represented MetaboLights and ChEBI at 18 scientific conferences, training sessions and knowledge-exchange events in 2014. These took place in the UK and other European countries as well as Brazil, Canada, China, India, Japan and the US. MetaboLights is now the recommended metabolomics repository for *Nature Scientific Data* and the *Metabolomics Journal*.

Future plans

Our research will focus on 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 select projects that are aligned with the goals of our services. We will work on extending the ChEBI database to offer greater utility 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.

We will undertake our MRC-funded work with the Phenome Centres at Imperial College London in 2015. This collaboration will be instrumental in making information on the metabolomes of a large number of human cohorts available to the public.

Selected publications

Griss J, Jones AR, Sachsenberg T, et al. (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* 13, 2765-2775

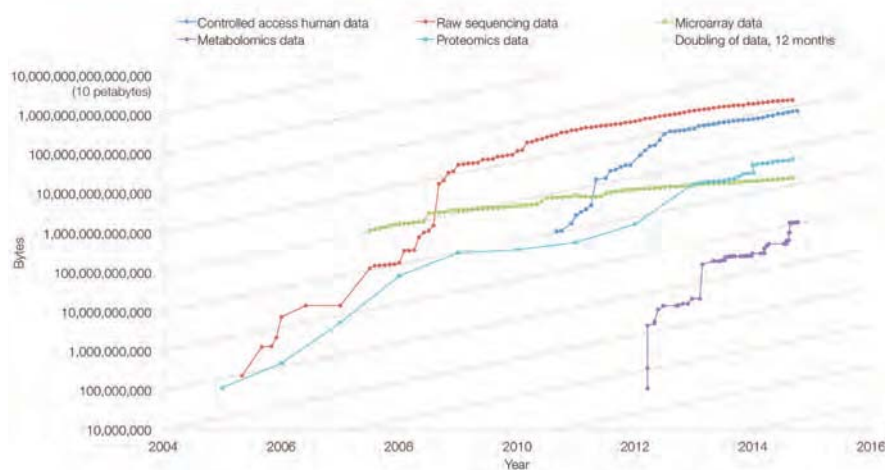
Jayaseelan KV, Steinbeck C. (2014) Building blocks for automated elucidation of metabolites: natural product-likeness for candidate ranking. *BMC Bioinformatics* 15:234

Rueedi R, Ledda M, Nicholls AW, et al. (2014) Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet* 10, e1004132

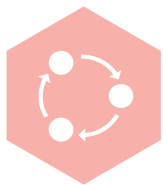
Truszkowski A, Daniel M, Kuhn H, et al. (2014) A molecular fragment cheminformatics roadmap for mesoscopic simulation. *J Cheminform* 6:45

Venkata C, Forster MJ, Howe PW, Steinbeck C. (2014) The potential utility of predicted one bond carbon-proton coupling constants in the structure elucidation of small organic molecules by NMR spectroscopy. *PLoS One* 9:e111576

Growth of data, by platform



Metabolomics data is doubling faster than any other data platform at EMBL-EBI.



Molecular systems

EMBL-EBI hosts several data resources that focus on the interactions and functional connectivity of biomolecules, including protein, oligonucleotide and small molecule components. These resources provide freely available data and information on the coordinated self-organisation and assembly of multi-component, regulated and emergent complex biological functions, drawing on resources such as ChEBI, UniProt and Ensembl.

We provide co-ordinated resources for systems biology at increasing levels of abstraction: physical molecular interactions in IntAct, connected interactions making up Reactome pathways, and quantitative, dynamic modelling of interaction networks in BioModels.

We follow an open-source, open-data approach, and participate actively in the development of community standards. We engage with research communities, journal editors and funding organisations to encourage the deposition of experimental data in our public, standards-compliant data resources so that they may be shared easily with scientists worldwide.

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.

<http://reactome.org>

BioModels

BioModels Database is a repository of peer-reviewed, published, computational models, primarily from the field of systems biology. BioModels allows biologists to store, search and retrieve mathematical models covering a wide range of systems. In addition, the database can be used to generate sub-models; models can be simulated online and can be converted between different representational formats. This resource also features programmatic access via web services.

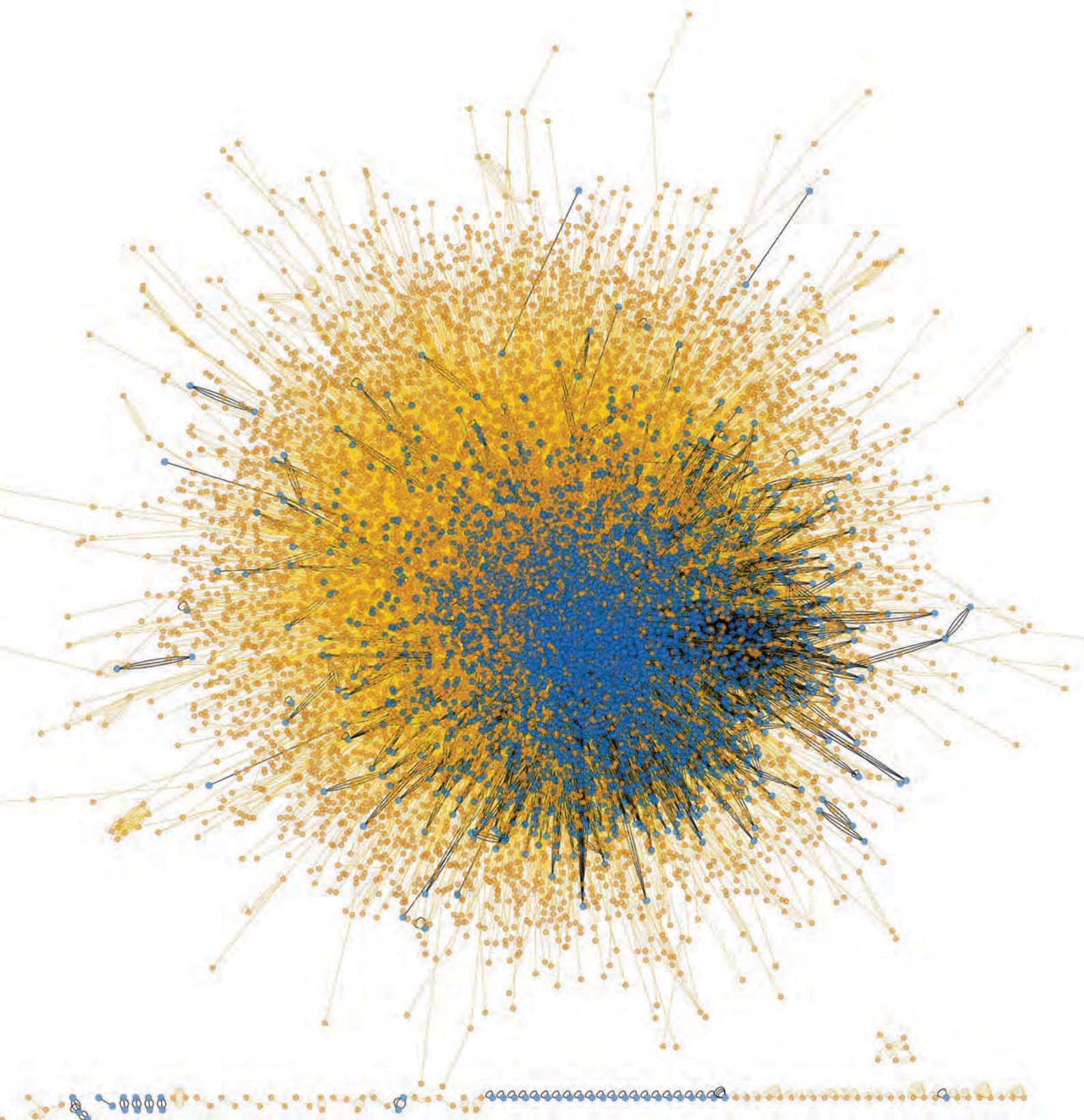
www.ebi.ac.uk/biomodels

Proteomics Services

Team Leader: Henning Hermjakob

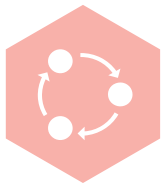
- *As part of the ProteomeXchange consortium, PRIDE team members processed almost 1000 submissions in 2014, 80% more than in 2013;*
- *Co-developed the qcML and mzTab community formats for quality assessment and summation of mass spectrometry based proteomics experiments (PRIDE team members);*
- *Released the Complex Portal, a new resource of manually curated molecular complexes (IntAct team members);*
- *Completely redeveloped the Reactome Analysis tool to allow genome-scale analysis within seconds;*
- *Celebrated the tenth anniversary of the BioModels database, and the entry of its 1000th literature-based computational model.*





Visualisation of human interactome data held in the IntAct database. Yellow: all interactions. Blue: Interactions present in *A proteome-scale map of the human interactome network* (Rolland T., et al., 2015). [Image: Pablo Porras, EMBL-EBI]





Proteomics 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.

The team is a major contributor to community standards, in particular the Proteomics Standards Initiative (PSI) of the international Human Proteome Organisation (HUPO) and systems biology standards (COMBINE Network). We provide public databases as reference implementations for community standards: the PRIDE proteomics identifications database, the IntAct molecular interaction database, the Reactome pathway database and BioModels Database, a repository of computational models of biological systems.

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

ProteomeXchange, IMEx and the Complex Portal

A major success in 2014 was the successful conclusion of the EU ProteomeXchange grant with a high-impact consortium publication (Vizcaino et al, 2014). As part of the ProteomeXchange consortium, PRIDE works with US partners (i.e. PeptideAtlas, ISB, Seattle; MassIVE, UCSD, San Diego) to co-ordinate data deposition and dissemination strategies for mass spectrometry data, providing a single entry point for deposition, a shared accession-number space and a deposition-metadata format. In 2014, ProteomeXchange processed 963 submissions, 80% more than in 2013. PRIDE downloads reached 127 Terabytes, with some datasets being downloaded more than 250 times. To support automated quality assessments and summarisation of proteomics experiments, we co-developed the qcML (Walzer et al, 2014) and mzTab (Griss et al, 2014) community formats.

The IMEx collaboration (Orchard et al., 2014) for the internationally co-ordinated curation of molecular interaction data continued to be a major activity. IMEx partners share formats, identifier spaces and curation strategies, and many share the web-based IntAct curation infrastructure directly. This avoids redundant development while retaining the value of each individual resource. IMEx partners include MINT (University of Rome), UniProt (SIB-Swiss Institute of Bioinformatics, EMBL-EBI), I2D (Ontario Cancer Institute), InnateDB (Collaborators in Australia, Canada and Ireland), Molecular Connections Ltd. (India) and MechanoBio (National University of Singapore).

In 2014 we released the Complex Portal (Meldal et al., 2014), a new resource based on the IntAct curation platform for curated molecular complexes. The Complex Portal, which was launched with over 1000 annotated complexes, is a joint effort with contributions from UniProt, PDBe, ChEMBL, and MatrixDB (University of Lyon).

Reactome and BioModels

Reactome provides review-style, curated, peer-reviewed human pathways in a computationally accessible form (Croft et al, 2014). In 2014, Reactome coverage increased to almost 8000 human proteins (sourced from UniProt), and our team focused on curating modified pathways in human disease. We released a completely redeveloped Reactome pathway analysis tool, which allows genome-scale over-representation analysis within seconds. This tool is now accessible via web services and links directly from PRIDE and the EU-funded BLUEPRINT project.

The BioModels database, a standards-compliant resource for systems biology models, celebrated its tenth anniversary and welcomed its 1000th literature-based model in 2014. We redeveloped the BioModels homepage, providing clearer entry points based on model status and Gene Ontology classification. We also further developed the JUMMP software platform, which is a flexible infrastructure for the next version of BioModels. JUMMP provides the basis for the IMI-funded Drug Disease Model Resources (DDMoRe) project's model repository, which was released in 2014. BioModels team members also released an update of identifiers.org, a system that enables the stable resolution of identifiers used to describe life-science objects.



Outreach and training

In 2014, the Proteomics Services Team contributed to 27 training events, which collectively reached 726 scientists. Sixteen of these were courses held in the UK, eight were hosted elsewhere in Europe and three were held outside of Europe. We contributed content to two new online training courses, participated in three webinars and presented in five Industry Programme workshops and events.

Future plans

We will continue to develop PRIDE to provide a cross-experiment view of protein identification data for human and model organisms. We will complete redevelopment of the PRIDE-Cluster algorithm, which will support the dramatic increase in proteomics data volume. We will reach out to researchers who submit data, implementing and sharing generic quality reports and performing 'spectral library' searches using the PRIDE Cluster algorithm.

We will redevelop the IntAct interface to emphasise graphical data presentation, building on pilot efforts in the Complex Portal. We will also initiate work to deepen the data model for interactions so that it allows reasoning across data, for example extending the scope of curation from undirected to directed interactions to enable causal reasoning on networks. We also plan to develop the complex portal into a reference resource for molecular complexes, similar to UniProt for protein sequence annotation, and ChEBI for chemical entities.

To establish Reactome as a primary entry point for the interpretation of large-scale biomolecular datasets, we will completely redevelop the pathway visualisation module and revamp the release process. This work, which follows

the successful refactoring of the web interface in 2013 and the pathway analysis tool in 2014, will make the underlying processes behind the resource much faster and more stable. Our curators will continue to map disease and dysregulation on Reactome pathways; the optimal presentation of modulation of complex pathways will be an active research topic.

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. These efforts greatly facilitate collaboration through the sharing of computational models.

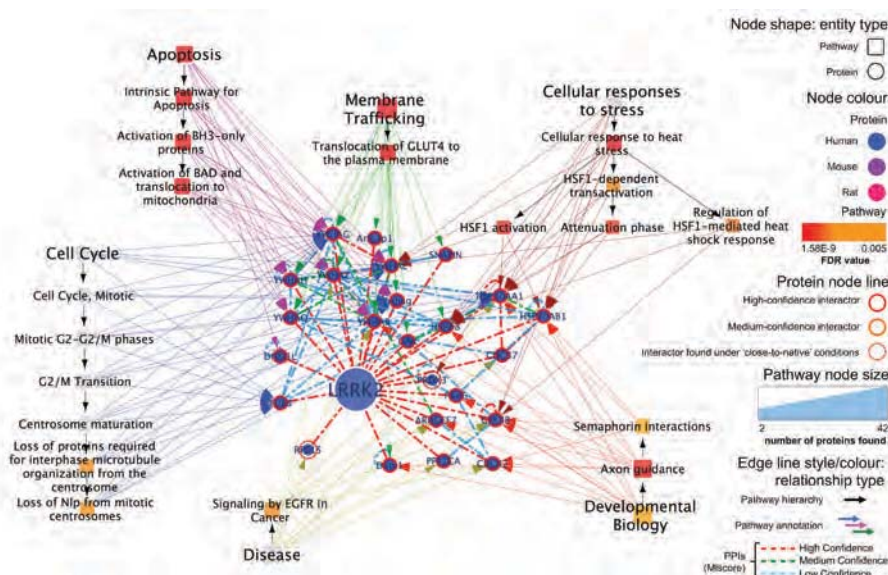
Selected publications

Meldal BH, Forner-Martinez O, Costanzo MC, et al. (2014) The complex portal - an encyclopaedia of macromolecular complexes. *Nucl Acids Res* 43:D479-D484

Griss J, Jones AR, Sachsenberg T, et al. (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* 13:2765-2775

Walzer M, Pernas LE, Nasso S, et al. (2014) qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Mol Cell Proteomics* 13:1905-1913

Vizcaíno JA, Deutsch EW, Wang R, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32:223-226



Integrated visualisation of the LRRK2 pathway and interactome context. LRRK2 is a protein of largely unknown function with strong genetic associations to Parkinson's disease. Manually curated LRRK2 high confidence partners and the interactions between them from IntAct are represented associated to Reactome pathways that appear to be significantly enriched for this group of proteins. Each pathway is associated with annotated proteins through colored arrows and Reactome's pathway hierarchy is shown using black arrows.



Cross-domain tools & resources

EMBL-EBI is unique in offering a comprehensive range of high-quality data resources and 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.

The scientific literature provides the essential context for research, and we endeavour to ensure publications are linked with their underlying data. In 2014 Europe PMC became EMBL-EBI's most highly accessed service, helping researchers all over the world to explore molecular data in new ways. This resource, which builds innovative layers on PubMed Central and other literature databases, represents the largest public-domain citation network in the world.

Our contributions to projects for the Centre for Therapeutic Target Validation (CTTV) included the creation of mapping tools to ensure our public resources can link meaningfully with proprietary disease knowledge bases, and a bespoke phenotype–disease annotation. These tools create the bridges necessary to work on target discovery at scale.

Europe PubMed Central

Europe PubMed Central contains more than 30 million abstracts and 3 million full text articles. The abstracts component includes all of PubMed, agricultural abstracts from Agricola and patents from the European Patent Office. About 900 000 of the full-text articles are open access, so they are free to read and to reuse in ways such as text mining. Europe PMC is supported by 26 European funding organisations, whose commitment supports their own Open Access mandates. Our goal is to provide fast and powerful access to the literature, as well as features and tools that place the article narratives in the wider context of related data and credit systems such as article citations.

europepmc.org

Gene Ontology

GO is a major bioinformatics initiative that aims to unify the representation of gene and gene-product attributes across all species. Groups participating in the GO Consortium include major model organism databases and other bioinformatics resource centres. At EMBL-EBI, the editors play a key role in managing the distributed task of developing and improving the GO ontologies.

www.ebi.ac.uk/go

Experimental Factor Ontology

EFO is a data-driven ontology that imports parts of existing community ontologies into a single framework. It is used for annotation, curation, query and visualisation of data by ArrayExpress, the Gene Expression Atlas, NHGRI GWAS Catalog, BioSamples database and CTTV.

www.ebi.ac.uk/efo

BioSamples Database

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

BioStudies Database

The EMBL-EBI BioStudies database is complementary to the EMBL-EBI BioSamples database. It groups datasets by study, provides metadata describing these studies and links to the actual data either in structured databases (for types of data for which such databases exist) or unstructured data files. The data may be stored either within or outside of EMBL-EBI.

www.ebi.ac.uk/biostudies

Literature Services

Team Leader: Johanna McEntyre

- Became the most highly accessed EMBL-EBI service;
- Implemented a new unified search system that allows searching across both abstracts and full text in a straightforward, intuitive way;
- Implemented a new export feature that allows search results to be selected and sent to file or email in a variety of useful formats (e.g. text, reference manager compatible, tab-delimited);
- Integrated unique author identifiers to the extent that Europe PMC now represents the most extensive public integration of ORCIDs in the world, with over 1 million articles now listing at least one ORCID;
- Updated the standard for encoding an article in XML (the JATS DTD) to include provision for citing data according to the Force11 Data Citation Principles.

www.europepmc.org

Functional Genomics Development

Team leader: Ugis Sarkans

- Released the first version of the BioStudies database in beta.

Samples, Phenotypes & Ontologies

Team Leader: Helen Parkinson

- Enhanced the Gene Ontology with new content for apoptosis, cilia and viruses;
- Worked with Roche to support industry use of the Gene Ontology in drug discovery;
- Improved representation of human intestinal parasites in the Gene Ontology;
- To support the Centre for Therapeutic Target Validation (CTTV), developed mapping tools and new ontology content to support disease annotations in EMBL-EBI databases including the European Variation Archive, UniProt and Reactome;
- Designed a phenotype-disease annotation representation for CTTV;
- In collaboration with Orphanet, released the Orphanet Rare Disease Ontology (ORDO) version 2.0;
- Improved access to EMBL-EBI's RDF platform via the AtlasRDF BioConductor package;
- In collaboration with BioMedBridges, delivered CMPO and Phenotator, an ontology and supporting tool for describing cellular phenotypes;
- Implemented human-mouse disease queries in the IMPC portal, connecting mouse models to clinical users;
- Delivered a BioConductor statistical analysis package, PhenStat, for analysis of high-throughput phenotyping data;
- Visualised the knockout mouse data for the NIH project Illuminating the Druggable Genome;
- Delivered a new user interface and supporting tooling for the NHGRI GWAS Catalog.

www.biomedbridges.eu
www.ebi.ac.uk/rdf
www.ebi.ac.uk/biosamples
www.ebi.ac.uk/efo
www.infrafrontier.eu
www.mousephenotype.org

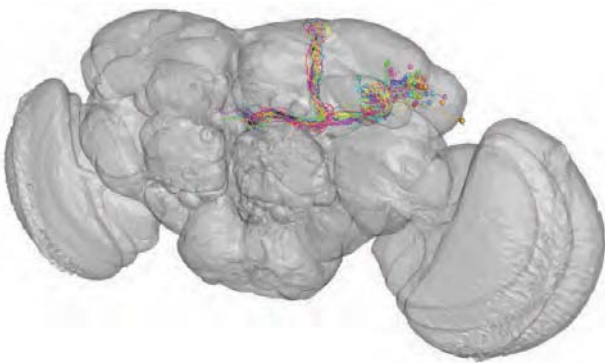


Samples, Phenotypes & Ontologies

The Samples, Phenotypes and Ontologies team has three major activity strands: BioSamples and semantic data integration, mouse informatics and the Gene Ontology Editorial Office. We focus on metadata integration, ontology development and supporting tooling, as well as development and delivery of content for the BioSamples database and mouse data for the biomedical research community.

Our team's activities diversified as the team grew in 2014, when we received new funding for two collaborative projects: From the Wellcome Trust funding for Virtual FlyBrain (VFB) and from the US National Institutes of Health (NIH) for the National Human Genome Research Institute (NHGRI) GWAS Catalog. VFB provides integrated access to fly brain anatomy, expression and genetic data. The NHGRI GWAS Catalog is a curated resource of disease-associated variants, and will move to EMBL-EBI in 2015 with content delivered by both EMBL-EBI and NHGRI.

Like the Literature Services team, we operate in the cross-domain space such that our work has overlap with all the major EMBL-EBI resources. We provide ontology resources such as the Gene Ontology and Experimental Factor Ontologies, and sample/phenotype resources such as the BioSamples database, Infrafrontier (formerly the European Mutant Mouse Archive, EMMA) and the International Mouse Phenotyping Consortium (IMPC), among others.



Virtual FlyBrain mushroom body intrinsic neurons, identified via neuron blast (see Manton et al., 2014). Raw data provided by the Chiang group at National Tsing Hua University, Taiwan (see Chiang et al., Current Biology 2011). Registration and statistical image analysis were performed by G. Jefferis of the Medical Research Council and M. Costa of the University of Cambridge.

Major achievements

Supporting target validation

In 2014 we put considerable effort into enhancing ontology services and integrating them with industry efforts such as the Centre for Therapeutic Target Validation (CTTV) and Roche drug discovery. The Gene Ontology now has content for apoptosis, cilia and viruses, and its representation of human intestinal parasites is much improved. We developed mapping tools and new ontology content to support CTTV's disease annotations in EMBL-EBI databases including the European Variation Archive, UniProt and Reactome. We also designed a phenotype-disease annotation representation for this public-private partnership.

Part of our work in 2014 involved creating visualisations for knockout mouse data in the context of Illuminating the Druggable Genome (IDG), a project funded by the NIH to explore the properties and functions of unannotated proteins in the most commonly drug-targeted protein families. Our efforts were part of the pilot program to create a Knowledge Management Center that will catalogue known information about these protein families to help investigators determine whether a given protein is of interest.

Rare diseases

In collaboration with Orphanet, the European portal for rare diseases and orphan drugs, we released the Orphanet Rare Disease Ontology (ORDO) version 2.0. The aim of this project is to provide a structured vocabulary for rare diseases that captures relationships between diseases, genes and other relevant features. While ORDO is updated monthly, version 2.0 was a major release that featured the addition of several new classes and relationships. We added epidemiology data such as annual incidence, case/family, prevalence at birth and lifetime prevalence. We also added gain/loss of gene function along with chromosomal location for the gene, subtypes for genetic material and geographical location. We also included annotations for database cross-reference mapping types, and more modes of inheritance.

Helen Parkinson

PhD Genetics, 1997. Research Associate in Genetics, University of Leicester 1997-2000.

At EMBL-EBI since 2000.



Semantic Web

Our team has been deeply involved in the EMBL-EBI RDF Platform, which provides a unified way to query across EMBL-EBI resources that provide access to their data using Semantic Web technologies. In 2014 we improved access to the platform via the AtlasRDF BioConductor package. This allows researchers who are exploring gene expression data to extend their query beyond genes to experimental factors (e.g. disease, cell type and compound treatments), pathways, proteins and small molecules. We also implemented functionality that allows users to enrich a gene list across Experimental Factor Ontology (EFO) using the Atlas background set.

BioMedBridges

BioMedBridges is a joint effort between 12 biomedical sciences research infrastructures to develop the shared e-infrastructure—the technical bridges—for data integration in the biological, medical, translational and clinical domains. In 2014 we collaborated with BioMedBridges to deliver the Cellular Microscopy Phenotype Ontology (CMPO), which provides a species-neutral, controlled vocabulary for describing phenotypic qualities relating to the whole cell, cellular components, cellular processes and cell populations. CMPO terms are now used to annotate phenotype descriptions from high-content screening databases and cellular image repositories. We also collaborated with BioMedBridges on a new ontology and supporting tool for describing cellular phenotypes: Phenotator, which had its first formal release in 2014.

Mouse informatics

We continued to develop the International Mouse Phenotyping Consortium web portal, and in 2014 we implemented new functionality for human–mouse disease queries. This helps clinical users find the most relevant mouse models easily.

We delivered a BioConductor statistical analysis package, PhenStat, for the analysis of high-throughput phenotyping data. High-throughput phenotyping generates large volumes of varied data including both categorical and continuous data. PhenStat provides statistical methods for the identification of abnormal phenotypes with an emphasis on high-throughput data flows. After testing and demonstrating the viability of the PhenStat with an application of 420 lines of historic mouse phenotyping data, we published the software, which offers dataset processing, statistical analysis and results output.

NHGRI GWAS Catalog

We delivered a new user interface for the NHGRI GWAS Catalog, and retooled it to better serve the needs of biologists who may not be experts in bioinformatics. We rebuilt the new Catalog to provide comprehensive search functionality, made possible by an underlying ontology and improved data modelling. We also laid the groundwork for moving this public resource to EMBL-EBI in 2015.

Future plans

In 2015 we will continue to work with disease communities to improve the utility of the open disease ontologies, and apply these ontologies to EMBL-EBI and CTTV databases. ‘GO slims’ are lists of GO terms that have been selected from the full set of terms available, and are used to generate a focused view of part of the GO or a broad overview of all of the top categories. Following this model, we will generate a ‘disease slim’ to support EMBL-EBI’s annotation, query and analysis needs. This will be for both in-house and community use to integrate disease data at EMBL-EBI.

The IMPC mouse phenotyping project will issue a major data release of phenotypic data for 1700 new knockout lines, supported with 98 000 images from knockout and control animals. This information will enable researchers to dissect aspects of function for uncharacterised genes and carry out fine mapping of animal phenotypes to human disease.

A major activity in 2015 will be hosting the NHGRI GWAS Catalog at EMBL-EBI, which will bring with it changes including a new user interface and data processing infrastructure. These will improve access for users and enable integration with resources such as Ensembl and EuropePMC.

Selected publications

Deans AR, Lewis SE, Huala E, et al. (2015) Finding Our Way through Phenotypes. *PLoS Biol* 13:e1002033. Published online in 2014.

INFRAFRONTIER Consortium (2015) INFRAFRONTIER—providing mutant mouse resources as research tools for the international scientific community. *Nucl Acids Res* 43:d1171–d1175. Published online in 2014.

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucl Acids Res* 43:d1049–d1056. Published online in 2014.

Kibbe WA, Arze C, Felix V, et al. (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucl Acids Res* 43:d1071–d1078. Published online in 2014.



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. Because the scientific literature represents the knowledge layer of research reporting, it is critical to link articles with underlying and related data in meaningful ways. We achieve this by working collaboratively with other service providers, publishers and databases, and by engaging with the scientific community. In this regard, the databases developed at EMBL-EBI and more widely through ELIXIR are of primary importance.

Europe PMC grows at a rate of over 1 million abstracts and 270 000 full-text articles per year, and as of December offered more than 30 million abstracts and around 3 million full-text research articles from PubMed and PubMed Central. It includes metadata for biological patents, clinical guidelines, PhD theses and research reports.

Europe PMC layers article-citation networks, links to underlying databases and makes text-mined terms of biological interest discoverable. It provides programmatic access via SOAP and REST web services, and allows users to bulk-download open-access articles via FTP. Users can also search over 50 000 biomedical research grants that have been awarded to approximately 20 000 PIs supported by Europe PMC's 26 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.

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 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.

Major achievements

In 2014 Europe PMC became the most accessed website at EMBL-EBI, visited by over 1 million unique IP addresses every month. We ensured that over 10 million searches and page views were served every month by this highly reliable resource, which frequently exceeded targets of 99.7% uptime per month.

We played a leading role in the development of ORCID, an open, non-profit, community-based effort to provide a registry of unique researcher identifiers. As of December 2014, over a million articles in Europe PMC had an ORCID claimed by at least one author and around 20 000 researchers had used the Europe PMC claiming tool to add articles to their ORCID record. We worked with the ORCID Foundation to improve both data management protocols and user experience design. EMBL adopted ORCIDs throughout the organisation in 2014, a process in which our team played a central technical role.

We implemented major improvements to Europe PMC's search and download capability. It is now possible to search by precise publication date, date range and 'first available' date. We also made it possible to search by article licence type, for example CC-BY, which is a critical value to determine the scope for reuse. Users can now select any set of search results for downloading with a few clicks, saving their results in the most suitable format for their work. For example, we now offer formats that are compatible with reference management software, and plain text for quick insertion into a document. Users can also easily export the full-text XML files of open access articles.

Integration

We devoted considerable efforts to completely reorganising the Europe PMC full-text article-processing pipeline, and to integrating full text articles with metadata. This brought immediate benefit to users, who can now search across all content combined and see the results in a single, unified search results list. This removes confusion caused by unaligned results lists (i.e. one for abstracts and one for full text), and makes it easier to find results from complete articles, rather

Johanna McEntyre

PhD in plant biology, Manchester Metropolitan University, 1990. Editor, Trends in Biochemical Sciences, Elsevier, Cambridge, United Kingdom, 1997. Staff Scientist, NCBI, National Library of Medicine, NIH, United States, 2009.

Team Leader at EMBL-EBI since 2009.



than just abstracts. This work was foundational for future developments that will require a unified concept of an 'article', rather than different models depending on the source of data.

To make it easier for institutional repositories to import Europe PMC content programmatically, we developed a Dublin Core response format for the web services. This helps ensure repositories can have accurate and up-to-date information about articles. Europe PMC also supports a full-text books and documents database now, which is useful to both researchers and institutional repositories.

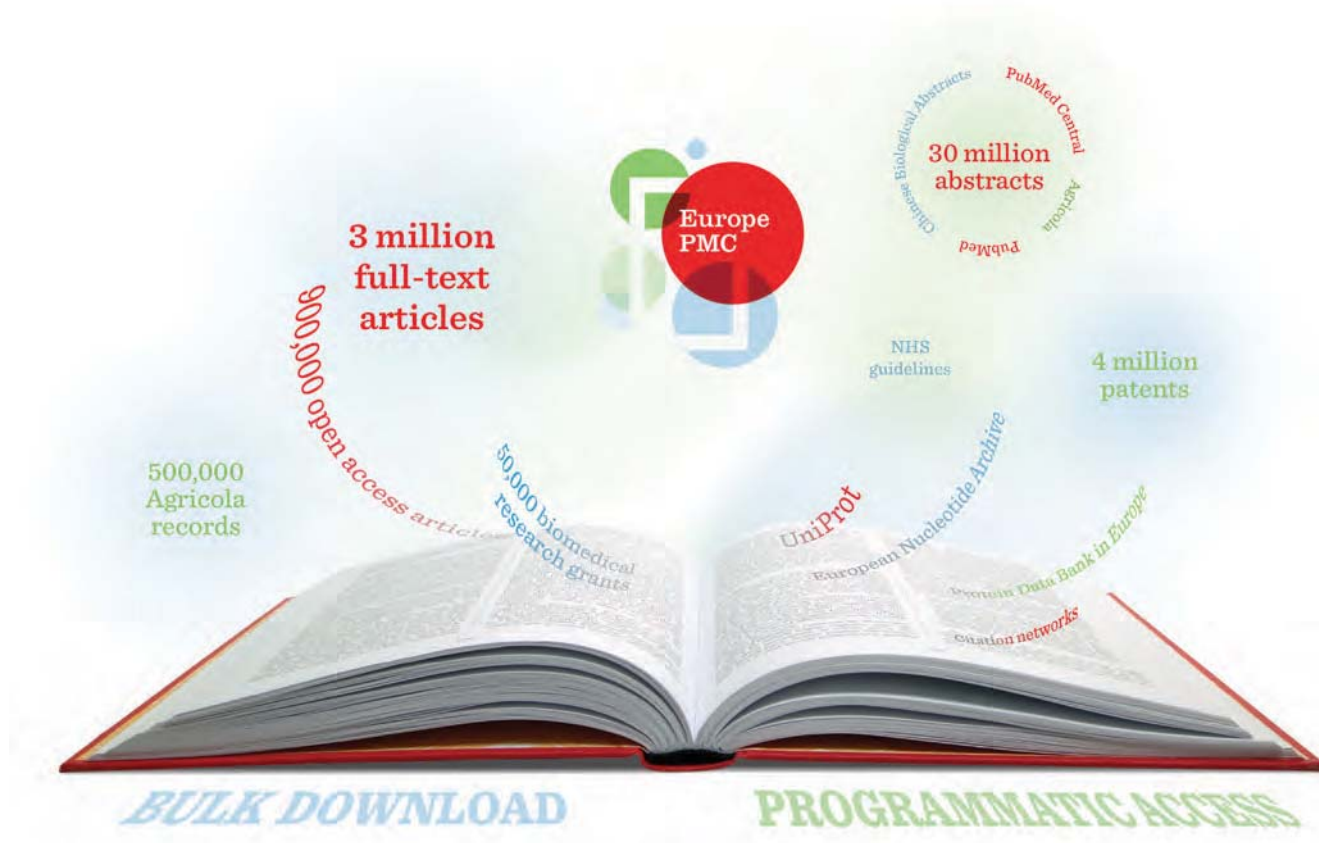
Our previous work on text mining data citations from full text articles and their accompanying supplemental files led to collaborations in the area of article-level metrics and alternative metrics ('alt-metrics'). We organised a workshop on the development of machine-readable data citations as described in the standard used by many publishers and Europe PMC to describe articles, NLM JATS DTD. The workshop was attended by several publishers and led to provision for data citation in the JATS DTD according to the Force11 Data Citation Principles.

Future plans

In 2015 we plan to re-establish the programme of work to develop a system for user accounts, which we reprioritised in order to deliver the unified search system. Users will be able to save their favourite searches, but perhaps more importantly we expect these accounts to form the basis of a social and bibliographic tools layer built on top of Europe PubMed Central content.

We also plan to develop functionality for delivering annotations generated from the Europe PMC text-mining pipeline directly into article displays. To achieve this, we will collaborate with leading text-mining groups to explore how we may also publish the results from third parties. This project will make use of the Embassy Cloud for groups that wish to run annotation software on new Europe PMC content on a regular basis. The key driver of these developments is the desire to integrate the literature with data usefully, feeding into EMBL-EBI databases both directly and via curation efforts.

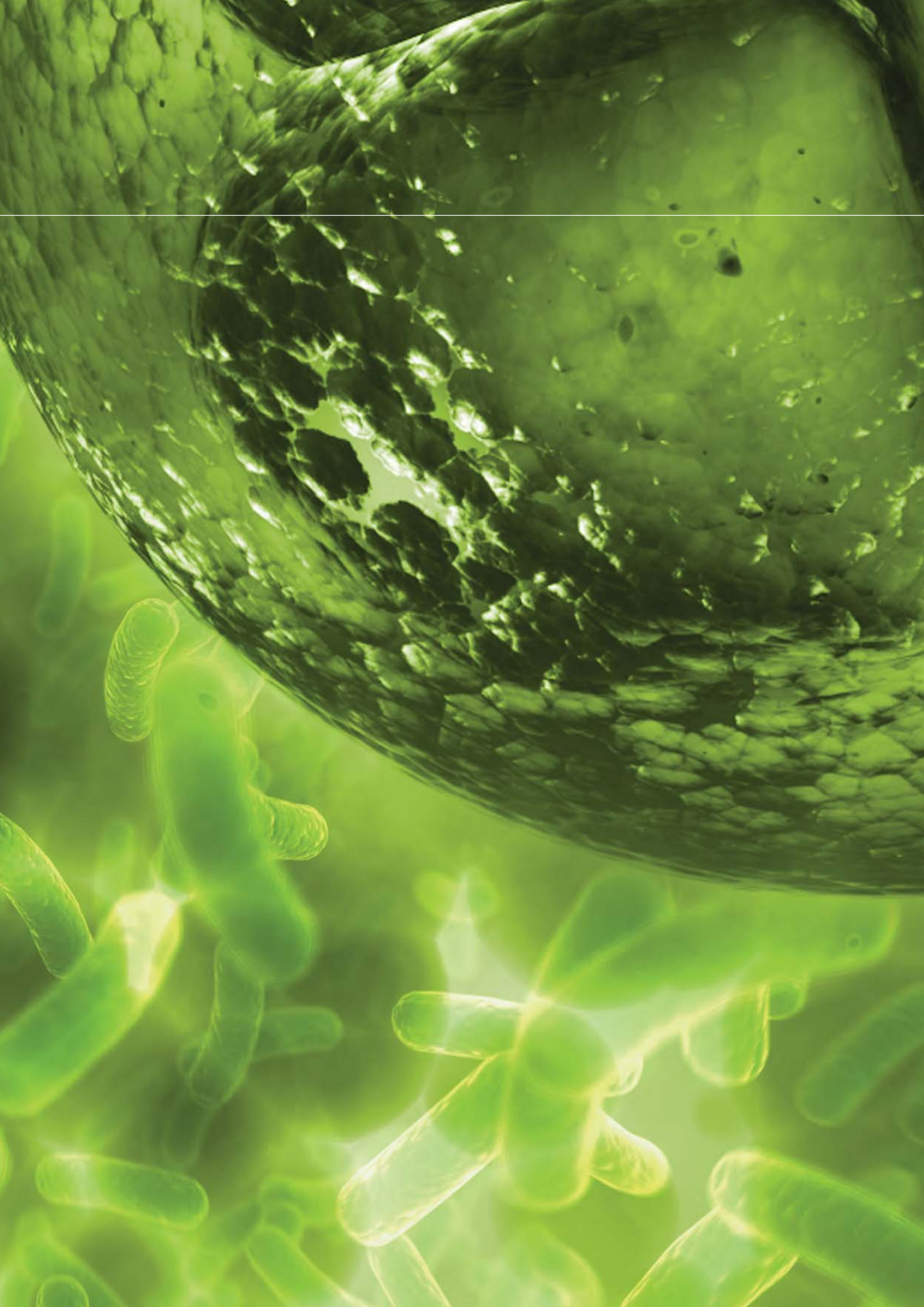
We expect these and related developments to drive the establishment of Europe PMC as a platform for text-based innovation over the coming years.



Europe PMC makes articles more discoverable and links them to the underlying data. In 2014, Europe PMC was the most widely used resource at EMBL-EBI, serving 10 million searches a month to 9 million unique IP addresses.



Research





Research highlights

Stem cell technology breakthrough

Paul Bertone and his colleagues resolved a long-standing challenge in biology by successfully ‘resetting’ human pluripotent stem cells to a fully pristine state, at the point of their greatest developmental potential. potential (Takashima et al., Cell).

Controlling protein shape

Sarah Teichmann and her collaborators uncovered a fundamental mechanism for regulating a protein’s shape (Perical et al., Science). Their findings have implications for the manipulation of proteins, with potential applications in biotechnology and drug development.

Evolutionary dynamics

Exploring the evolutionary coupling between transcription factor binding and gene expression in mammals, the Flicek team found that some transcription factor-dependent genes may be regulated mainly by a single transcription factor across evolution (Wong et al., Genome Research).

Kidney cancer in central Europe

As part of the International Cancer Genome Consortium, the Brazma team analysed data from large-scale DNA and RNA sequencing of renal cell carcinoma patients in Europe (Seelo et al., Nature Communications) and found a clear association between cancer incidence and exposure to aristolochic acid—an ingredient in some herbal remedies.

Analysing large drug screenings in cancer cell lines

The Saez-Rodriguez group collaborated on methods to analyse large drug screenings in cancer cell lines (Gobbi et al., Bioinformatics), offering significant improvements to efficiency in the study of large genomics datasets.



Warped Linear Mixed Models

The Stegle group was instrumental in creating WarpedLMM: a powerful new method that vastly improves the accuracy of Genome-Wide Association Study (GWAS) analyses (Fusi et al., Nature Communications). WarpedLMM helps researchers create a more accurate picture of the genetics of quantitative traits.

Introducing immunosteroids

The Teichmann group discovered that Type 2 T-helper cells produce a steroid that regulates their own proliferation (Mahata et al., Cell Reports). Their findings have implications for the study of cancers, autoimmune diseases and parasitic infections.

Where transcription meets translation

The Marioni group demonstrated that mRNA codon pools are highly stable over development and simply reflect the genomic background, while precise regulation of tRNA gene families is required to create the corresponding tRNA transcriptomes. Their findings reveal a stable molecular interaction interlocking transcription and translation.

Cell lineage trees

The Goldman group developed a method for inferring cell lineage trees from whole-genome single-cell sequencing data. The method offers better accuracy, correcting the vast majority of sequencing errors.

Comparing enzymes by function

The Thornton group created EC-BLAST: software that makes it possible to compare a potentially useful enzyme against thousands of well-known reactions (Rahman et al., Nature Methods). It creates a binary 'reaction profile' for enzymes that are relevant to green biotech, drug discovery and many other areas.



Research achievements 2014

Collaboration is the cornerstone of our research programme. Our approaches range from the purely mathematical to fully interdisciplinary, with two groups in 2014 running both wet and dry labs and all our researchers working with experimentalists from study design through analysis and publication.

The research landscape at EMBL-EBI has evolved from a niche area of science on the outskirts of molecular biology to a well-integrated discipline at the heart of biological research. Computational methodologies now allow us to explore a much wider variety of data and systems – on many more scales – than ever before. From individual molecules to complexes and whole organisms, bioinformatics connects different bodies of knowledge to bring meaningful questions to light.

Breakthrough in stem-cell technology

The Bertone group, in collaboration with colleagues in Germany, Japan and elsewhere in the UK, resolved a long-standing challenge in stem cell biology by successfully ‘resetting’ human pluripotent stem cells to a fully pristine state, at the point of their greatest developmental potential (Takashima et al., *Cell*). They used reprogramming methods to express the NANOG and KLF2 genes, which reset the cells, and maintained the cells indefinitely by inhibiting certain biological pathways. Working with the EMBL Genomics Core Facility, they produced comprehensive transcriptional data for all the conditions explored in order to compare reset human cells to genuine mouse ES cells. The human cells produced were genetically normal and capable of differentiating into any adult cell type. This marks a significant advance for human stem cell applications, such as drug screening of patient-specific cells and regenerative tissue grafts.

Controlling protein shape

The Teichmann group was instrumental in uncovering a fundamental mechanism for regulating a protein’s shape (Perica et al., *Science*). To determine how a protein morphs from an active (RNA-binding) configuration to an inactive one in two very different environments, they looked at a family of bacterial RNA-binding proteins that control a basic process in metabolism, one of which was an extremophile. They found that the process is controlled by mutations – not at the site where the binding happens, but at a distance, where they act indirectly to change the protein’s shape. Undertaken initially as a purely computational study, the work involved experiments in biophysics and structural biology as well as elastic network modelling. The findings have implications for the manipulation of proteins, with potential applications in biotechnology and drug development.

Single-cell genomics uncovers immunosteroids


Using data produced in the Single-Cell Genomics Centre, the Teichmann group discovered that Type 2 T-helper cells produce a steroid that regulates their own proliferation (Mahata et al., *Cell Reports*). Deep statistical analyses performed on a very large, comprehensive RNA-sequencing dataset led them to the genes involved in pregnenolone production at the point when Th2 cells are being produced. Their deduction that Th2 cells themselves were involved in immunosuppression has implications for the study of cancers, autoimmune diseases and parasitic infections.

Comparing enzymes by function

Bringing many disciplines to bear on a common problem in protein research, the Thornton group created software that makes it possible to quickly compare a potentially useful enzyme against thousands of well-known reactions (Rahman et al., *Nat Methods*). The new software, EC-BLAST, replaces a previously manual process. The computer scientists, chemists, biologist and physicist who developed EC-BLAST spent five years working through roadblocks to quantifying the comparison of enzymes already well described by the Enzyme Commission (EC). EC-BLAST provides an overview of the world of biochemical reactions through a series of algorithms that draw on a knowledge base of structures, chemical transformations, bond changes, stereochemistry and other enzyme features. It creates a binary ‘reaction profile’ for enzymes that will be of use to people working in green biotech, drug discovery and many other areas.

Better methods, better science

In collaboration with scientists at Microsoft Research, the Stegle group created a powerful new method that vastly improves the accuracy of Genome-Wide Association Study (GWAS) analyses (Fusi et al., *Nat Commun*). The new software, WarpedLMM, automates processes previously carried out manually by individuals, and makes it possible to attribute a greater proportion of phenotypic differences between individuals to genetics. WarpedLMM is a practical improvement on the widely used Linear Mixed Models, and can help create a more accurate picture of the genetics of quantitative traits.



Current technologies allow whole-transcriptome sequencing of spatially identified cells, but lack the throughput required to characterise complex tissues. In 2014 the Marioni group developed and validated a high-throughput method that facilitates identification of the precise spatial origin of cells assayed using scRNA-seq (Achim et al., accepted). Their approach compares complete, specificity-weighted mRNA fingerprints of a cell with positional gene-expression profiles derived from a gene-expression atlas, and can be applied to the study of any system with a sufficiently high-resolution reference gene-expression database.

The Goldman group developed a method for inferring cell lineage trees from whole-genome single-cell sequencing data. Their unique method explicitly models mutations and errors as two separate processes and offers better accuracy, correcting the vast majority of sequencing errors. The group also explored the question of how many cells need to be harvested to infer reliably the history of early embryonic cell divisions in mice and humans (Behjati et al., *Nature*). By simulating sub-sampled cell-lineage trees, they found they could estimate the number of cells needed to reconstruct cell divisions in the first few generations of cells with high probability.

Evolutionary dynamics

EMBO Postdoctoral Fellow Emily Wong in the Flicek team led work on the evolutionary coupling between transcription factor binding and gene expression in mammals (Wong et al., *Genome Res*). The results of this study highlight a significant tolerance to evolutionary changes in transcription-factor binding intensity in transcriptional networks in mammals, and suggest that some transcription factor-dependent genes may be regulated mainly by a single transcription factor across evolution.

In collaboration with the Odom lab at the University of Cambridge Cancer Research UK–Cambridge Institute (CRUK CI), the Marioni group published a study exploring the interface between transcription and translation, based on assays of mRNA and tRNA gene expression levels during mouse development (Schmitt et al., *Genome Res*). They demonstrated that mRNA codon pools are highly stable over development and simply reflect the genomic background, while precise regulation of tRNA gene families is required to create the corresponding tRNA transcriptomes. Their findings reveal a stable molecular interaction interlocking transcription and translation.

Chemogenomics

The ChEMBL database of bioactive entities provides a rich resource for pharmaceutical research, and is made stronger by the team's research activities. In 2014, a study led by PhD student Rita Santos in the ChEMBL team developed a general theoretical model of drug resistance and drug combinations that offers novel applications in improving drug safety and countering drug resistance in areas such as oncology and antibiotics (Rita Santos, PhD thesis).

Cancer

The Brazma research group, as part of the International Cancer Genome Consortium, analysed data from large-scale DNA and RNA sequencing of renal cell carcinoma patients in Europe and achieved insights into the genetic architecture of clear-cell renal-cell carcinoma (Scelo et al., *Nature Commun*). The consortium found a clear association between cancer incidence and exposure to aristolochic acid—an ingredient in some herbal remedies. These findings have implications for public health, particularly in Romania.

The Saez-Rodriguez group, in collaboration with groups at the Wellcome Trust Sanger Institute and the Netherlands Cancer Institute, developed methods to analyse large drug screenings in cancer cell lines (Gobbi et al., *Bioinformatics*). A major focus of their work was the comprehensive genomic characterisation of cancer cell lines, and an assessment of their value as models of actual tumours. Their BiRewire method offers significant improvements to efficiency in the study of large genomics datasets.

The Enright group, in collaboration with colleagues at the University of Cambridge and Addenbrooke's hospital, proposed a new, non-invasive method with the potential to diagnose childhood solid tumours (Murray et al., *Oncogenesis*). Building on previous work showing that microRNAs are useful diagnostic indicators for such tumours, they conducted a small test study from 53 serum samples and identified a panel of six miRNAs of potential importance for the clinical management of neuroblastoma and other childhood solid tumours.

Predocs to postdocs

In 2014 we benefited from the presence of 45 PhD students, welcoming eight newcomers. Fifteen students successfully defended their theses and were awarded PhDs from the University of Cambridge: Stephan Beisken, Samuel Croset, Mar Gonzales-Porta, Myrto Kostadima, Chen Li, Sergio Martinez-Cuesta, John May, Sarah Parks, Jean-Baptiste Pettit, Rita Santos, Christine Seeliger, Robert Sugar, Camille Terfve, Sander Timmer and Ying Yan.



Research summaries

Beltrao group

- Awarded a European Research Council (ERC) starting grant to study the functional relevance of protein phosphorylation using a combined computational and genetic approach;
- Contributed to a project led by the Thomas Benzing lab at Cologne University to study kidney-protein phosphorylation.

Birney group

- Along with Nick Goldman, published a high-profile paper on a method to store digital information in DNA.
- Publication of the first transcription factor QTL study, the CTCF transcription factor in human LCL lines
- We published the 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.

Brazma team

- Co-led a European renal cancer project in which our work revealed novel pathways and genes affected by recurrent mutations and abnormal transcriptome patterns including focal adhesion, components of extracellular matrix and genes encoding FAT cadherins (Scelo et al., Nature Commun).

Enright group

- Developed new techniques for the diagnosis of childhood tumours from microRNA profiling of serum;
- Published new work on the secretion of microRNAs via extra-cellular vesicles;
- Developed new computational techniques for the prediction and characterisation of long non-coding RNAs in mouse and *D. melanogaster*.

Flicek team

- Reported the first comprehensive results comparing the evolutionary changes between transcription factor binding and gene expression in mammals.

Goldman group

- Contributed to the initial publication of the ferret genome, focusing on the ferret as model organism to study human respiratory disease;
- Developed a maximum-penalized-likelihood method to estimate the distribution of selection coefficients;
- Developed methods to infer cell-lineage trees from whole-genome single-cell sequence data;
- Implemented novel multiple sequence alignment visualisation software;
- Studied worm behaviour under different environmental conditions using statistical models.

Marioni group

- Accounted for confounding variables in scRNA-seq data;
- Facilitated high-throughput spatial single-cell transcriptomics;
- Explored the interface between transcription and translation.

Overington team

- Studied the differential expression of all known, pharmacologically relevant and ADMET genes in mouse development, from pre-birth to natural old age, and analysed changes in the expression patterns of this gene set that could point towards differential efficacy and safety in paediatric and geriatric human populations;
- Developed the Functional Therapeutic Chemical Classification System resource, which specifically addresses drug repositioning;
- Validated two cases of target prediction for potential tuberculosis drug leads, including enzymology and structural biology studies;
- Developed a general theoretical model of drug resistance and drug combinations that offer novel applications in improving drug safety and countering drug resistance in areas such as oncology and antibiotics.

Saez-Rodriguez group

- Characterised cancer cell lines molecularly as models of primary tumours, and predicted drug efficacy in cancer from multiple types of data (genomic, epigenomic and transcriptomic);
- Developed a method to infer signalling activity from gene-expression data, and applied it to predict drug efficacy and toxicity;
- Developed a method to integrate signalling and metabolic processes, and applied it to understand deregulation of metabolism in cancer.

Stegle Group

- Developed new linear mixed model approaches to map transformed phenotypes (Fusi et al., 2014);
- Ran the first EMBO training course on genotype-phenotype mapping at EMBL-EBI in summer 2014;
- Developed new, powerful methods to dissect the transcriptional heterogeneity at the level of single cells.

Steinbeck team

- Developed a Java library for NMR signal processing;
- Completed a quantum mechanics pipeline for NMR spectra prediction;
- Released Metingear, a desktop application for creating and curating metabolic reconstructions;
- Released MassCascade, a workflow plug-in for processing metabolomics liquid chromatographic mass spectrometry data;
- Enhanced the desktop and command-line versions of the SENECA structure elucidator.

Teichmann group

- Discovered a steroid-producing T cell type involved in immune homeostasis (Mahata et al., 2014);
- In collaboration with the Clarke group, showed that mutations can act allosterically to change quaternary structure of protein complexes (Perica et al., 2014);
- Analysed the role of protein flexibility in the evolution and assembly of protein complexes (Marsh and Teichmann, 2014).

Thornton group

- Published our overview analysis of ligase enzymes, identifying complexities in the C-N forming ligases (EC 6.3), which include enzymes with very different reactions. The cofactor binding domains involved in ligation are used in many different ligase reactions, with different substrate binding domains. For the ligases, 'changing the substrate' is the dominant mode of evolutionary change.
- Published the EC-BLAST software tool, which compares enzyme reactions and produces a hit list of the most similar reactions;
- Completed an analysis of the chemistry and evolution of the isomerases;
- Completed a Genome Wide Association Study (GWAS) on longevity in flies, which has been submitted for publication;
- Completed an analysis of the complexity of enzyme reactions and the EC nomenclature.



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 that is 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

During 2014 the group was awarded an ERC Starting Grant to study the functional importance of protein phosphorylation. Previous research has shown that phosphorylation diverges quickly during evolution. One possible explanation for these fast evolutionary changes would be that a fraction of phosphosites serves no function in existing species. This funding will allow us to develop and apply genetic approaches to study the relevance to fitness of different phosphosites. This exciting new line of research will be highly complementary to the proteomic- and computational-based approaches that the group has been using.

The group was also awarded an EIPOD postdoctoral fellowship, for Haruna Imamura. Different pathogens are known to modulate a host's signalling pathways in order to promote infection or inhibit defence mechanisms. Haruna's project will explore how human proteins change in their PTMs during *Salmonella* infection. This project will be developed in collaboration with the groups of Nassos Typas and Jeroen Krijgsveld at the EMBL Genome biology unit.

Protein function can be regulated by PTMs by changing different properties such as their localisation, interactions or activity. In previous years we showed that conservation of PTMs within domain families could identify regions of these domains that were important for the regulation of their function. In 2014 we participated in a study of kidney protein phosphorylation. Our analysis contributed for the identification of important regulatory phosphosites in the prohibitin homology domain of the podocin protein.

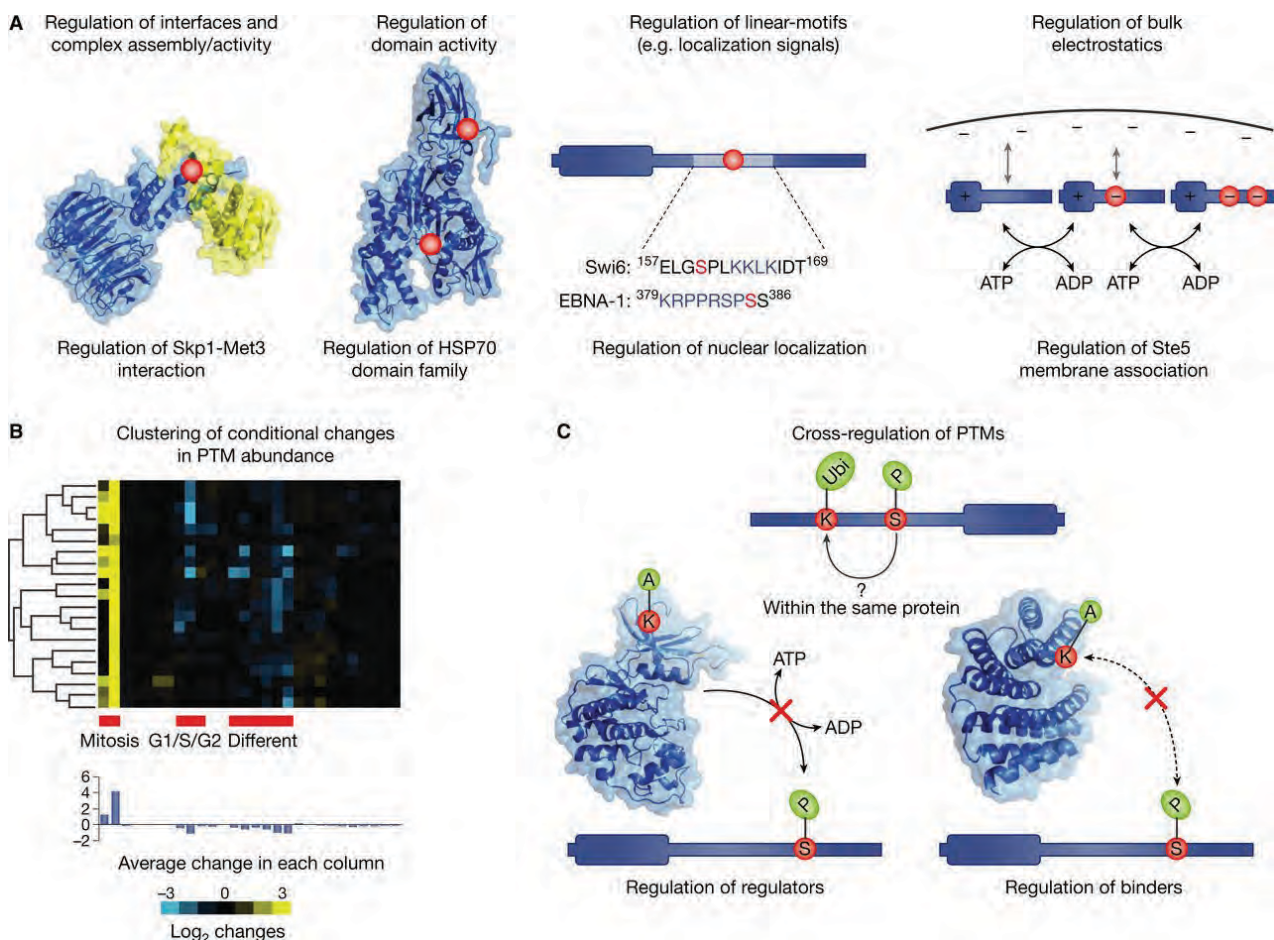


Future plans

In 2015 we will continue to study the evolution of cellular interaction networks with a specific focus on post-translational regulatory networks, genetic and chemical-genetic networks. Our PTM-related work has focused closely on the study of conditional dependent regulation as well as phylogenetic based approaches to determine the age of a PTM. The genetic studies have primarily examined the condition dependence of genetic-interaction networks.

Selected publications

Rinschen MM, Wu X, König T, et al. (2014) Phosphoproteomic analysis reveals regulatory mechanisms at the kidney filtration barrier. *J Am Soc Nephrol* 25:1509-1522



Functional role of post-translational modifications. PTMs act to change the activity of proteins through different mechanism and in response to different conditions. (A) Different mechanism used by PTMs to regulate protein activity. (B) Example of conditional regulation of phosphorylation sites. (C) Mechanism of cross-regulation between different PTM types. [Beltrao et al., *Mol Sys Biol* 2013]



Birney group

Nucleotide Data

DNA sequence remains at the heart of molecular biology and bioinformatics. Dr Birney is joint Associate Director of EMBL-EBI and shares strategic oversight of bioinformatics services. He also has a modestly sized research group, focused on 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 four years a considerable focus has been on compression, with theoretical and now 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—including the application of new sequencing technologies and the interaction with imaging techniques.

We are also interested in the interplay of natural DNA sequence variation with cellular assays and basic biology. Over the past five years there has been a tremendous increase in the use of genome-wide association to study human diseases. However, this approach is very general and need not be restricted to the human disease arena. 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 both molecular (e.g. RNA expression levels and chromatin levels) and basic biology traits in a number of species where favourable populations are available including human and *Drosophila*. In humans we are exploring both molecular traits and physiological traits, such as 4D resolution of human hearts in healthy volunteers. We hope to expand this to a variety of other basic biological phenotypes in other species, including establishing the first vertebrate near-isogenic wild panel in Japanese Rice Paddy fish (Medaka, *Oryzias latipes*).

Major achievements

Our group carries out two major types of projects, one of which is a series of molecular and other phenotype association studies. We are working in a variety of human systems to explore association of molecular events in both normal and disease samples. The former is focused on studies of the human heart structure and physiology in collaboration with MRI researchers and Cardiologists at Imperial College. The latter (disease samples) is a collaboration with Francis Collins at the National Institutes of Health in the US through a shared student. For both studies we are working closely with the Stegle research group at EMBL-EBI, using their new association methods that can handle both population confounders and other multiple phenotype scenarios.

We continue to develop the resources around Medaka fish, and have demonstrated that our selected population looks appropriate for establishing a population reference panel.



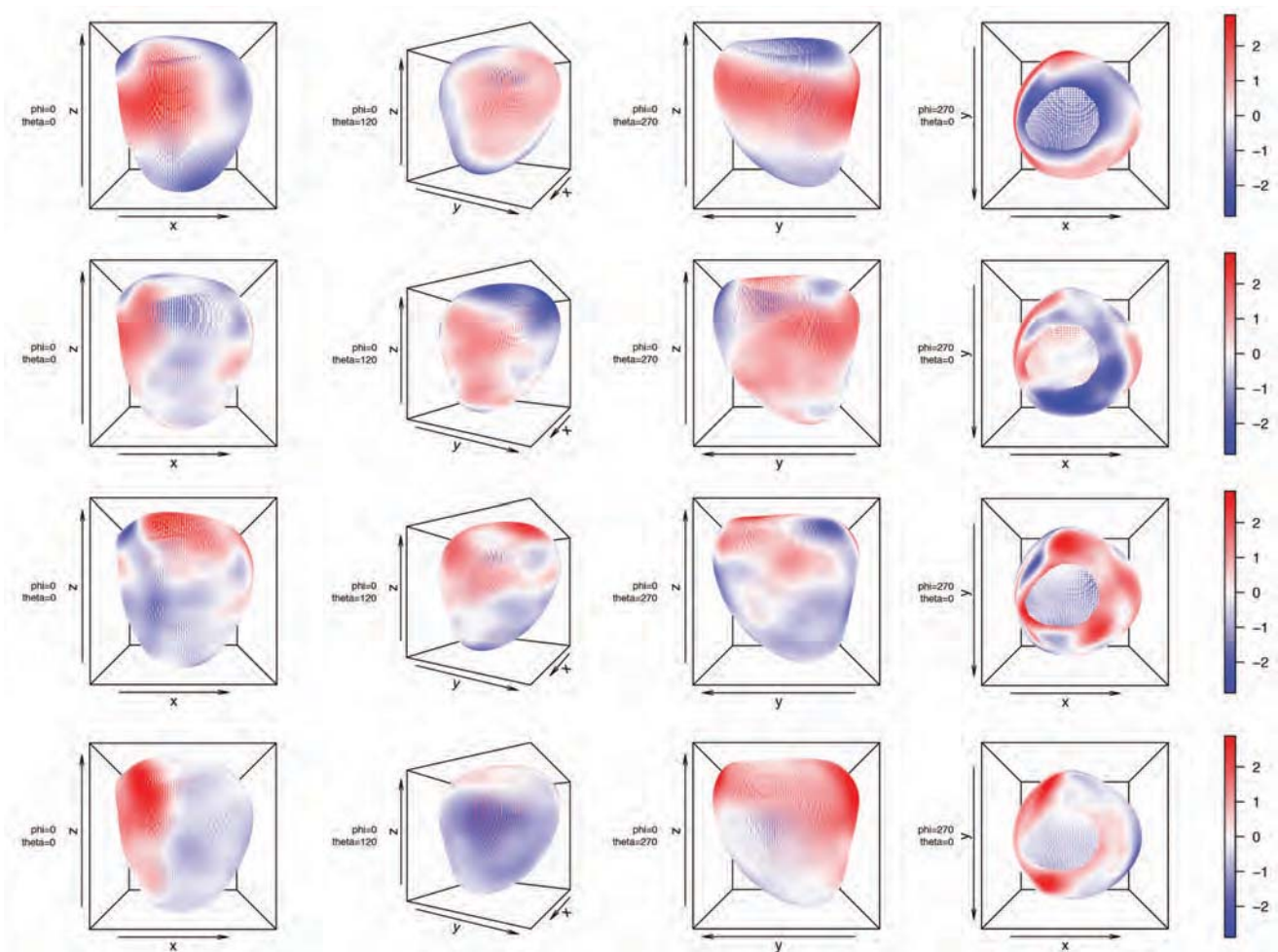


Other projects in our group explore broader associations of molecular functional information, such as information generated by the ENCODE and Epigenome Roadmap projects with cancer data (the BASIS project on Breast Cancers) and working on leveraging human genetics to help impact drug discovery via a project with the Centre for Therapeutic Target Validation.

Dr Birney also engages in policy-level discussions on the use of genomic information in human health, and recently co-authored a paper on the risks and benefits of incidental findings in human genomics.

Future projects and goals

In 2015 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.



Visualisations of low dimensional projections of sources of variance in healthy human hearts on idealised 3D model of the human heart.



Enright group

Functional Genomics and Analysis of Small RNA Function

Complete genome sequencing projects are generating enormous amounts of data, and while progress has been rapid a significant proportion of genes in any given genome are either un-annotated or possess a poorly characterised function.

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

Clinical genomics

In our long-standing collaboration with the Department of Pathology at the University of Cambridge and Addenbrooke's Hospital, we explored the roles of microRNAs and other non-coding RNAs in paediatric germ cell tumours and solid tumours such as neuroblastoma. We showed previously that microRNAs are useful diagnostic indicators for such tumours from analysis of miRNA levels in tumour biopsy samples; however, performing similar analyses from serum derived from blood is far less invasive. In a small test study from 53 serum samples, we identified a panel of six miRNAs of potential importance for the clinical management of neuroblastoma and other childhood solid tumours.

Long non-coding RNAs

Working closely with the Furlong group at EMBL Heidelberg and the O'Carroll at EMBL Monterotondo, 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.

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. We also worked with Frédéric Quignon at the University of Lorraine to explore miRNAs expressed in the developing larva of the Honey bee (*Apis mellifera*) and possible changes resulting from the effects of pesticide contamination.

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 new tool for the assessment of 3' Uridylation of miRNAs and other changes such as adenylation or ADAR editing. Thanks to a BBSRC-funded project, we also developed a new graphical user interface for our miRNA pipelines that will enable simple access to cutting-edge next-generation sequencing analysis via a graphical interface. We plan to release these new tools in 2015.

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.



Future plans

In 2015 we will expand our analysis of tumour-secreted miRNAs, and work very closely with the Furlong and O'Carroll groups to explore the functional roles of lncRNAs. We will embark on a project to explore the process of genomic integration of Human papilloma virus (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 tools, Chimera and the miRNA pipeline user interface, over the course of the year.

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. 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.

Selected publications

Cossetti C, Iraci N, Mercer TR, et al. (2014) Extracellular vesicles from neural stem cells transfer IFN- γ via Ifngr1 to activate Stat1 signaling in target cells. *Mol Cell* 56:193-204

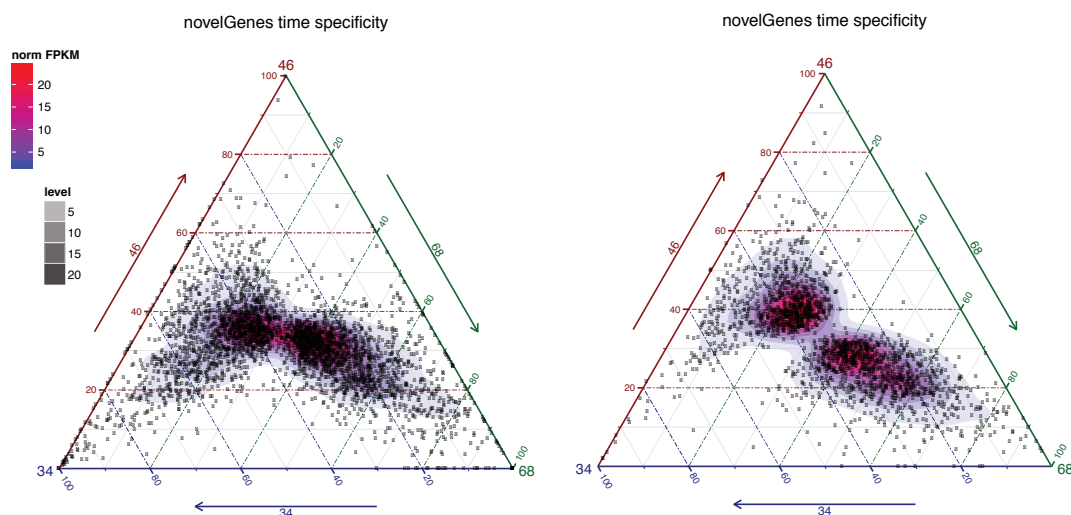
Camps C, Saini HK, Mole DR, et al. (2014) Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *J Mol Cancer* 13:28

Murray MJ, Bailey S, Raby KL, et al. (2014) Serum levels of mature microRNAs in DICER1-mutated pleuropulmonary blastoma. *Oncogenesis* 3:e87

Bussotti G, Notredame C, Enright AJ (2013) Detecting and comparing non-coding RNAs in the high-throughput era. *Int J Mol Sci* 14:15423-15458

Novel *D. Melanogaster* lncRNAs

Two sample time-course ternary plots of expression (with Furlong lab, EMBL Heidelberg)



Expression of novel lncRNAs predicted in *D. melanogaster*. Each ternary plot shows expression levels of de novo predicted lncRNAs across three time points from both unsorted cells (left) and sorted cells (right).



Goldman group

Evolutionary Tools for Genomic Analysis

Evolution is the historical cause of the diversity of all life. Our group's research focuses on the development of data analysis methods for the study of molecular sequence evolution and for the exploitation of evolutionary information to draw powerful and robust inferences about phylogenetic history, molecular evolutionary processes and genomic function.

The evolutionary relationships between all organisms require that we analyse molecular sequences with consideration of the underlying structure relating those sequences.

We develop mathematical, statistical and computational techniques to reveal information present in genome data, to draw inferences about the processes that gave rise to these interrelationships and to make predictions about the biology of the systems whose components are encoded in those genomes.

Our three main research activities are: developing new evolutionary models and methods, providing these methods to other scientists via stand-alone software and web services, and applying such techniques to tackle biological questions of interest. We participate in comparative genomic studies, both independently and in collaboration with others; this typically involves the analysis of next-generation sequencing (NGS) data. This vast source of new data is providing enormous gains in understanding genomes, and brings with it many new challenges.

Major achievements

Our group contributed to the publication of the ferret genome. The ferret is an important model organism for respiratory disease research. We used our state-of-the-art orthology-calling software to identify homologous genes between the ferret and a set of 33 other organisms for which the complete genome sequence is known. We examined the evolutionary distances from human to ferret and mouse for each ortholog present in the three species. We found that, in general, there is less divergence between human and ferret genes linked to respiratory disease.

We investigated the structural and functional determinants of selective constraint in mammals on the level of genes and domains. We contributed evolutionary analysis to two studies: conservation analysis of a transcription factor implicated in megakaryopoiesis in a study by the BLUEPRINT consortium, and analysis of the evolutionary dynamics of genome regions that are differentially methylated between different tissues.

Estimation of the distribution of selection coefficients of mutations ($S=2Ns$) presents important challenges in molecular evolution. We developed a maximum penalized-likelihood method to estimate the distribution in protein-coding genes from phylogenetic data. We found

that the new method regularises the estimates of amino acid fitnesses for small, relatively uninformative datasets, and does a better job of recovering the large proportion of deleterious mutations. We also found that as the number of taxa in the phylogeny or the level of sequence divergence increases, the distribution of S can be more accurately estimated. Our analysis of three widely studied protein-coding genes for which large datasets are available showed that the new method recovers a large proportion of deleterious mutations in these data, even under strong penalties, confirming the distribution of S is bimodal in these real datasets.

As most cell divisions introduce a small number of mutations in the genome, it is now possible, in principle, to infer phylogenies of single cells from sequencing data. Known as cell lineage trees, these phylogenies are becoming useful in developmental biology and cancer research. We developed a method for inferring cell lineage trees from whole-genome single-cell sequencing data. This problem differs from standard phylogenetic reconstruction problems due to very low mutation rates and large numbers of sequencing errors. Unlike other approaches, our method explicitly models mutations and errors as two separate processes. In simulations, we found that our method is more accurate than current approaches, and corrects the vast majority of sequencing errors. In 2014 we began applying this method to real data, in collaboration with the Stratton and Macaulay groups at the Wellcome Trust Sanger Institute.

We investigated the experimental design problem of how many cells need be harvested to infer the history of early embryonic cell divisions in mice and humans reliably. Real-life studies will only ever be able to sample and sequence a tiny fraction of the trillions of cells in a fully developed mammal. By simulating sub-sampled cell-lineage trees, we can estimate the number of cells needed to reconstruct cell divisions in the first few generations of cells with high probability. The results we obtained in 2014 guided the design of a large-scale single-cell study that is being carried out in the Stratton group at the Sanger Institute.

We continued our efforts to provide interactive visualisation of large sequence datasets. In collaboration with the London-based design studio Science Practice, we implemented Sequence Bundles, our previously published visualisation method, into a feature-rich standalone application for explorative analysis and visualisation of

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.



Alvis Sequence Bundle visualisation of the HAD family. The Bundle shows the different sequence groups in different colours. Horizontal dependencies are immediately visible. All Ciona sequences (green) have Met in position 241 and have a Tyr residue in position 232 exclusively (not shown) and a Val at position 249. This information is not available from the standard sequence logo. Top: green shaded markers indicate which sites are most likely responsible for the grouping. Site 241 (orange triangle) is most significant (see Seifried et al., 2014).

multiple sequence alignments. This application, Alvis, allows for simultaneous visualisation of phylogenetic trees and alignments and can identify and display phylogenetic information such as homoplasies and synapomorphies in a concise manner.

Our previous experience with modelling horizontal dependencies between adjacent genomic sites as a way to study the development of cancer tumours, as used in the MEDICC software, led to an interdisciplinary collaboration with Dr André Brown at Imperial College, London. Using a similar finite-state transducer framework, we provided statistical modelling of behavioural phenotypes in *C. elegans*. Our results illustrated the complexity of worm behaviour under different environmental conditions and provided a systematic way of extracting behavioural motifs characterising genetic and environmental variants.

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 concentrate on linked areas that draw on our expertise in phylogenetics, genomics and NGS. Basic to all our work are the fundamentals of phylogenetic analysis, an area in which we will continue to devise improved tests for detecting positive selection, investigating the use of non-reversible models of sequence substitutions and developing data analysis methods to detect and represent the discordant evolutionary histories of different genes in

an organism's genome. We are committed to keeping abreast of evolving NGS technologies and exploiting them for new experiments. Particularly intriguing is the new possibility of sequencing single cells, opening the way to studies that can trace the development of the different parts of a single living organism. We will continue to look to medical applications of NGS and phylogenetics as a source of inspiring collaborations, and have new collaborations underway bringing molecular evolutionary methods and whole-genome NGS of pathogens into a clinical setting where they may be applicable in 'near real time' to help inform doctors' decisions and treatment choices.

Selected publications

Behjati, S. et al. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 513:422–425

Parks, S. L., and N. Goldman. 2014. Maximum likelihood inference of small trees in the presence of long branches. *Systematic Biology* 63:798–811

Schwarz, R. F. et al. 2014. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comp Biol* 10:e1003535

Tamuri, A. U., N. Goldman, and M. dos Reis. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–271



Marioni group

Computational and Evolutionary Genomics

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.

Within this context, we focus on work in 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 2014 we built on our previous work in the field of single-cell transcriptomics, collaborating with the Stegle and Teichmann groups to develop a method that accounts for confounding factors such as the cell cycle on the heterogeneity of gene expression (Buettner et al., 2015). The approach uses latent variable models to account for such hidden factors, allowing us to identify otherwise undetectable subpopulations of cells that correspond to different stages during the differentiation of naïve T cells into T-helper-2 cells. This computational strategy can be used to identify cellular subpopulations and to tease apart different sources of gene expression heterogeneity in single-cell transcriptomes.

To achieve a full understanding of cell type identity in a multicellular tissue, each cell's expression profile must be integrated with its spatial location within the tissue under study. Although current technologies allow whole-transcriptome sequencing of spatially identified cells, they lack the throughput required to characterise complex tissues. In 2014 we developed and validated a high-throughput method that facilitates identification of the precise spatial origin of cells assayed using scRNA-seq within the tissue of interest (Achim, Pettit et al., accepted). 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 can apply our approach to the study of any system with a sufficiently high-resolution reference gene-expression database.

In 2014 we continued our work developing novel approaches for studying gene-expression levels using bulk RNA-sequencing data. In collaboration with the Odom lab at the University of Cambridge Cancer Research UK–Cambridge Institute (CRUK–CI), we published an important study exploring the interface between transcription and translation, based on assays of mRNA and tRNA gene expression levels during mouse development (Schmitt, Rudolph et al., 2014).

The Marioni group was strengthened by the appointment of John Marioni as an Associate Faculty Member of the Wellcome Trust Sanger Institute, in recognition of his contribution to the development of single-cell 'omics approaches on the Wellcome Genome Campus. 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.

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.



Future projects and goals

Our group will continue to focus on developing computational tools for understanding the regulation of gene-expression levels. We will focus on developing methods for analysing single-cell RNA-sequencing data, which has the potential to reveal novel insights into cell-type identity and tumourigenesis.

We will extend the model introduced by Buettner and colleagues to better order cells along a differentiation trajectory (i.e. ordering cells in 'pseudotime') and to further tease apart the contributions of different factors to heterogeneity in gene expression across cells. We will also continue to investigate how Bayesian approaches can be used to better identify highly variable genes across cell types, and to develop robust computational approaches for assaying the degree of stochastic, allele-specific expression across single-cell populations.

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., accepted), using these data to identify cell types and examine heterogeneity in expression at the spatial level (Pettit et al., 2014).

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 – as part of our work on a Wellcome Trust Strategic Award-funded project led by Wolf Reik at the Babraham Institute. Moreover, we will continue to apply our models in numerous biological contexts, such as the study of heterogeneity in mouse embryonic stem cells, cancer biology and non-model systems to study evolution.

Selected publications

Achim K, Pettit JB, Saraiva LR, et al. (2015) Single-cell expression profiling and spatial mapping into tissue of origin. *Nat Biotechnol* (accepted)

Buettner F, Natarajan KN, Casale FP, et al. (2015) Computational analysis of cell-to-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* (in press); DOI: .10.1038/nbt.3102

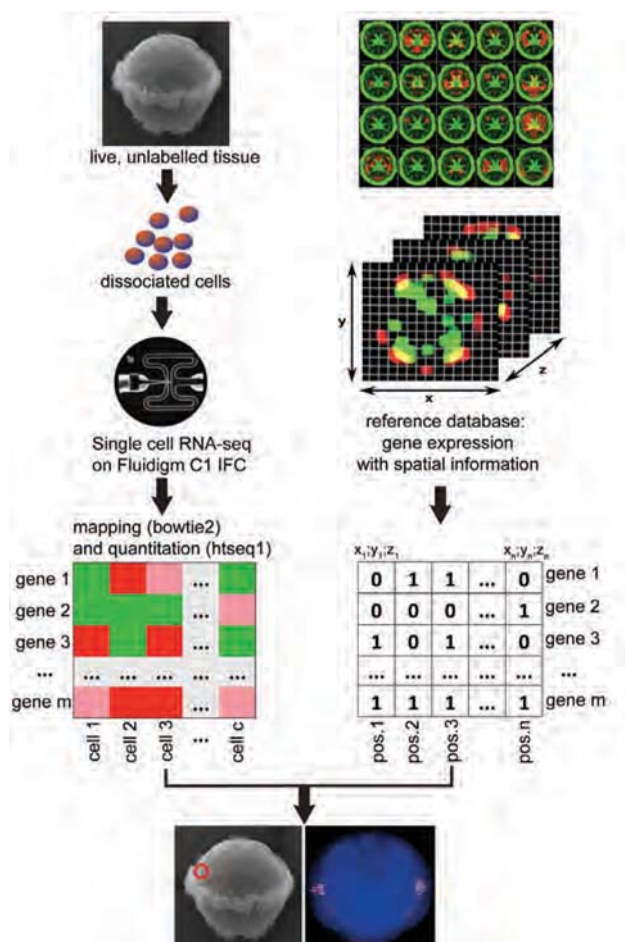
Pettit JB, Tomer R, Achim K, et al. (2014) Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput Biol* 10:1003824

Schmitt BM, Rudolph KL, Karagianni P, et al. (2014) High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Res* 24:1797-1807

Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* (in press); DOI: 10.1038/nrg3833

An approach for integrating single-cell RNA-sequencing with spatial information. We used an existing gene expression atlas (right) to link single-cell RNA-seq data (left) from cells extracted from the developing brain of *P. dumerilii* (top left) with the spatial coordinates. The gene expression atlas was binarized, resulting in a matrix of n positions that each comprise of presence/absence values for m genes.

For each sequenced cell c , expression data for m genes was compared to the n positions in the reference matrix and matched with the respective position(s) based on highest similarity. An example of the likely position for one cell is indicated in the bottom panel: by the red circle in the bottom left (ventral view of *P. dumerilii* larva) and red voxels in the apical view at bottom right.





Saez-Rodriguez group

Systems
Biomedicine

Our group aims to achieve a functional understanding of signalling networks and their deregulation in disease and seeks to apply this knowledge to 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 2014 we developed 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 and the Wessels group at the Netherlands Cancer Institute (NKI). A major focus was the comprehensive genomic characterisation of cancer cell lines, and an assessment of their value as models of actual tumours. This was done by comparing the genomic features of the cell lines to those of primary tumours.

We further developed our CellNOpt tool for logic modelling and applied it to various questions, from studying crosstalk mechanisms in yeast (with the Aebersold group at ETH-Zurich) to analysing the differences in signalling circuits in multiple sclerosis patients (as part of the European consortium CombiMS).

In collaboration with the Schultz group at EMBL-Heidelberg, we analysed single-cell live-imaging data and used it to build mechanistic models of PI3K signalling to understand the regulation of PIP3.

Our group developed a novel methodology to build logic signalling networks from phosphoproteomic data generated from mass spectrometry shotgun data, 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 School of Medicine.

We worked on the integration of models of signal transduction, gene regulation and metabolism. These processes within the cell are highly interconnected, and only an integrated view can explain how they are orchestrated and to elucidate their deregulation in human disease.

In addition to our basic research activities, we again helped organise DREAM Challenges (Dialogues in Reverse Engineering Assessment of Methods; www.dreamchallenges.org). DREAM is a community effort that uses 'challenge' events to advance the inference of mathematical models of cellular networks.



Future projects and goals

In 2015 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 on modelling signalling networks using phosphoproteomics 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 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

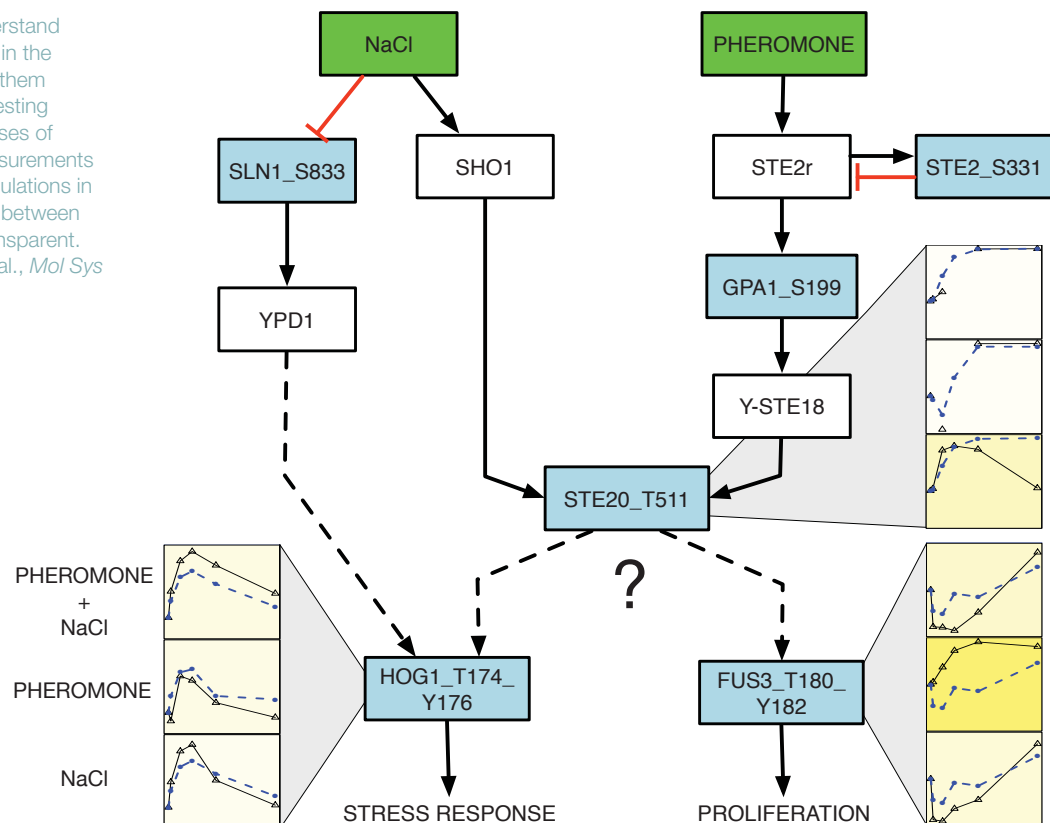
Vaga S, Bernardo-Faura M, Cokelaer T, et al. (2014) Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast.. *Mol Sys Biol* 10:767

Gobbi A, Iorio F, Dawson KJ, et al. (2014) Fast randomization of large genomic datasets while preserving alteration counts. *Bioinformatics* 30:i617-i623

Costello JC, Heiser LM, Georgii E, et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 32:1202-1212

Egea JA, Henriques D, Cokelaer T, et al. (2014) MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinf* 15:136

Logic modeling to understand mechanisms proposed in the literature by converting them into logic models and testing them against time-courses of phosphopeptides. Measurements are shown in black, simulations in blue and disagreement between data and simulation transparent. [Adapted from Vaga et al., *Mol Sys Biol* 2014]





Stegle group

Statistical Genomics & Systems Genetics

Our interest lies in computational approaches to unravel the genotype–phenotype map on a genome-wide scale. How do genetic background and environment jointly shape phenotypic traits or causes diseases? How are genetic and external factors integrated at different molecular layers, and how variable are these molecular readouts between individual cells?

We use statistics as our main tool to answer these questions. To make accurate inferences from high-dimensional ‘omics datasets, it is essential to account for biological and technical noise and to propagate evidence strength between different steps in the analysis. To address these needs, we develop statistical analysis methods in the areas of gene regulation, genome wide association studies (GWAS) and causal reasoning in molecular systems.

Our methodological work ties in with experimental collaborations, and we are actively developing 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 and the proteome, and we derive new tools for single-cell biology.

Major achievements

In 2014 we developed and applied methods for linking genetic variation data and phenotype. In collaboration with researchers at Microsoft Research, we devised new methods to model phenotype data on an unknown linear scale. By combining principles from genetic association studies with non-linear regression models, we improved the genetic analysis of quantitative traits, thereby revealing more clearly how genetic differences shape phenotypic diversity (Fusi et al., 2014).

In addition to developing new methods for genetic association analysis, we started work in the area of single-cell genomics. Together with colleagues at the Babraham Institute, we showed how single-cell epigenomes can be profiled using a combination of experimental advances and statistics. (Smallwood et al., 2014). In parallel, we collaborated with the Marioni and Teichmann groups at EMBL-EBI to devise new ways to dissect transcriptional heterogeneity between single cells (Buettner et al., 2015). Our approach, for the first time, enables modeling both known and unknown factors that underlie single-cell transcriptome variation. This method has already helped to identify new sub-clusters of cells in single-cell RNAseq studies and will be an important building block for our future endeavours.



Oliver Stegle

PhD in Physics, University of Cambridge, 2009.
Postdoctoral Fellow, Max Planck Institutes
Tübingen, 2009–2012.

Research Group Leader at EMBL-EBI since
November 2012



Future plans

In 2015 we will continue to develop innovative statistical approaches to analyze data from high-throughput genetic and molecular profiling studies. We are particularly interested in following up our recent efforts to model single-cell variation data. A major challenge in this area will be the integration of multiple modalities in single-cell genomics, for example linking single-cell epigenome variation with single-cell RNAseq. We are particularly interested in applying these methods to data from the Human Induced Pluripotent Stem Cell Initiative (HiPSci), in which we are a partner.

Selected publications

Fusi N, Lippert C, Lawrence ND, Stegle O (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat Commun* 5:4890

Smallwood SA, Lee HJ, Angermueller C, et al. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11:817-820

Cubillos FA, Stegle O, Grondin C, et al. (2014) Extensive cis-regulatory variation Robust to environmental perturbation in Arabidopsis. *Plant Cell* 26: 4298-4310

Buettner, F. et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotech* (in press); DOI: 10.1038/nbt.3102

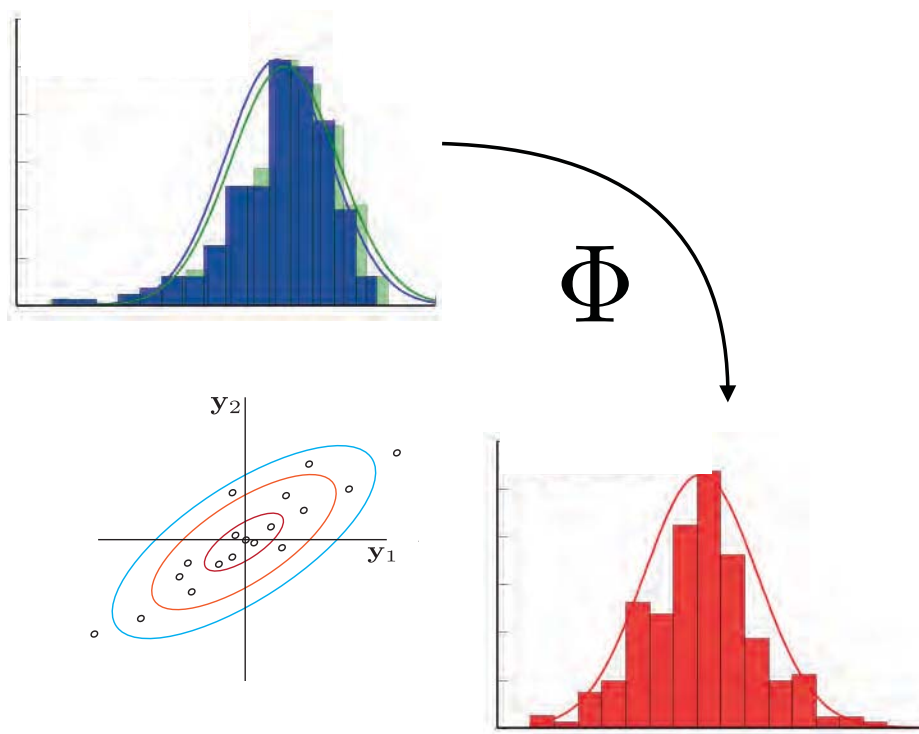


Illustration of the WarpedLMM approach for improved genetic mapping of quantitative traits. This integrative statistical model jointly discovers a non-linear parameterization of the phenotype data while testing for genetic associations. This joint modeling approach helps to greatly increase power and interpretation of genome-wide association studies.



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.

We also explore the regulation of gene expression during switches in cell state, and use mouse T-helper cells as a model of cell differentiation. We combine computational and wet-lab approaches at both EMBL-EBI and the Wellcome Trust Sanger Institute.

The wealth of genome-scale data available for sequences, structures and interactions provides an unprecedented opportunity to investigate systematically principles of gene and protein interactions. 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. Our two main areas are transcription factors and the regulation of gene expression; and physical protein-protein interactions and protein complexes.

Differences in genes and their spatio-temporal expression patterns determine the physiology of an organism: its development, differentiation and behaviour. Transcription factors regulate this process by decoding DNA elements and binding to DNA in a sequence-specific manner. Our group has developed a prediction pipeline (transcriptionfactor.org) that identifies repertoires of transcription factors in genomes.

We are very interested in elucidating transcriptional regulatory networks that orchestrate T-helper-cell differentiation and plasticity. Using the T-helper-cell system, we explore the hierarchy and kinetics of molecular events that contribute to changes in gene expression, and whether the kinetics of these interactions is graded or switch-like. In 2014 we discovered a steroid-producing T-cell subtype that can inhibit T-cell proliferation and B-cell class switch, and proposed that de novo steroid biosynthesis in immune cells has a role in immune homeostasis (i.e. switching off an immune response after an infection challenge).

Our group investigates principles that govern the folding and assembly of protein complexes. Using the informative power of genomic, proteomic and structural data, we capture the critical changes in sequence and structure that distinguish protein-complex formation from the sea of functionally neutral changes. The 3DComplex.org database is a research tool for our work in this area. Our in silico, phylogeny-based methods predict critical ancestral mutations involved in changing protein complexes, and we test these using wet-lab biophysical and biochemical techniques.

Major achievements

In 2014, our group used single cell RNA-sequencing to dissect a new subpopulation of T cells that makes its own steroids (Mahata et al., 2014). While steroids have long been used to suppress the immune system, it was not thought that immune cells themselves are steroidogenic; rather, steroid biosynthesis was thought to be restricted to the adrenal gland and gonads. This signalling system is likely to be a way that immune cells switch off an immune response at the end of an infection, and could potentially be used by tumours as a mechanism of immune evasion.

In our work on structural bioinformatics of protein complexes, our major breakthrough in 2014 was to show a mechanism for 'allosteric' mutations to modify protein quaternary structure (Perica et al., 2013). These mutations act from a distance, and in the case of the pyrR protein family, they act by rewiring amino acid contacts to stabilise one pre-existing conformation over another. The same population shift mechanism used by allosteric ligands in this protein family is 'hijacked' in protein evolution, to effect a switch between dimeric and tetrameric states (see Figure).

Related to dynamics in protein complexes, we charted global trends of protein flexibility in protein subunits in different types of complexes (Marsh and Teichmann, 2014). We show that intrinsic protein flexibility is greater when proteins form asymmetric (i.e. heterologous) inter-subunit interfaces. This leads to a strong association between subunit flexibility and homomeric complexes with cyclic and asymmetric quaternary structure topologies (as opposed to dihedral symmetries), and to greater flexibility in complexes with more non-homologous subunits. We extended these principles to protein evolution, showing that eukaryotic proteins tend to be more flexible than prokaryotic proteins, and associate with more distinct binding partners.

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 Career Track Programme Leader, MRC Laboratory of Molecular Biology, 2001-5 and MRC Programme Leader, 2006-12. Fellow and Director of Studies, Trinity College, since 2005. Principle Research Associate at the Department of Physics/Cavendish Laboratory, University of Cambridge, 2013-2016.

Group Leader at EMBL-EBI and Sanger Institute since 2013.



Future plans

We will continue our projects in structural bioinformatics of protein complex assembly and expand our programme in genomics of gene expression. We will focus on single-cell transcriptomics of the dynamics of immune response to pathogens. This will reveal the full spectrum of CD4+ T-cell types, and lead to new discoveries. Further, we will explore the evolution of an infection time course in terms of expansion and contraction of diverse cell types, and their cell-cell interactions. These in vivo experiments, together with in vitro T-cell and ES-cell experiments, will inform us about the cellular circuitry and decision-making in switching from one cell type to another. To gain more insight into cellular switches, we will integrate high-throughput high-content imaging with single-cell RNA sequencing.

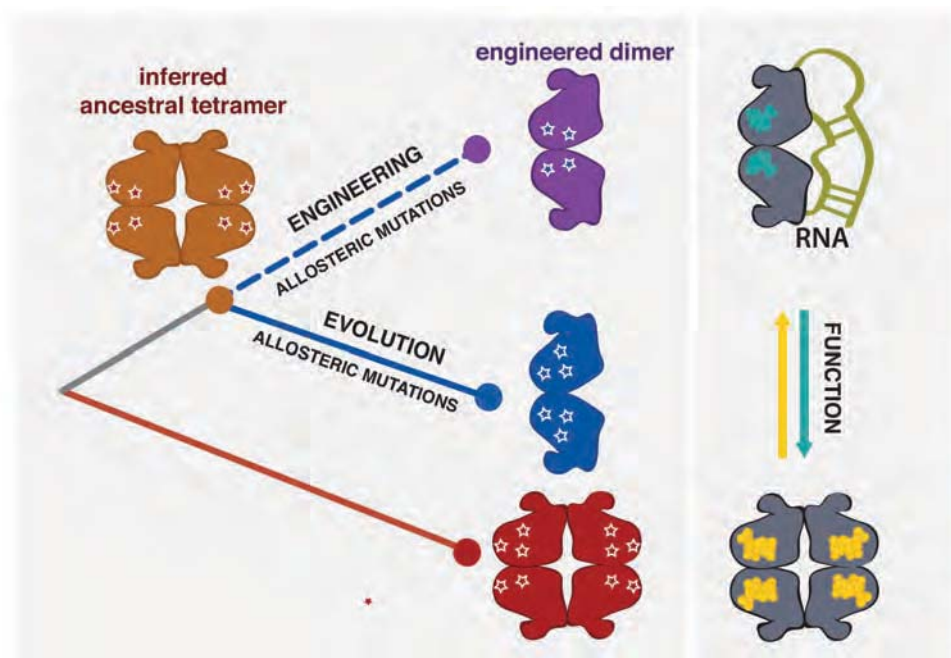
We will also pursue methods development in single-cell bioinformatics approaches. This is an exciting field still in its infancy, and there are many open questions that require new statistical and computational techniques. Together with the Marioni and Stegle groups at EMBL-EBI, we are keen to find new 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.

Selected publications

Mahata B, Zhang X, Kolodziejczyk AA (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep* 7:1130-1142

Marsh JA and Teichmann SA (2014) Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol* 12:e1001870

Perica T, Kondo Y, Tiwari SP, et al. (2014) Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* 346:1254346



Allosteric mutations can change oligomeric state by employing the same conformational dynamics as the allosteric ligands. PyrR homologues differ by mutations all of which are outside of the tetrameric interface. A subset of these allosteric mutations can be used to engineer a shift in oligomeric state in the ancestral PyrR. Allosteric mutations act by introducing conformational change in a manner analogous to the allosteric ligands.



Thornton group

Proteins: Structure, Function and Evolution

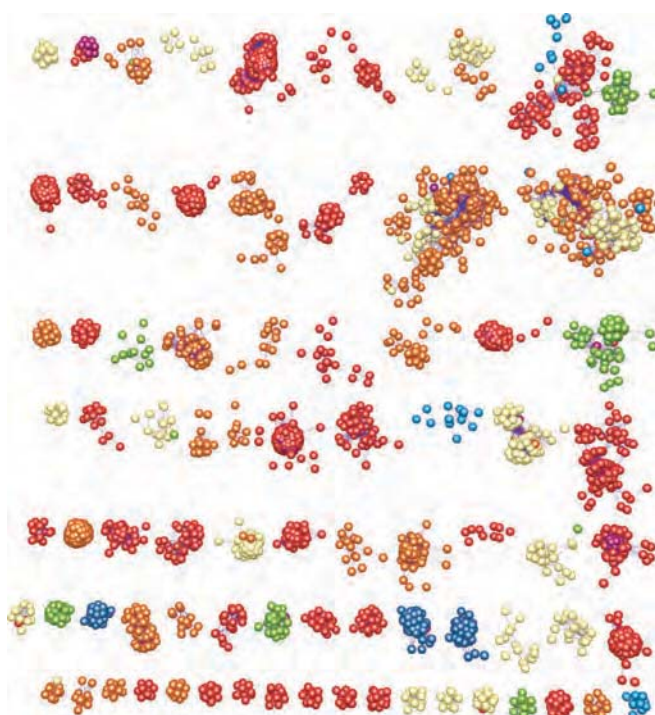
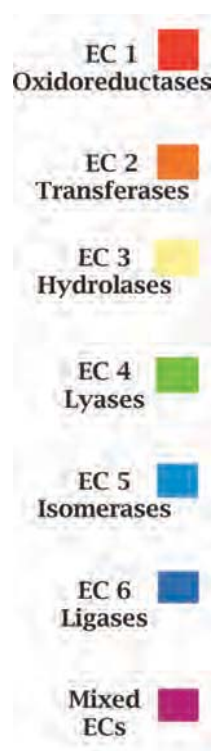
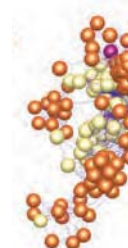
The goal of our research is to understand more about how biology works at the molecular level, with a particular focus on proteins and their 3D structure and evolution.

We explore how enzymes perform catalysis by gathering relevant data from the literature and by developing novel software tools that allow us to characterise enzyme mechanisms and navigate the catalytic and substrate space. In parallel, we investigate the evolution of these enzymes to discover how they can evolve new mechanisms and specificities. This 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 life span.

Major achievements

Clustering the universe of biochemical reactions: We developed EC-BLAST (Rahman et al, 2014), a new algorithm and web tool for quantitative similarity searches between enzyme reactions at three levels: bond change, reaction centre and reaction-structure similarity. It uses bond changes and reaction patterns for all known biochemical reactions derived from atom-atom mapping across each reaction. In order to get an overview of the universe of the chemical reactions, we derived the network of the 5073 representative reactions using a combination of bond and reaction-center scores. There are 785 individual clusters with more than one member (when a P value cutoff of <0.01 is used), of which 715 are pure clusters, whose members all have the same primary EC number. The figure shows the largest clusters, all with at least ten members. A few clusters (~7.6%) contain a mixture of reactions from different primary classes, shown in different colours, highlighting cases of shared chemistry between these enzymes, despite their disparate EC classification. Our study highlighted the complexity of enzymatic catalysis and the need for well-structured and accurate databases of enzyme reactions.



Characterizing the universe of enzyme reactions using EC-BLAST. Clustering of 5073 representative reactions, using a combination of bond and reaction-center similarity scores. Each sphere represents one reaction, colored by primary IUBMB EC (Enzyme Commission) class. All reaction similarity clusters with $P < 0.01$ and cluster size of more than ten reactions are shown.

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 since 2001.



EC-BLAST has the potential to improve enzyme classification, identify previously uncharacterised or new biochemical transformations, improve the assignment of enzyme function to sequences and assist in enzyme engineering.

The chemistry and classification of the isomerases: We explored the ability of EC-BLAST to capture the overall chemistry of the isomerases (EC 5) and to reproduce their EC classification (Martinez Cuesta et al., in preparation). The isomerases are a small class of enzymes that catalyze geometrical and structural rearrangements between isomers. Our results revealed that isomerase reactions are chemically diverse and difficult to classify automatically using the hierarchical approach adopted in the manual EC classification. Although racemases and epimerases (EC 5.1) and cis-trans isomerases (EC 5.2) can be easily grouped according to changes of stereochemistry in the substrate, the overall chemistry of intramolecular oxidoreductases (oxidations and reductions within the substrate - EC 5.3), intramolecular transferases (transfer of a chemical group within the substrate - EC 5.4) and intramolecular lyases (cleavage of bonds intramolecularly - EC 5.5) is diverse and defies automated classification on the basis only of bond changes and reaction centres. In addition the subclass 'other isomerases' (EC 5.99) sits apart from the rest of subclasses and exhibits even greater diversity. Strictly, there are just three types of isomerisation: enantiomerism (EC 5.1), cis-trans isomerism (EC 5.2) and structural isomerism (rest of subclasses). The matrix of bond changes shows that the division into sub-classes 5.3, 5.4 and 5.5 is complex and requires knowledge of the mechanism of the enzyme, not just the reactions performed. In addition, classification of the chemistry of some similar isomerase reactions needs a better approach e.g. oxidosqualene cyclases and pseudouridine synthases in intramolecular transferases acting on 'other groups' (EC 5.4.99).

GWAS study of longevity in flies: We used the *Drosophila melanogaster* 'Genetic Reference Panel' (DGRP) to map the genetic basis of natural variation in lifespan of virgin female flies from 197 lines, each genotyped for approximately two million common single nucleotide polymorphisms (SNPs) (Ivanov et al., submitted). We found considerable genetic variation in lifespan in the DGRP, with a broad-sense heritability of 0.413. The genome-wide association study had little power to detect signals at a genome-wide level in a single-SNP and gene-based analysis. Polygenic score analysis revealed that only a small proportion of the inter-line lifespan variation (~4.7%) was explicable in terms of additive effect of common SNPs ($\geq 2\%$ minor allele frequency). However, several of the top genes are involved in processes previously shown to impact ageing (e.g. carbohydrate-related metabolism, regulation of cell death, proteolysis). Other top-ranked genes are of unknown function and provide promising candidates for experimental examination. Genes in the Target of Rapamycin pathway (TOR) contributed to the highly significant enrichment of this pathway among the top-ranked 100 genes. Gene Ontology analysis suggested that genes involved in

carbohydrate metabolism are important for lifespan. This analysis suggests that our understanding of the genetic basis of natural variation in lifespan from induced mutations is incomplete.

Future plans

Our work to understand enzymes and their mechanisms using structural and chemical information will include a study of 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 and trying to identify small molecules that might modulate lifespan in the model organisms.

Selected publications

Rahman SA, Cuesta SM, Furnham N, et al. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat Methods* 11:171-174

Tullet JM, Araiz C, Sanders MJ, et al. (2014) DAF-16/FoxO directly regulates an atypical AMP-activated protein kinase gamma isoform to mediate the effects of insulin/IGF-1 signaling on aging in *Caenorhabditis elegans*. *PLoS Genet* 10:e1004109

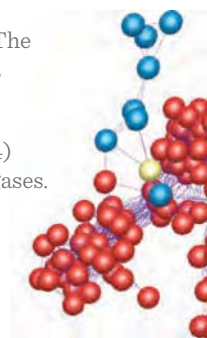
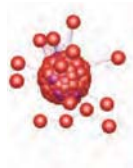
Papatheodorou I, Petrovs R, Thornton JM (2014) Comparison of the mammalian insulin signalling pathway to invertebrates in the context of FOXO-mediated ageing. *Bioinformatics* 30:2999-3003

Martinez Cuesta S, Furnham N, Rahman SA, et al. (2014) The evolution of enzyme function in the isomerases. *Curr Opin Struct Biol* 26:121-130

Holliday GL, Rahman SA, Furnham N, Thornton JM (2014) Exploring the biological and chemical complexity of the ligases. *J Mol Biol* 426:2098-2111

Links

EC-BLAST:
<http://www.ebi.ac.uk/thornton-srv/software/rbl/>





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.

In 2014, EMBL-EBI had its status as a University Partner Institute of the University of Cambridge reconfirmed, and our PhD students continue to be registered with and awarded their degrees by the University. EMBL-EBI benefited from

the presence of 45 PhD students in 2014, welcoming eight newcomers. Fifteen students successfully defended their theses and were awarded PhDs; four of them are showcased below. We also extend our congratulations to Stephan Beisen, Samuel Croset, Myrto Kostadima, Chen Li, Sergio Martinez-Cuesta, John May, Sarah Parks, Christine Seeliger, Robert Sugar, Sander Timmer and Ying Yan. EMBL-EBI Group Leaders John Marioni and Julio Saez-Rodriguez each celebrated their first PhD students' successful defences.

EMBL International PhD Programme events at EMBL-EBI in 2014

- *PhD Student Seminar Day at EMBL-EBI;*
- *Bioinformatics course for second-year EMBL PhD Programme students, organised and run by EMBL-EBI students;*
- *16th EMBL PhD Symposium, 'Inspired by Biology', at EMBL-Heidelberg;*
- *PhD student-led Lunchtime Seminar series;*
- *'Primers for Predocs' course at EMBL-EBI;*
- *Statistics training course, jointly held with the Wellcome Trust Sanger Institute at EMBL-EBI.*



EMBL-EBI predocs, March 2015. From top left: Rachel Spicer, Maria Xenophontos, Kevin Gori. Hannah Meyer, Mitra Barzine, Claudia Hernandez. Nils Eling, Anath Prakash, Emanuel Goncalves. Matt Jeffryes, Paolo Casale, Tom Rensch. Ewan Johnston, Julio Saez-Rodriguez (Research Group Leader), Christof Angermuller, Nils Kolling. Michael Menden, Billy Coleman-Smith, Oliver Stegle (Research Group Leader), Damien Arnol, Dimitrios Vitsios, Jose Bach Hardie, Greg Slodkowitz.

Jean-Baptiste Pettit

Spatial analysis of complex biological tissues from single cell gene expression data



Supervisor: John Marioni

Tissues within complex multi-cellular organisms have historically been defined in terms of their anatomy and function. More recently, experimental approaches have shown that different tissues express distinct batteries of genes, thus providing an additional metric for characterising them. These experiments have been performed at the whole tissue level; however, it is becoming apparent that even within putatively homogeneous tissues there is significant variation in gene expression levels between cells, suggesting that additional cell subtypes, defined by distinct expression profiles, might be obscured by 'bulk' experimental approaches. Jean-Baptiste developed a computational approach, based upon Markov Random Field models, for clustering cells into cell types by exploiting their gene expression profiles and location within the tissue. He demonstrated the efficacy of this approach using simulations, before applying it to identify known and novel cell types using a 169-gene expression atlas generated using in situ hybridisation within the brain of the ragworm *Platynereis dumerilii*, an important model for understanding how the Bilaterian brain evolved.

Jean-Baptiste also studied the related problem of characterising the full and spatially resolved transcriptome of every cell within the *P. dumerilii* brain. Current technologies allow whole-transcriptome sequencing of spatially identified cells but lack the throughput needed to characterise complex tissues. Jean-Baptiste, in conjunction with colleagues in the EMBL Developmental Biology Unit, developed a high-throughput method to identify the spatial origin of cells assayed by single-cell RNA-sequencing within a tissue. The approach is based on comparing complete, specificity-weighted mRNA profiles of a cell with positional gene expression profiles derived from the gene expression atlas he profiled. He showed that this method allocates cells to precise locations in the brain of *P. dumerilii* with a success rate of 81%. This method is applicable to any system that has a reference gene expression database of sufficiently high resolution.

Mar Gonzales-Porta

RNA sequencing for the study of splicing



Supervisor: Alvis Brazma

The human genome contains about 20 000 protein-coding genes, which can generate well over 100 000 different transcripts via alternative splicing. It has been argued that this is how the human genome generates protein diversity that is far larger than the number of its protein-coding genes. Mar explored this theory using public sequencing-based gene expression data, investigating whether human tissues do express large numbers of transcripts at significant levels. She found that on average, a human tissue expresses only between 10 000 and 15 000 genes to significant levels and, most importantly, that only one major transcript is expressed to significant levels for most of these genes. Her work also revealed that the human transcriptome is heavily dominated by these major transcripts. Mar used proteomics data to show evidence that the major transcripts are predominantly translated into proteins. She applied her methods to RNA data from a renal cancer project of the International Cancer Genome Consortium (CAGEKID) and showed that transcript switching in cancer can lead to loss of gene function without changes in overall gene expression.

Rita Santos

Differential drug response as a function of age



Supervisor: John Overington

Medicines are generally developed to be effective and safe in adult populations, and only a small subset of drugs has been formally studied for efficacy and safety in paediatric or geriatric populations. As a consequence, 45% of prescribed drugs in paediatric hospitals are given off-label, and approximately 20% of drugs given to the geriatric population fall into the same category. It is not easy to extrapolate doses, therapeutic indications, safety and efficacy from one age group to another: physical size, weight, body lipid fraction, metabolic state, drug target and differences in the molecular metabolic machinery expression can all affect drug response and safety. Understanding how some of these factors change at a molecular level has been a challenge to the safe translation of medicines from one population to another. Rita studied pharmacologically relevant gene expression changes as a function of age in human populations, in five primary organs (liver, brain, kidney, lungs and heart), to increase understanding of some of the age-related differential drug responses reported in the literature. In parallel, she also investigated how these changes translated to animal models (macaque, rat and mouse) commonly used in drug development.

Rita found that for drug targets associated with reported age-related differential drug responses, when given to either paediatric or geriatric populations, there was an overall tendency for those targets to decrease their expression with age. This pattern was particularly striking among ion-channels, and in contrast this pattern was not seen in rodents, where the opposite trend was detected. Taken together, these results could be used to identify drugs likely to have differential age-related responses and potentially to tailor therapeutic doses.

Camille Terfve

Modelling High Content Proteomics Data in a Signalling Context



Supervisor: Julio Saez-Rodriguez

During her PhD in the Saez-Rodriguez group, Camille Terfve studied how hundreds of proteins work within a cell in an orchestrated manner to process extracellular information. These proteins are highly interconnected in very complex networks. Camille developed methods to build mathematical models to understand how these networks work, and how they are affected by small-molecule inhibitors that are used as therapeutic agents. She focused her work on the use of mass spectrometry proteomics, a methodology of unrivalled coverage in terms of proteins and their alterations, but also a complex technology with various limitations.

She developed novel computational approaches and applied them to different cases, from the metabolic syndrome to the effect of small-molecule inhibitors on breast cancer cells.



Support





Summaries from support teams

Our technical services provide staff and visitors with safe, secure and authorised access to technical services that meet their current needs and will evolve in the future to meet their new requirements.

These teams are at the front line of many of the challenges in the life sciences that stem from the data deluge. Their goals are to provide a performant, scalable infrastructure for the long-term storage of publicly archived data, and to undertake rapid analyses of these datasets, making the results public as soon as possible.

Web production

- Completed the second migration of web services to the new data centre in Hemel Hempstead;
- Ensured the EBI Search engine was fit to index the 1 billion entries held in all EMBL-EBI data resources;
- Deployed a new, very fast RESTful API for the EBI Search and used it to integrate new services (e.g. RNACentral.org);
- Handled a sustained increase in the usage of EMBL-EBI sequence-analysis services, which totalled more than 100 million jobs in 2014;
- Maintained EBI Search and many of the institute's most heavily used computational tools (i.e. ranking amongst the top 10 services in the annual User Surveys).

Web development

- Built and deployed the Data Submission Wizard, which guides data depositors to the right resource for sharing their research outputs;
- Automated the process of displaying scholarly publications on the website, drawing on ORCID unique author identifiers;
- Contributed user-experience design to major projects, including the Centre for Therapeutic Target Validation (CTTV).

Systems and networking

- Accommodated a 23% increase in demand for compute from EMBL-EBI users, growing the Hinxton data centre from 12 000 cores in January to 17 000 in May;
- Coordinated efforts to grow our infrastructure to handle virtualisation of EMBL-EBI's database servers, with 94 Oracle virtual machines (VM) in use across the three data centres as of November;
- Installed new storage appliances to accommodate 25% growth in use of the virtual infrastructure;
- As coordinator of Embassy Cloud activities, invited 12 organisations to access the VMware-based cloud on a trial basis;
- Completed an ELIXIR pilot project with CSC Finland, providing a lightpath (i.e. ethernet circuit) over European Academic networks between CSC and EMBL-EBI;
- As part of the International Cancer Genome Consortium Pan-Cancer initiative, an 'Enlighten Your Research' collaboration and the Tara Oceans consortium, investigated open-source software solutions for the required scaling up of the EMBL-EBI cloud infrastructure from hundreds to thousands of cores;
- Oversaw doubling of storage on the EMBL-EBI Hinxton site from 3.7 petabytes in January to 7.4 petabytes in November (average storage utilization, 65%).
- Implemented the EMBL-EBI security committee's recommendations on user accounts;
- Completed the first phase of a LAN upgrade in Hinxton and integrated the new EMBL-EBI South Building into the campus network;
- Began to retire all tape-based backups, introduced our first Object Storage system and built software for a long-term data archiving facility;
- Migrated the Oracle and MySQL databases and Delphix from physical to virtualised infrastructure;
- Designed a resource-accounting method that is accessible via the web, which allows GTLs to control the database resources in use by their groups.



Support summaries

Training and industry engagement are essential for ensuring our services are relevant and useful to researchers in all sectors, and that they can be utilised to maximum effect. These activities also give rise to fruitful collaborations that help us look at biological problems in new ways.

Our External Relations and administration teams support the institute in countless ways, enabling our scientists and engineers in their careers and promoting their successes to a global audience.

Training

- *Actively involved 152 members of personnel in 231 events, reaching an audience of over 8500 people in 34 countries on six continents;*
- *Supported our colleagues to create 40 new courses in Train online, EMBL-EBI's free, web-based training resource;*
- *Accommodated doubling of Train online usage from 2013: 132 890 unique IP addresses accessed it and 5194 users registered;*
- *Through our train-the-trainer programme, supported the Australian Bioinformatics Network, Leicester University and new networks of trainers in Brazil and Africa, enabling them to deliver 12 courses reaching 345 people;*
- *Worked with Health Education England to define training requirements in the UK National Health Service for clinical bioinformatics;*
- *Contributed to the complete redevelopment of on-course®, the EMTRAIN course portal;*
- *Supported two new EMBL-EBI ambassadors to raise awareness of the institute to the scientific community;*
- *Together with Prof. Christine Orengo at UCL and the Genomes3D Consortium, co-led a project to develop a workflow-based approach to training in structural bioinformatics;*
- *Collaborated on developing a coordinated approach to course and conference planning throughout EMBL, one outcome of which was the launch of a new series of joint Wellcome Trust-EMBL conferences.*

Industry programme

- *Welcomed Biogen Idec (Cambridge, MA, US) as a new programme member (January 2014);*
- *Organised ten workshops on topics prioritised by our members;*
- *Held two workshops in the US, hosted by Novartis and Biogen Idec;*
- *Organised four quarterly strategy meetings;*
- *Jointly organised an SME Forum meeting with One Nucleus, and hosted the event in the EMBL-EBI South Building;*
- *Organised an outreach event in Buenos Aires, Argentina.*

Innovation & translation

- *Negotiated and concluded a contract between GSK, the Wellcome Trust Sanger Institute and EMBL-EBI, through EMBLEM, to establish the Centre for Therapeutic Target Validation (CTTV);*
- *Secured an agreement to share all CTTV data openly;*
- *Initiated discussions and identified other potential opportunities for further strategic alliances with industry.*

External Relations

- Organised EMBL-EBI's 20th anniversary celebrations, which comprised full event management, alumni and staff engagement, film production, creation of a bioinformatics timeline; design of décor and merchandise (e.g. flags, tee shirts) creation of web pages, video coverage and distribution of materials;
- Wrote and distributed 29 press releases and other news stories, supported service news and other writers, hosted several journalists and visiting film crews;
- Engaged with users on social media platforms including Twitter (all EMBL-EBI news shared with ~11,400 followers), Facebook (highlights and social news shared with 5,312 followers), LinkedIn (jobs and headlines posted to 4,291 followers) and Sina Weibo (experimental phase);
- Hosted and prepared materials for nine high-level visits to the Director and Associate Directors, and supported internal communication;
- Created and distributed print and digital publications including the EMBL-EBI Annual Scientific Report, Overview brochure, Industry Programme brochure and others, and contributed to EMBL-wide publications;
- Published cover artwork in two high-profile journals;
- Provided communications and branding support for strategic alliances such as ELIXIR and CTTV;
- Consulted on service and project branding;
- Created materials for conference exhibitions;
- Carried out staff inductions, which included taking photos for the website, introducing intranet resources and templates, reviewing news sites/social media and voicing core values and key messages to new members of staff.

Administration

- Organised and supported the data-centre tender exercise, acquiring facilities until the end of the current LFCF funding in 2019;
- Developed processes and procedures for measuring impact and benefits derived from LFCF funding;
- Further developed the internal EMBL-EBI budgetary process;
- Continued efforts to attract high-quality staff through targeted recruitment and advertising, and to improve their induction into EMBL-EBI;
- Contributed to the implementation of EMBL procedures and to the development of new reporting software;
- Sustained our Health & Safety practices and procedures.



Web Production

The Web Production team manages the web infrastructure comprising more than 1100 web servers running on 540 virtual machines. These provide platforms for web service development and robust, secure frameworks for deploying public bioinformatics services.

We are responsible for the global EBI search engine, which indexes more than 1 billion entries from all the EMBL-EBI data resources, and for the Job Dispatcher framework and corresponding SOAP/REST web services for programmatic access, which ran more than 100 million jobs during 2014. We also facilitate access to infrastructure services such as project-management and user-support core software.

Major achievements

A major data-centre move took place from 15 October to 23 December 2014, during which public services were migrated from two data centres in London to one in Hemel Hempstead. Our web administrators managed the relocation of virtual infrastructure, providing continued service to our users during the move. This involved replicating live services and fail-over capability in the Hinxton data centre whilst the individual data centre IT infrastructure was in transit.

The EBI Search engine in 2014 indexed over 1 billion data entries from all data resources at EMBL-EBI. The Job Dispatcher framework, through which users run Blast, InterProScan and Clustal Omega services, handled more than 110 million job requests during 2014 (compare to 65 million jobs in 2014).

The annual User Survey carried out in 2014 showed that the EBI Search and tools provided by Job Dispatcher rank amongst the top ten most used services – an outcome supported by our web stats and an increase in demand from service teams.

Uninterrupted service

In 2014 the Technical Services cluster completed the move of public services from Oliver's Yard and Power Gate in London to Gyron in Hemel Hempstead. This move involved physical consolidation of the IT infrastructure into a single location. During the move EMBL-EBI service usage increased considerably, and our team helped maintain a steady level of response to service requests. We deployed live services in our Hinxton facility for high-usage services (e.g. EBI Search, Job Dispatcher, uniprot.org, europepmc.org) and enabled a light-weight fail-over infrastructure for other services that made use of our testing and staging environments. Approximately 540 Virtual Hosts and 1100 distinct service end-points were involved in this operation.

Increased usage

During 2014 user traffic comprising web, FTP and Aspera totalled 14 terabytes a day. Job Dispatcher service ran more than 110 million jobs, an increase of 59% from 2013. The most popular services were InterProScan, Clustal Omega, pfam.xfam.org, PDB.org and InterPro. External users integrating these tools into their workflows and pipelines include Blast2GO, Galaxy, .Net Bio, BlastStation, BioServices, KEIO Bioinformatics Web Services, Yabi, BlastStation, STRAP, T-Coffee, CCP4, Geneious and GMU-metagenomics.

The number of datasets available for sequence searching via EMBL-EBI grew from 60K to 70K in 2014. This figure is based on the number of class- or taxonomy-specific files being produced to generate Blast and InterProScan services.

As part of the EMBL-EBI Tools Review committee, our team undertook to phase out underutilised services and to introduce new ones. We prepared to retire services such as maxsprout, dalilite, wublast and clustalw2 in 2015 and set up a process by which tools may be retired to alleviate unnecessary maintenance and ensure EMBL-EBI is providing users with scientifically accurate, well-maintained and up-to-date services.

Search engine API

In 2014 we deployed a new search engine API based on the RESTful web services model. This API is scalable, fast and easy to implement in third-party portals. RNAcentral (www.rnacentral.org) is a great example of this API in action.

To ensure maximum efficiency, we fully virtualised all search engine components in 2014, achieving a 20% reduction in the EBI Search memory footprint. We also improved the functionality of facets in the EBI search engine, which help our users filter results in a number of ways (e.g. using ontologies, controlled vocabularies, taxonomy, keywords, dates, molecule types). We overhauled facet functionality in the core Lucene libraries of the search system, which resulted in faster filtering of results and more consistent term-binding and quality assurance (e.g. spotting potential annotation errors).

Rodrigo Lopez

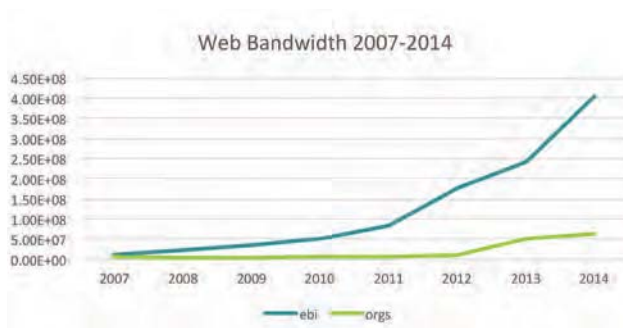
MSc Veterinary Medicine, Oslo Veterinaerhoyskole, 1984. MSc Molecular Biology and Toxicology and Informatics, University of Oslo, 1987.

At EMBL-EBI since 1995.



Reporting on usage

Our team fields all questions pertaining to web usage. In 2014 we retired the Business Intelligence web-log retrieval server and replaced it with a more scalable solution (based on the PIG language and Hadoop cluster). This solution provides usage reports by country, and service-based reports with accurate geolocation information. The Asia-Pacific region is now correctly represented, and in 2014 we could ascertain that countries in this region were responsible for 10% of our web traffic.



Web bandwidth, 2007-2014. During 2014 user traffic comprising web, ftp and Aspera totalled 14 terabytes a day.

Outreach, training and support

In 2014 we adopted a new help-desk system that works well with our feedback and development pipelines, and that empowers other service teams and research groups to handle their own support requests autonomously. Requests handled by our team were primarily concerned with programmatic access and data acquisition; technical problems with services; best practices; and training on specific resources.

We contributed to 11 bioinformatics training events, both in the context of the EMBL-EBI Training Programme and independently at external conferences and workshops. Our trainers provided comprehensive overviews of how best to access EMBL-EBI resources, discussed tools and techniques for sequence manipulation and searching (e.g. analysis of proteomic and immunogenetic sequences); walked people through different aspects of multiple sequence alignment; and provided developer training on the use of Web Service APIs.

Future plans

In 2015 we will identify and explore usage patterns associated with query types (e.g. gene names, database identifiers, chemical and biological names, taxonomies), which will inform future development. We will carry out the first full-scale analysis and, working closely with user experience designers in the Web Development team, will optimise the display and navigability of the EBI Search.

In 2015 we will further improve the interoperability of EMBL-EBI services with the EBI Search and will enhance data logistics to keep up with data growth. Based on the findings of the Tools Review committee, we will reduce the number of services we maintain and begin to defer requests. Our web administrators will focus on the consolidation and reorganisation of web resources in the new data centre and our web platforms specialists will move existing users of the SOAP web services to the new RESTful API. Importantly, we will reassess our capacity to provide web infrastructure services to the broad range of collaborations and projects we have supported in the past.

Selected publications

Lopez R, Cowley A, Li W, McWilliam H (2014) Using EMBL-EBI Services via Web Interface and Programmatically via Web Services. *Curr Protoc Bioinform* 48:3.12.1-3.12.50

Jones P, Binns D, Chang HY, et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236-1240

Silvester N, Alako B, Amid C, et al. (2014) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res* 43:D23-D29



Web Development

The Web Development team designs, develops and maintains the internal and external websites relating to EMBL-EBI's core activities, develops and maintains affiliated websites and acts as a central consultancy for web development and User Experience Design (UXD) at the institute.

We maintain the global EMBL-EBI website, its content database, the Intranet and training portals. We also develop and maintain over 30 ancillary web portals and services, including BioMedBridges, 1000genomes, INSDC, BioMedBridges and HGNC.

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

User Experience

EMBL-EBI resources are developed according to our users' needs, and every year we carry out a detailed User Survey. In 2014 analysis of over 1800 responses to our survey provided valuable insights that we shared with resource teams and consulted on change requests as appropriate.

A long-standing challenge for data depositors is determining which resource best suits their experimental outputs. We collaborated with archive service teams on the design, development and testing of a dynamic Data Submission Wizard, an intuitive interface that guides users through straightforward questions about their data and directs them to the most appropriate resource. This tool offers efficiencies for individual users and for publishers, who currently manage a range of submission instructions that require periodic updating.

We contribute user-experience expertise to the EMBL-EBI Tools Review committee, and in 2014 concluded a thorough exploration of search result display and discoverability of bioinformatics services and tools indexed by www.ebi.ac.uk. We will help implement changes to the website based on the results of this review in 2015.

Global website

In 2014 we consulted with stakeholders including EMBL-EBI teams, interested individuals, partner institutes (e.g. ELIXIR, GOBLET) and standards bodies (e.g. SASI) involved in organising or promoting courses, public events, seminars, webinars and other meetings. To ensure such events are

displayed to best advantage – and that they can be displayed easily on partner websites such as those hosted by ELIXIR Nodes – we developed a centralised resource for event information. The portal gathers images, descriptive text and facts pertaining to people, venues and programmes in a standardised manner, which allows for the creation of attractive, reusable content on www.ebi.ac.uk. Developing a common 'minimum exchange' set of fields and a common vocabulary for life sciences topics, in coordination with leading life sciences experts across Europe, was central to the success of this project. The new events portal will be launched in spring 2015.

Analytics are essential to any website's toolkit, and EMBL-EBI's unique position as a data provider for discovery activities adds a layer of challenge to protecting privacy whilst gaining content-development insights. In 2014 we carried out a trial of the open-source PiWik analytics set, which provides event tracking, user profiling, page transitions and environment information. A thorough review of PiWik in 2015 will establish whether it can be used to help inform decisions on improving websites and services.

On the technical front, we retired a number of project websites, moving them from the higher-maintenance dynamic Drupal sites to static HTML sites. This reduced the burden of maintaining Drupal updates and security patches for projects that had come to an end and had only minimal information-display requirements. We also upgraded several portals from Drupal 6 to Drupal 7, as Drupal 6 will no longer be supported as of 2015. We continued to work with service teams throughout the year to facilitate their adoption of web guidelines.

The Technical Services cluster started to create a coherent portal that provides, among other things, centralised information for staff to identify what services we offer and who to contact with queries. This portal will allow dissemination of some content to www.ebi.ac.uk.

Integration with embl.org

In 2014 we developed a new system that utilises staff ORCID IDs to populate publication information on appropriate areas of the website automatically. Because of the complexity of credit systems (e.g. publications arising from work done elsewhere), this involved our developing an interface that allows group and team leaders to decide whether a publication is excluded from their list. These activities bring us in closer contact with

Brendan Vaughan

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.



the EMBL IT teams, with whom we began working to converge broader technologies such as SAP and Converis, the grant and publications reporting software.

Developer Experience

EMBL-EBI developers use a variety of operating systems and environments and in 2014 our team created a new local development environment using Vagrant and Virtual Box. This environment replicates the production environment locally, effectively acting as a 'sandbox' that uses the same build instructions as the production servers. This secures an identical platform and has the added benefit of displaying to the developer their preferred local tools.

EMBL-EBI developers work in a rapidly changing technical environment and require on-going training and knowledge exchange. Addressing this need, in 2014 we organized a JavaScript training course that was substantially oversubscribed. The course was so well received that we will offer another in 2015.

Supporting strategic partnerships

In 2014 we completed the massive restructuring and redevelopment of the European Genome-phenome Archive (EGA) portal, and handed over complete responsibility for the service to EMBL-EBI's Variation team. The resource has gone on to be jointly developed by EMBL-EBI and CRG Barcelona.

We developed the global website for the Centre for Therapeutic Target Validation (CTTV) in Drupal, and continue to maintain and update the site. Our team hosts three new positions that are dedicated to the CTTV: a UX designer, a JavaScript developer and a BioJS developer. The CTTV has tight timelines for early releases, so our long-standing UX designers filled the gap between opening the positions and hiring. The beta release of the CTTV platform in 2015 will be based largely on the work of non-CTTV staff members, who put considerable efforts into the platform before transitioning new staff.

In 2014 we continued to engage with teams throughout EMBL and with external partners, collaborating and consulting on endeavours including ELIXIR, BioMedBridges, Europe PMC, the European Nucleotide Archive (ENA), the Protein Data Bank in Europe (PDBe), the ORCID foundation, the EMBL-EBI Staff Association, ChEBI and the EMBL-EBI Industry Programme.

Future plans

We are looking forward to launching the consolidated Events portal, in 2015 and testing event feeds to partner websites. We will also conduct substantial development work on the main website based on instructions and recommendations from the Tools committee, and transition the CTTV website from Drupal to the Embassy Cloud.

Drupal 8 will be released in 2015, and our team is fully trained and ready to make the transition. We will make use of our impact assessment to effect changes and improvements to EMBL-EBI websites. We have also trained in the BEHAT testing model, an open-source, behaviour-driven development framework for PHP, and will incorporate this technology into our coding cycle.

A thorough review of PiWik in 2015 will establish whether this software can be used to help inform decisions on improving websites and services.

Selected publications

McInerny GJ, Chen M, Freeman R, et al. (2014) Information visualisation for science and policy: engaging users and avoiding bias. *Trends Ecol Evol* 148-157

Corpas M, Jimenez R, Carbon SJ, et al. (2014) BioJS: an open source standard for biological visualisation - its status in 2014. *F1000Res* 3:55



Systems and Networking

The Systems and Networking team manages EMBL-EBI's IT infrastructure, which includes compute and database servers, storage, virtualization, private clouds, desktop systems, telephones and networking. We also provide database administration, support EMBL-EBI staff in their daily computer-based activities and manage the campus Internet connection. The team works closely with all project groups, maintaining and planning their specific infrastructures. We play a key role in the LFCF-funded frameworks.

Major achievements

Data centre migration

EMBL-EBI's lease with its data centre providers expired in 2014, and a new provider was selected following a public tender process. Our team was responsible for the migration of all services from two data centres in London to a single data centre in Hemel Hempstead. Our team successfully completed this enormous task without causing any unplanned downtime to any EMBL-EBI service. This was thanks to the thorough planning that we began in 2009, when we originally architected the systems to run in two external data centres.

The migration of EMBL-EBI's hardware required that we move 425 units of storage and compute; re-connect 850 power cables; re-connect and re-configure 3400 copper and fibre patch cables; provide a new network comprising 34 switches, two load balancers, four firewalls, two boarder routers, four SOIP switches; re-cable six Isilon clusters, a total of 170 nodes, which included 340 Infiniband connections; move approximately 22 petabytes of storage and 9500 compute cores.

Computing

We organised several training days for EMBL-EBI staff who use our hardware clusters. The largest of these clusters is the ebi cluster, based on 22068 CPU cores, 762 hosts and 97 terabytes of memory. As of December 2014, we manage 29166 CPU cores, 1693 hosts and 115 Terabytes of memory in our clusters.

Our computing team put significant efforts into developing 'elastic' clusters, which enable us to adjust clusters dynamically in order to maximise the efficiency of our equipment. These clusters will go into production in 2015.

We enabled the first-ever EMBL-EBI parallel file-system on the ebi compute cluster in 2014. We also retired a lot of old hardware and installed new servers in their place.



EMBL-EBI's public-facing data centre is up and running at Gyron in Hemel Hempstead, UK. The move from London was effected without any significant down time.

Networking

In 2014 we completed the transition to Janet6, the UK's national research and education network. EMBL-EBI now has three ten-gigabyte IP links to Janet6 (one active, two redundant) and two ten-gigabyte L2 links for connectivity to Hemel Hempstead.

We refreshed a large part of the local area network (LAN) in the Hinxton data centre and in the Flint Cross site (our disaster-recovery site). We also refreshed the network infrastructure for the European Genome-phenome Archive, a secure resource for clinical research data that is co-developed by EMBL-EBI and CRG Barcelona. We also took the opportunity to refresh the network in the East Wing.

We designed a completely new networking infrastructure, and built it in our new data centre space. The new design has several advantages, notably: There is no single point of failure; dual paths are available for all devices that support this feature; there is a 160 gigabyte/second path between any pair of ports on the LAN; there is a completely separate management network, with out-of-band (OOB) connectivity; the network supports over 800 10G optical ports and more

Petteri Jokinen

MSc in Computer Science 1990,
Helsinki University.

At EMBL-EBI since 1996.



than 300 1-gigabyte copper ports; there is 'single pane of glass' for production ports (two virtual chassis); and there are approximately 200 serial-over-IP ports for remote serial access.

Storage

We upgraded the 'ebi' cluster to have two parallel file systems: IBM GSS (mainly for research use) and Panasas (for production). We retired nearly all of the out-dated, small SAN systems and installed several larger, virtualised SAN systems (three EMC VNX arrays, three HP 3PAR arrays, Tegile hybrid system). We also provided a long-term storage archive, which was activated in January 2014.

Our team implemented a new version of the FIRE architecture, which is used to implement the Sequence Archive. It now supports object-based storage in addition to NFS and parallel file-systems. We also implemented an Object Tape Archive, which will hold the second copy of SRA on tapes. As of December 2014, the total occupied space for real data in our shared filesystems is 25 petabytes (compare to 15 petabytes in 2013).

Future plans

In 2015 we plan to improve high-performance IO, database storage, the FIRE archive and the Object Tape Archive. Our goal is to get the elastic clusters operational in Hemel Hempstead and in Hinxton as early as possible in 2015.

We will complete the Hinxton LAN retirements, improve monitoring, reporting and documentation of network and data centres. We will also provide any support and help needed for staff to implement projects initiated by the Technical Services cluster.





Training

EMBL-EBI provides an extensive user training programme to ensure that our users can access EMBL-EBI's data efficiently and get the most out of their own datasets when comparing them with the public record. Our programme is coordinated and funded centrally, and has regular input from technical experts throughout the institute.

This arrangement sets EMBL-EBI's courses apart. Our training activities offer unique interactions between service developers and users, and provide invaluable input to 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.

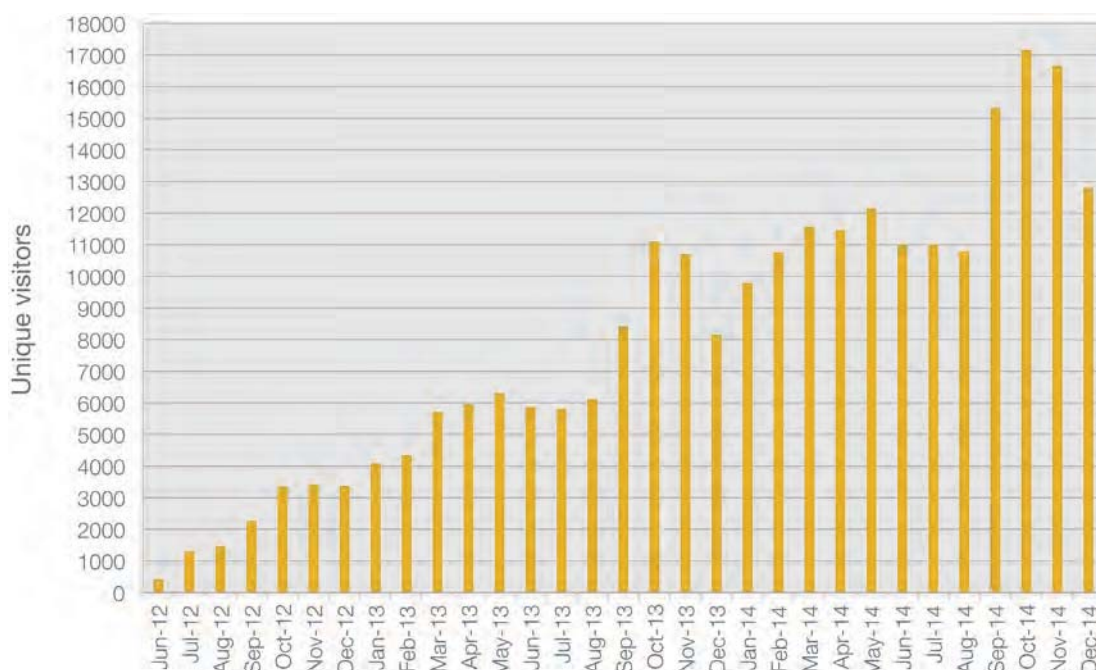
Major achievements

In 2014, EMBL-EBI staff orchestrated 167 events throughout the world and contributed to a further 64 organised by others. These included training courses at EMBL-EBI, off-site training events, conference exhibitions, careers fairs and workshops. Our new courses included animal-genome informatics and genotype-to-phenotype mapping. Our off-site workshops took us to Brazil, India, Japan, Norway, Portugal and Italy, among other countries. 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.

Over the past two 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 2014 we extended our reach to Africa in a collaboration with the H3Africa initiative. We also ran our first trainer-support workshop for scientists in Brazil, as part of a BBSRC-funded Brazil Partnering Award. Our five 'apprentice trainers' led training activities on the animal-genome informatics course, immediately consolidating what they had learned.

Train online, EMBL-EBI's web-based training resource, is now the fastest growing part of our training programme. In 2013 we had 82 000 unique users (based on unique IP addresses, which may represent all users at a single institute); in 2014 this grew to 132 890 unique IP and over 5000 registered users. Keeping pace with demand, we supported our colleagues throughout EMBL-EBI to develop 40 new courses.



Usage statistics for Train online, June 2012 to Dec 2014.

Cath Brooksbank

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.



European training infrastructure

We are a partner in ELIXIR's UK node, which currently focuses on training. Working with the structural bioinformatics sector, Christine Orengo, Cath Brooksbank and the Genome3D consortium made an inventory of existing online training materials in structural bioinformatics and designed a workflow-based approach to organising training materials on the basis of common research questions. Consequently, structural bioinformatics has been identified as the top priority for further development in the ELIXIR-UK work plan and will serve as a test case for the pilot Training e-Support System – ELIXIR's aggregator for training information.

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 2014 we launched the website for LifeTrain – a pan-European framework for continuing professional development in the biomedical sciences. We contributed significant input into the complete redevelopment of on-course®, EMTRAIN's comprehensive online course catalogue. Short courses continue to be the fastest growing part of on-course, with the EMBL-EBI training team making a major contribution to this growth.

LifeTrain's core principles have influenced our work on other projects. For example, we contributed to a task force, commissioned by Health Education England (HHE) and chaired by EMBL-EBI Director Janet Thornton, to provide input into the development of clinical bioinformatics training for the UK National Health Service. Our contribution comprised a matrix capturing the bioinformatics competencies that will be required by healthcare professionals to make the most of data emerging from clinical sequencing projects for the benefit of patients. This will be used to inform the development of HEE's clinical bioinformatics training.

Face-to-face training in action in one of EMBL-EBI's new IT training rooms.

Future plans

As EMBL-EBI's user base continues to diversify, scalability of the EMBL-EBI training programme becomes a major concern. Online learning has the potential to address this issue, and in 2015 we will continue to extend our online training offering, incorporating more webinars and videos into our portfolio. We look forward to working with our many collaborators to continue delivering both established and new hands-on courses to our users, and to build training capacity through train-the-trainer initiatives.

Selected publications

Brooksbank C, Bergman MT, Apweiler R, Birney E, Thornton J (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res* 42 (database issue), d18–d25

Welch L, Lewitter F, Schwartz R, et al. (2014) Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput Biol* 10, e1003496

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



Innovation & Translation

We collaborate in different ways with companies of all sizes to support Europe's thriving life-science industry.

The central pillars of these activities are the EMBL-EBI Industry Programme, founded in 1996; shared grants through the Innovative Medicines Initiative; and focused, large-scale, public-private partnerships.

In 2014 we launched a new programme to identify, implement and manage long-term, pre-competitive and translational projects and strategic partnerships between EMBL-EBI and industry. The first of these is the CTTV, a joint project between EMBL-EBI, the Wellcome Trust Sanger Institute and GSK that was directly enabled by LFCF funding to expand available space and computational infrastructure at EMBL-EBI.

EMBL-EBI Industry Programme

The EMBL-EBI Industry Programme has been a vibrant part of EMBL-EBI since 1996, providing regular contact with commercial users and informing the institute's priorities. We support pre-competitive collaboration through a subscription-based programme for larger companies representing the pharmaceutical, agri-food, nutrition, healthcare and consumer goods sectors.

In 2014 our Industry Programme ran ten knowledge-exchange workshops for its members, including two in the US. We held a major outreach event in EMBL's new associate member state, Argentina, and co-hosted an event for small and medium-sized enterprise (SMEs) with One Nucleus and the Council for European Bioregions on the Wellcome Genome Campus.

Industry Programme members in 2014

- *Astellas Pharma Inc.*
- *Bayer Pharma AG*
- *Boehringer Ingelheim*
- *Eli Lilly*
- *GlaxoSmithKline*
- *Merck Serono S.A.*
- *Novartis Pharma AG*
- *Pfizer Inc*
- *Syngenta*
- *Unilever*
- *AstraZeneca*
- *Biogen Idec*
- *Bristol-Myers Squibb*
- *F Hoffman-La Roche*
- *Janssen Research & Development, LLC*
- *Nestlé Institute of Health Sciences*
- *Novo Nordisk*
- *Sanofi-Aventis R&D*
- *UCB*

Finding new drug targets

Picking the wrong target is a major reason why medicines fail in their development, and the CTTV aims to address this by improving the validity of drug-target data. Using genome-scale experiments and analysis, the partners are working to provide evidence on the biological validity of therapeutic targets and to provide an initial assessment of the likely effectiveness of pharmacological intervention. The Centre is committed to sharing its data openly with the scientific community.

In 2014 Ewan Birney, Associate Director and Senior Scientist at EMBL-EBI, was Interim Head of the CTTV and Ian Dunham its Scientific Director. The CTTV involves staff from each of its partner institutions, interacting in our Innovation and Translation suite and participating in the full scientific life of the campus.

EMBLEM: enabling new partnerships

Birgit Kerber and Jason Munding, EMBL Enterprise Management Technology Transfer GmbH (EMBLEM)

EMBLEM supports EMBL-EBI by identifying opportunities for collaborations with industry and by ensuring these public-private endeavours benefit both the organisations involved as well as the wider scientific community. In 2014, EMBLEM played an important role in setting up the CTTV, and securing key agreements for the Single Cell Genomics Centre and Embassy Cloud.

The CTTV agreement, signed in 2014, is well aligned with EMBL-EBI's core values. It is pre-competitive, combining ground-breaking experimental and computational approaches to enable transformational change in the selection and validation of targets; it is non-exclusive, aiming to attract other external members; and it is research enabling, with the potential to extend beyond the human target validation arena. EMBLEM negotiated and concluded the agreement and, critically, was instrumental in securing an agreement to share all CTTV data openly.

CTTV activities in 2014

Initial work in the CTTV core focused on integrating data sources that affect the validity of a target in a single infrastructure, allowing seamless interrogation of these data. BioMedBridges, an ELIXIR project coordinated by EMBL-EBI,

played a key role in these endeavours. In preparation for the launch of a public CTTV service in 2015, our Web Production team provided extensive usability research and user-driven web design.

Computational pipelines activities centre on ensuring that data are described in a machine-readable form and can be queried in a standard way that allows access to pertinent information including phenotypic relationships and distinctions. Our Samples, Phenotypes and Ontologies team worked to establish ontologies as infrastructure, providing sets of common terms and defined synonyms and specifying relationships between terms across disparate disease areas, with an initial focus on inflammatory bowel disease and cancer.

To understand what types of information are most needed for CTTV research activities, the partners initially undertook specific projects in three therapeutic areas: oncology, inflammatory bowel diseases and respiratory disease. In each of these, data generation was coupled with targeted analysis to further target validation. Exploratory projects were initiated to explore new generic research areas for target validation.

Embassy Cloud

Organisations of all sizes find it challenging to provide adequate infrastructure to manage Big Data. Embassy Cloud, formally launched in 2014, provides private, secure, virtual-machine-based workspaces within the EMBL-EBI infrastructure, hosted in our Tier 3+ secure data centre in Hemel Hempstead. Users can make optimal use of their own customised workflows, applications and datasets, with direct access to EMBL-EBI data, services and compute. This is a practical and cost-effective alternative to replicating services and downloading vast datasets for local use. Embassy Cloud users can access their workspace from anywhere, reducing the need for capital investments in hardware and related operational costs.

In 2014 the Embassy Cloud received significant hardware investment to support international collaborations such as the International Cancer Genome Consortium (ICG) and the EU-funded COMPARE project.

Single-Cell Genomics Centre

In 2014, with the help of EMBLEM, we signed a formal collaboration with the Wellcome Trust Sanger Institute and Fluidigm Corporation to accelerate the development of new methods for the analysis of single-cell genomics data. The Single Cell Genomics Centre (SCGC) began working with onsite Fluidigm senior staff, who ensure that the centre has early access to the latest equipment, workflows and methods for genomics and proteomics research. This partnership enables our researchers to develop new experimental and computational methods, which in turn will help the technology mature more quickly. For example, the Teichmann group discovered that immune cells produce steroids to regulate themselves – knowledge based on mRNA-seq data from single cells. Fluidigm technology also enabled the Marioni group to develop a novel statistical method that shows how single-cell mRNA sequencing can be used to pinpoint true differences between cells in apparently homogeneous samples.

Innovative Medicines Initiative projects in 2014

DDMoRe: The Drug Disease Model Resources consortium

In 2014, BioModels developers at EMBL-EBI contributed to DDMoRe's launch of a new repository for computational models of disease used to describe interactions between drugs and patients. Featuring a flexible format and following a new standard, this public, open-access resource will make it easier for researchers to share and reuse models of drug action and disease progression using their own software.

EMTRAIN

We participate in EMTRAIN, a project to establish a European platform for professional development in the biomedical sciences. In 2014 the project launched the LifeTrain framework for continuing professional development. Our Training Programme provided substantial input to the complete redevelopment of the EMTRAIN online course catalogue, and were major contributors to its growth.

European Autism Interventions (EU-AIMS)

In 2014 our Functional Genomics Development team provided data management solutions for 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.

European Medical Information Framework (EMIF)

Our Functional Genomics Development team integrated their R cloud scientific computation infrastructure into the EMIF medical data exploration system in 2014. This will help the project enable the reuse of patient health records in clinical research.

Open Pharmacological Concepts Triple Store

Open PHACTS, a project in which our ChEMBL team is a regular contributor, integrates pharmacological data across diverse resources, reducing barriers to drug discovery in industry, academia and for small businesses.

eTOX

The eTox project, which also includes our ChEMBL team, is developing innovative in silico strategies and novel software tools to better curate toxicity data and use it to predict toxicity in early stages of the drug development pipeline.

Electronic Health Records for Clinical Research (EHR4CR)

EHR4CR is designing a scalable, cost-effective approach to interoperability between electronic health record systems and clinical research.



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 the institute's priorities.

We support pre-competitive collaboration through a subscription-based programme that welcomes larger companies that make use of the data and resources provided by EMBL-EBI. Member companies represent the pharmaceutical sector and, increasingly, 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 forum supports SMEs through subsidised outreach activities, held primarily in the new EMBL-EBI South Building. In the past we have held these at host institutes in EMBL member states, and they have been very well received. ELIXIR will be taking these activities forward, raising awareness of the utility of public data resources for SMEs throughout Europe.

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 many 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 2014 we ran ten of these workshops, which are a major benefit of membership: eight at EMBL-EBI and two in Cambridge, Massachusetts. 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 Novartis-hosted workshop covered the ENCODE project and epigenomics, and the Biogen Idec-hosted event focused on causal inference.

The opening of the South Building in 2013 has provided opportunities for working with companies in new ways. In March 2014 we ran the first of our new UK series of annual events for small and medium-sized enterprise: the SME Bioinformatics Forum. This was jointly organised with One Nucleus, which is 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.

In 2014 Argentina became EMBL's newest associate member state, and we were asked to organise one of the first bioinformatics events for industry to mark the occasion. In collaboration with the Argentine Ministry of Science, Technology and Productive Innovation (MINCYT) and the Argentine Chamber of Biotechnology (CAB), we organised an industry outreach meeting in Buenos Aires that was very well attended and received.

The CTTV, also located in the South Building, was launched in 2014 by the new Translation and Innovation programme and represents a significant extension of EMBL-EBI's interactions with GSK. The CTTV, which will lead to many new tools and datasets being available in the public domain, is discussed in detail in the Innovation and Translation section.

Future plans

The Industry Programme will continue to adapt and seek innovative methods of interaction with our industry partners 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. We see our interactions with industry partners becoming stronger as we work together to address shared challenges and opportunities created by Big Data. In addition to finding cost-effective paradigms to manage expanding volumes of data, we will collaborate to establish methods to ensure 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.

Dominic Clark

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
(secondment 2006–2009).



Industry workshops

- *In silico ADMET prediction*
- *Causal inference: algorithms, methods and resources (USA)*
- *Bio-JS (Bio Java Script)*
- *Systems pharmacology*
- *Biologics informatics*
- *RNAseq and ChIP-seq data analysis*
- *Shared data, shared costs*
- *ENCODE and epigenomics (USA)*
- *Nutrition information, ontologies and nutrigenomics*
- *Informatics for rare diseases research and repurposing: precompetitive opportunities.*

Selected publications

Shirai H et al. (2014) Antibody informatics for drug discovery. *Biochimica et Biophysica Acta* 1844:2002–2015

Hendrickx DM, Aerts HJ, Caiment F, et al. (2014) diXa: a Data Infrastructure for Chemical Safety Assessment. *Bioinformatics* 2014.

Rebholz-Schuhmann D, Grabmüller C, Kavaliauskas S, et al. (2014) A case study: semantic integration of gene-disease associations for type 2 diabetes mellitus from literature and biomedical data resources. *Drug Discov Today* 19:882–889

Innovative Medicines Initiative projects

Our programme helps its partners and EMBL-EBI researchers establish collaborations in the context of the Innovative Medicines Initiative. Some of these are listed below.

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.





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 audiences both in person and using a wide range of media.

We support the work of EMBL-EBI's many ambassadors, in particular the institute's Director, Associate Directors and team leaders, in fostering good relations with policymakers, funders, potential collaborators and service users throughout the world. We welcome visiting delegations of scientists, politicians and industry representatives, and work with leadership to refine the delivery of key messages. We endeavour to raise the profile of the EMBL-EBI brand by generating high-quality content and disseminating it through the press, the global website, social media and printed publications. Our goal is to convey the value of EMBL-EBI to targeted audiences in a clear and professional manner.

Major achievements

As EMBL celebrated its 40th anniversary in 2014, EMBL-EBI marked its 20th with a day of celebrations on the Wellcome Genome Campus. For our team, the event was the culmination of a full year of planning and production. We created a film featuring interviews with founders and key influencers; compiled and created a 19-metre-long timeline of important events in bioinformatics; engaged with alumni and staff to encourage networking and share photographs; designed and produced branded tee shirts, mugs, banners, flags and many other materials; liaised with an array of vendors; created web pages for the event and through it shared videos and photos from the day. Following the event, we reused the materials we created in many ways.

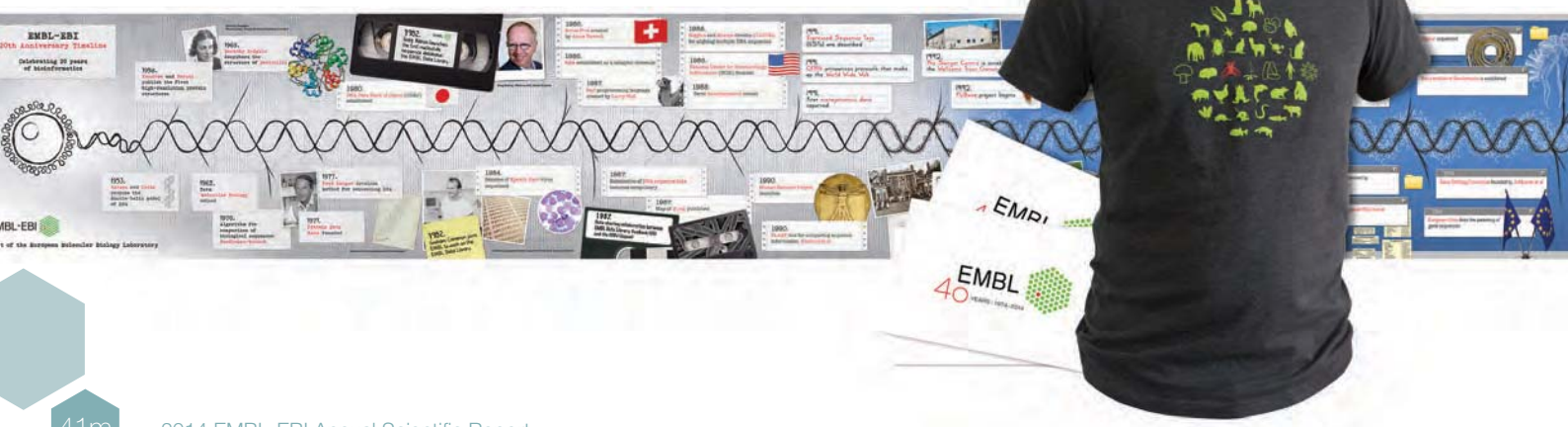
In 2014 we wrote and distributed 29 press releases and other news stories, covering breakthrough research, new service

launches and major developments such as the opening of the CTTV. We supported writers throughout the organisation by editing service news and blog posts, and hosted several visiting journalists and film crews. We distributed news stories directly to journalists, and signposted to news content on our website and the EMBL news site using social media platforms (Twitter, ~11,400 followers; Facebook, 5,312 followers; LinkedIn, 4,291 followers).

Our team created and distributed print and digital publications including the EMBL-EBI Annual Scientific Report, overview brochure, Industry Programme brochure and others, and contributed to EMBL-wide publications including *Research at a Glance* and the EMBL Annual Report. We also designed materials for conference exhibitions, including programmes highlighting EMBL-EBI activities at Plant and Animal Genomes and Intelligent Systems for Molecular Biology. To address internet availability challenges at external conferences, we created a pilot EMBL-EBI app to encourage visitors to browse a tablet to see what's on offer at EMBL-EBI.

We use graphic design to help our institute's leaders convey the excitement and impact of bioinformatics to specialist and non-specialist audiences alike. In 2014 one such project was a large-scale visual presentation on using synthetic biology for digital data storage to a non-scientific audience at the World Economic Forum's annual meeting at Davos. We also created artwork that was featured on the cover of *Developmental Cell*.

External Relations is the central point of contact for brand and visual identity, and in 2014 we provided



Lindsey Crosswell

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.



communications and branding support for strategic alliances such as ELIXIR and CTTV in the form of web content consultation, key message development, printed publications, graphics and slide presentations. We contributed to the rebranding of the Wellcome Genome Campus (to be rolled out in 2015), and consulted and provided design for services including Embassy Cloud, the CTTV, RNAcentral, European Variation archive and several others.

When new staff arrive at EMBL-EBI, we take the opportunity to make them aware of our Institute's mission and core values, show them the outreach materials we share via the intranet and point them to different sources of EMBL news. We also take their photo for the web and for future use in publications if required. In 2014 we inducted and photographed approximately 175 members of staff and visitors.

We collaborate with our colleagues in Heidelberg on many programmes and initiatives, including alumni relations and development. In 2014, recognising the need to explore new sources of revenue for the institute, we collaborated on a strategy for creating a culture of philanthropic giving to the institute. We supported our Directors and senior scientists in the hosting of two breakfasts at the Royal Society aimed at engaging influential figures from outside of the world of

science in the work of the institute. We continue to pursue opportunities to engage more widely to attract philanthropic support which will enable us to broaden the scope of EMBL-EBI's research and services.

Future plans

In 2015 our team will continue to promote the institute to many audiences through diverse channels, and to support EMBL-EBI's leaders and many ambassadors in their outreach efforts. In close collaboration with the Web Development team and our colleagues in Heidelberg, we will endeavour to build a shared vision for EMBL's digital content strategy so that we can make the most of our global website through high-impact written and visual content.

Publications

Crosswell, LC and Thornton, JM (2014) Databases and data sources: EBI and ELIXIR. In: Anders Brahme, Editor. *Comprehensive Biomedical Physics*. Elsevier, 175-190





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.

Our activities span budgetary, financial and purchasing matters; human resources; grants and external funding management; facilities management; health and safety; on-going support, along with our Systems and Networking colleagues, for the UK Large Facilities Capital Fund (LFCF); and pre- and post-doctoral programmes. We coordinate and integrate administrative activities throughout EMBL-EBI to promote interactions with the wider scientific community, for example by organising meetings and courses and arranging travel for our extremely mobile staff.

Major achievements

LFCF project delivery

Our team provides project management for major funding projects. In the UK's Life Sciences Strategy, announced in 2011, provision was made under the LFCF for a programme of work designed to help meet the growing demand for EMBL-EBI services and, in the context of ELIXIR, to support life science research and its translation to medicine and the environment, the bio-industries and society. This programme encompasses the construction of the EMBL-EBI South Building, home of the ELIXIR Hub, and the delivery of biological data services from robust and reliable 'Tier III' data centres.

Data Centres

In 2014 our team devoted substantial efforts to the competitive tendering of data-centre space for 2015-2019. This successful procurement exercise led to the migration of EMBL-EBI's services to Hemel Hempstead, and we supported our colleagues in Systems and Networking and many Service Teams in undertaking this well-orchestrated endeavour. We also deepened our work on impact-and-benefit assessment arising from the LFCF funding. This included liaising with the EC-funded EvarRIO project on reporting the impact of research infrastructures, and commissioning Charles Beagrie Ltd to undertake a preliminary case study as a precursor to a more details and rigorous study.



EMBL-EBI's data centres supply many petabytes of molecular data to scientists around the world, 24/7. Gyron Ltd was awarded the contract for EMBL-EBI's data centres from 2015 to 2019. Pictured: Mark Green, Head of EMBL-EBI Administration, with Gyron's Robin Balen and Dan Clark. Photo credit: Robert Slowley

Finance

Our grants office contributes to the development of EMBL-wide administration of external funds and plays a significant role in the review of existing grants. We handled 128 applications with call deadlines in 2014 (compare to 78 in 2013). This excludes applications without call deadlines.

Our finance and purchasing team members continued to exercise sound financial control, following the introduction of EMBL-wide scanning processes for invoices, and the more efficient system introduced early in the year bedding in well.



Mark Green

Fellow of the Chartered Institute of Internal Auditors. At EMBL since 1997; joint appointment with EMBL-EBI.

At EMBL-EBI since 2003.



Human Resources

Our human resources team members continued to improve their working processes and to develop internal documentation that helps us provide accurate and consistent advice to staff and supervisors. In 2014 we introduced annual performance assessments, accompanied by training for supervisors. We recruited 120 new members of staff in 2014 (compare to 67 in 2013).

We participated in a wide range of cross-campus activities and initiatives including EMBL-EBI/Wellcome Trust Sanger Institute meetings, Health and Safety, Campus Library and the 'Sex in Science' programme.

We work closely with EMBL Administration in Heidelberg to ensure that all staff have the administrative support they need. We have an active voice in the overall development of strategic objectives for administration and identification of opportunities for improving efficiency, for example joint agreements with recruiting agencies.

Future plans

We will continue to develop longer-term strategic financial plans taking account of EMBL, external and LFCF funding. We will help implement the new EMBL Business Warehouse/Objects software that will facilitate analysis and reporting of financial and personnel data and contribute to the EMBL-wide effort to replace some of our paper processes with electronic workflow and approvals. We will continue to foster good interactions among a wide diversity of stakeholders such as the BBSRC, Wellcome Trust and NIH. During the course of 2015 we will implement a new automated leave-recording system as part of the on-going HR Blueprint project. We will undertake an analysis of 'exit interviews', and introduce a newcomers survey to evaluate our recruitment and induction processes. We will continue to support the LFCF procurement process by extending the current framework of equipment suppliers for a further year, whilst continuing to investigate whether we can make use of other public frameworks between September 2015 and 2019. We will continue our endeavours to measure impacts and benefits, and begin writing the business case for continued LFCF funding beyond 2019.





Facts and figures





0.36521

0.85234546888966

0.5624911256657858858/565

0.36521

0.2521540889

0.5624911256657858858/565

0.85234546888966

0.029632454

0.4923322

0.02654225983

5.767

0.95623

3.568

4.261

4.6223

3.568

0.95623

0.0235

0.35854

0.28679970

0.7047345332

0.6823

0.0235

0.485

3.56

0.95623

1.5656

0.35854

0.0235

1.89744

0.752

0.4576955

1.568

5.7542



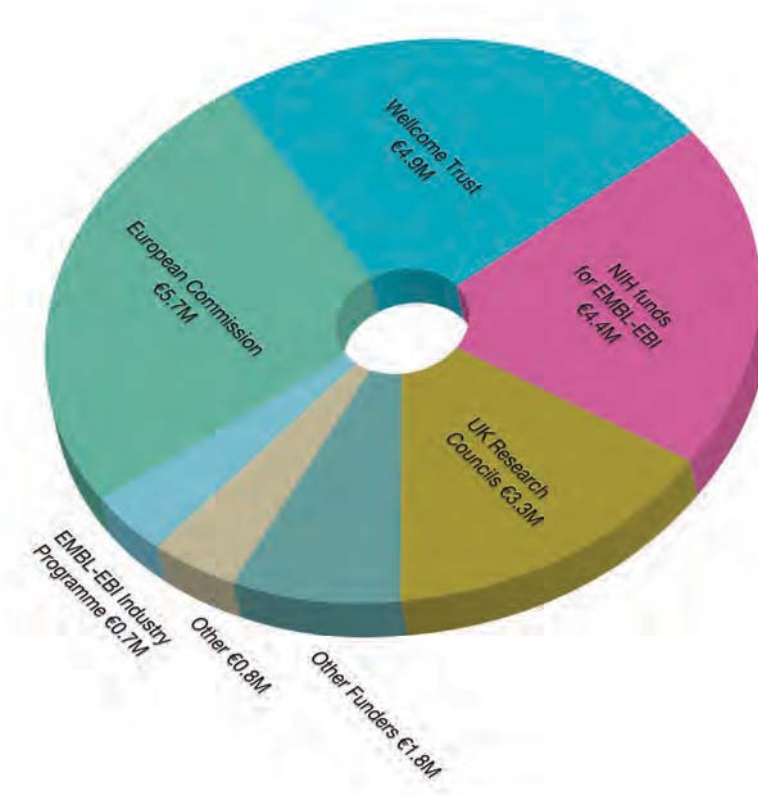
Funding and resource allocation

EMBL-EBI funding remained stable in 2014. This continued support of our member states and other funding bodies in 2014 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 2014. The 'external funds' shown here represent both funds that were available for our use in 2014 and those earmarked for subcontractors as part of our grant funded activities.

Sources of funding

Funding for EMBL-EBI, excluding sums earned earmarked for project subcontractors, in 2014 was €58.3 million and comes primarily from EMBL member states (€36.7 million). Our major sources of external funding include the European Commission (€5.7 million), the Wellcome Trust (€4.9 million), the US National Institutes of Health (€4.4 million to EMBL-EBI for direct use) the UK Research Councils (€3.3 million) and the EBI Industry Programme (€0.7 million). We also benefit from a large number of grants from various other sources (total, €2.6 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 2014 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

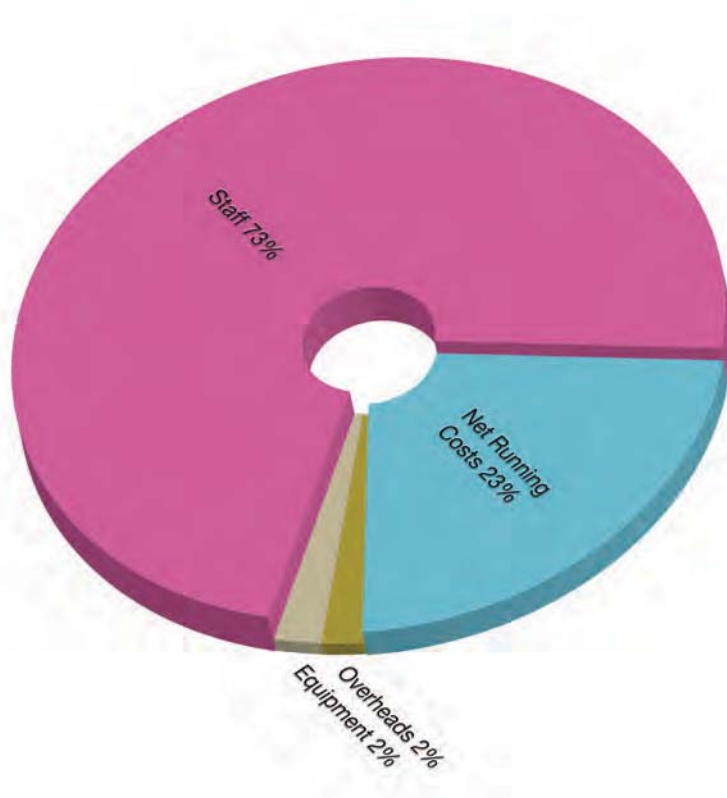
The UK Government's Large Facilities Capital Fund has provided for the EMBL-EBI South Building (officially opened in 2013), which houses ELIXIR and an Innovation and Translation Suite, and for the on-going use of Tier III data centres and the equipment to enable data service provision.

Year	Data centre capacity	Funding for Technical Hub (EBI South Building)	Total funding received
2012	€ 0.3 M	€ 6.7 M	€ 7.0 M
2013	€ 7.3 M	€ 15.1 M	€ 22.4 M
2014	€ 7.4 M	€ 1.7 M	€ 9.1 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 2014 (€58.3 million, excluding sums expended by grant subcontractors).





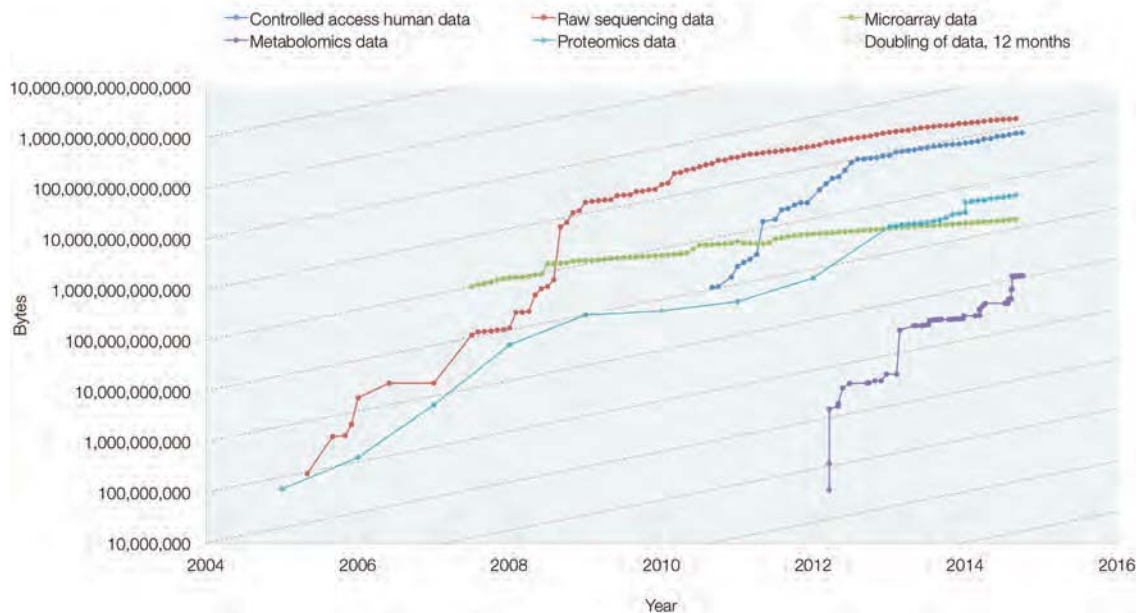
Growth of core resources

At the end of 2014, we saw in excess of 11 million requests to our websites on an average weekday. Compared year-on-year, this figure shows continued growth in the use of our websites. But the absolute number of hits can be influenced heavily by the way the web site is implemented, so we approach these figures with caution.

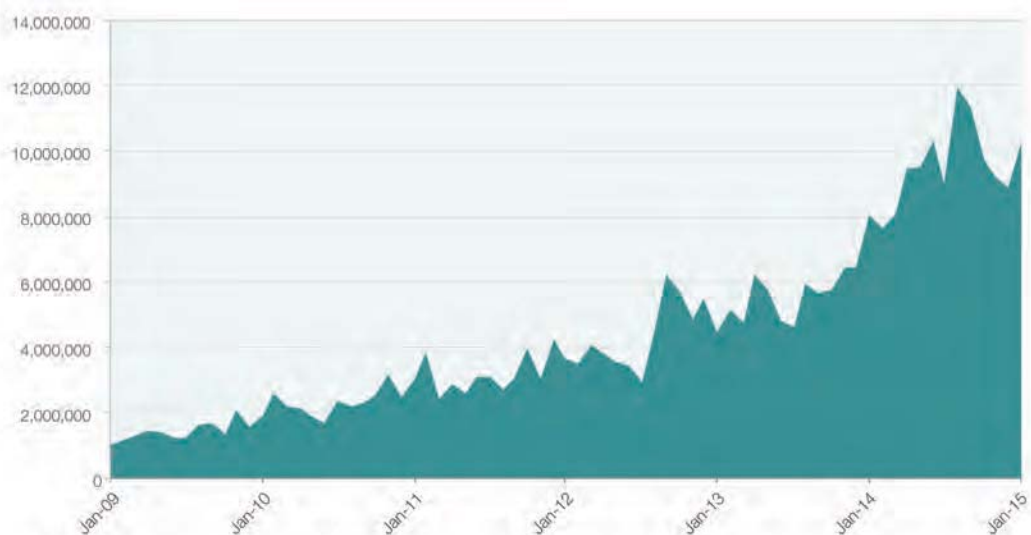
Extracting information on the number of users based on web logs is also very difficult, as it is not uncommon for one IP address to represent a whole organisation. But we continue to see steady growth in usage and in the number of computers accessing our services over time. Our web logs indicate that in 2014, people at 2 million unique IP addresses were using our websites every month.

More than 110 million jobs were run in 2014 using the jdispatcher framework. The average number of jobs per month was 9.2 million (compare to 5.4 million in 2013). The most popular services were InterProScan, Clustal Omega, Pfamscan, and NCBI Blast+.

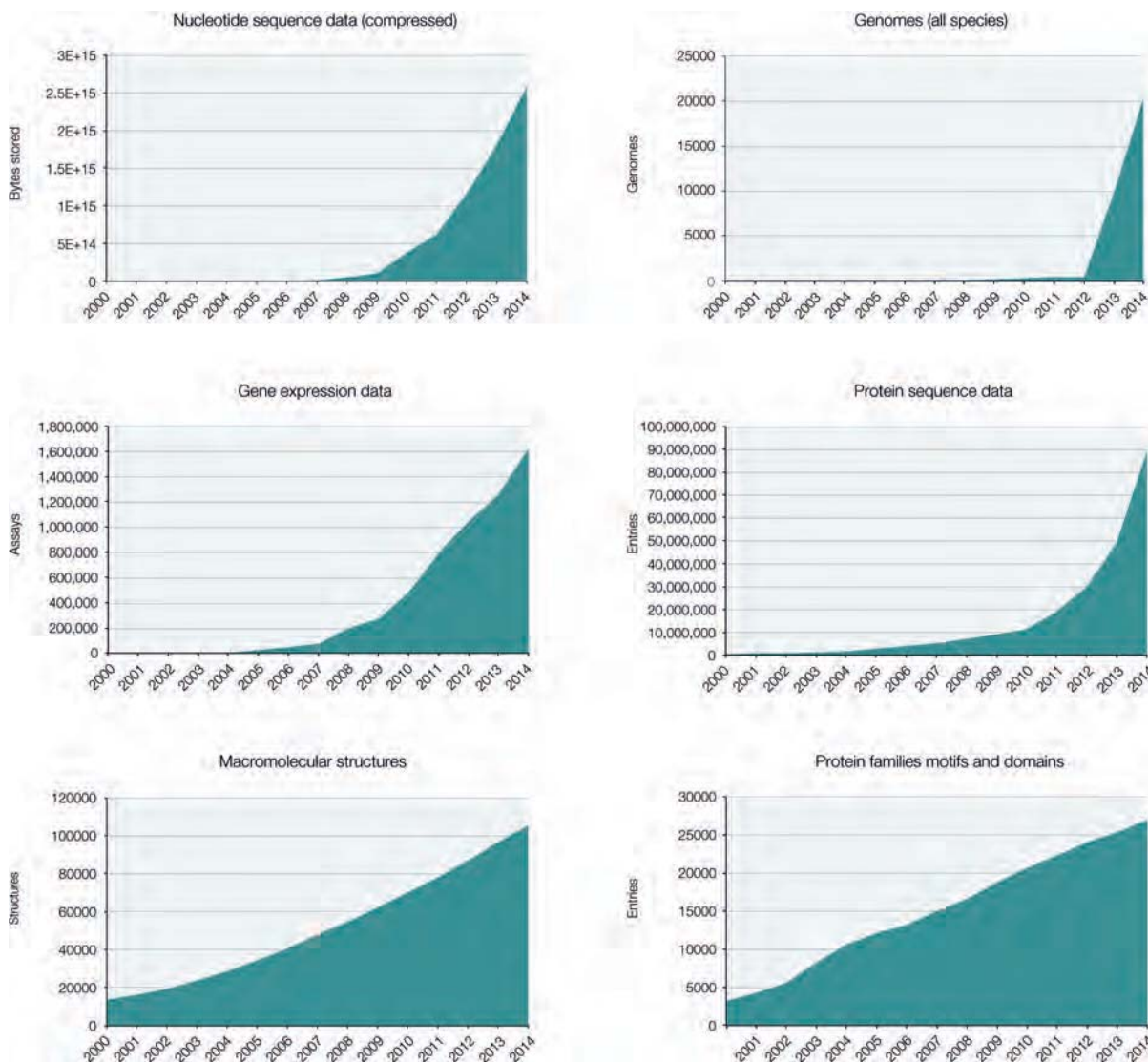
The number of datasets available for searching was over 120 000. This figure is based on the number of class- or taxonomy-specific files being produced to generate Blast and InterProScan services.



Growth of data platforms at EMBL-EBI, 2004 through 2014



Requests per day on EMBL-EBI services, 2009 through 2014



In 2014 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. In particular, 2014 was the first full year during which nucleotide sequence data compression (CRAM) was available and in use by heavy submitters.

- *Compressed nucleotide sequence data: 2.58 petabytes stored (compare to 1.8 petabytes in 2013). A petabyte is 1×10^{15} bytes;*
- *Genomes, all species and strains: 20 343 (compare to 11 010 in 2013);*
- *Gene expression assays: 1.62 million (compare to 1.25 million in 2013);*
- *Protein sequences: 89.1 million (compare to 49.2 million in 2013);*
- *Macromolecular structures: 105 444 (compare to 96 574 in 2013);*
- *Protein families, motifs and domains—entries in InterPro: 27 002 (compare to 25 326 in 2013).*

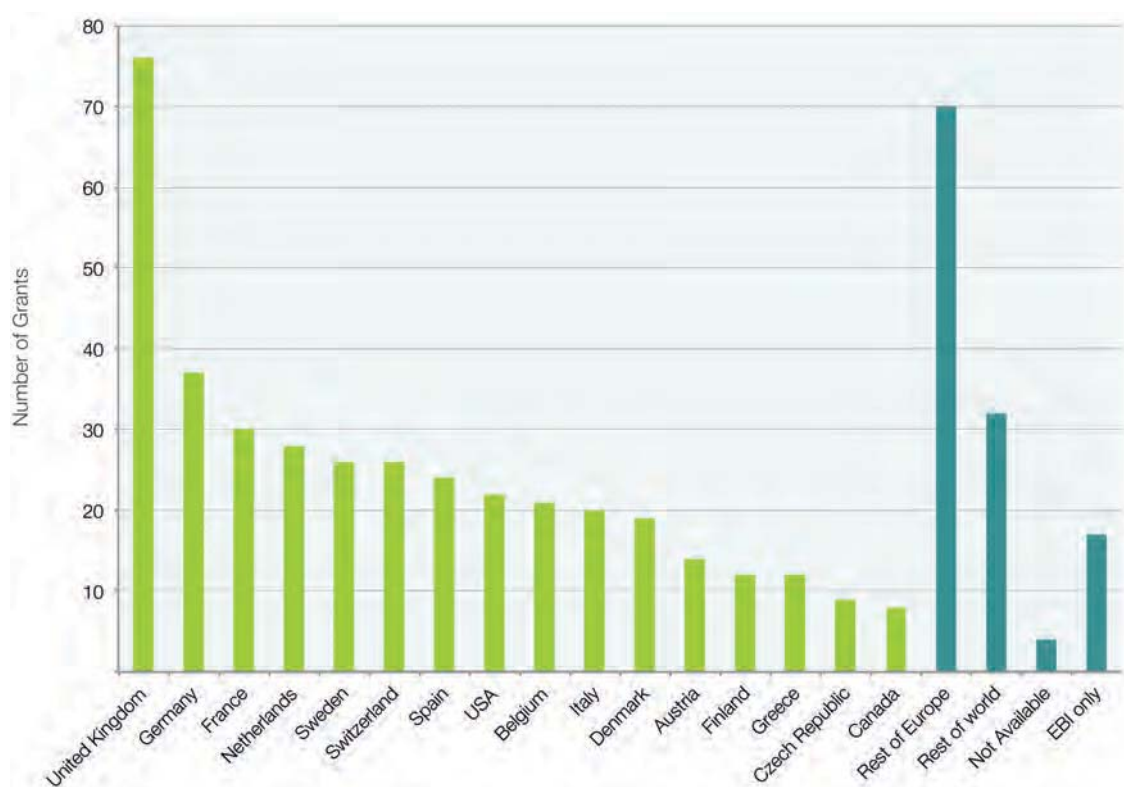


Scientific collaborations

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 enjoys strong, productive partnerships with a large network of academic peers throughout the world.

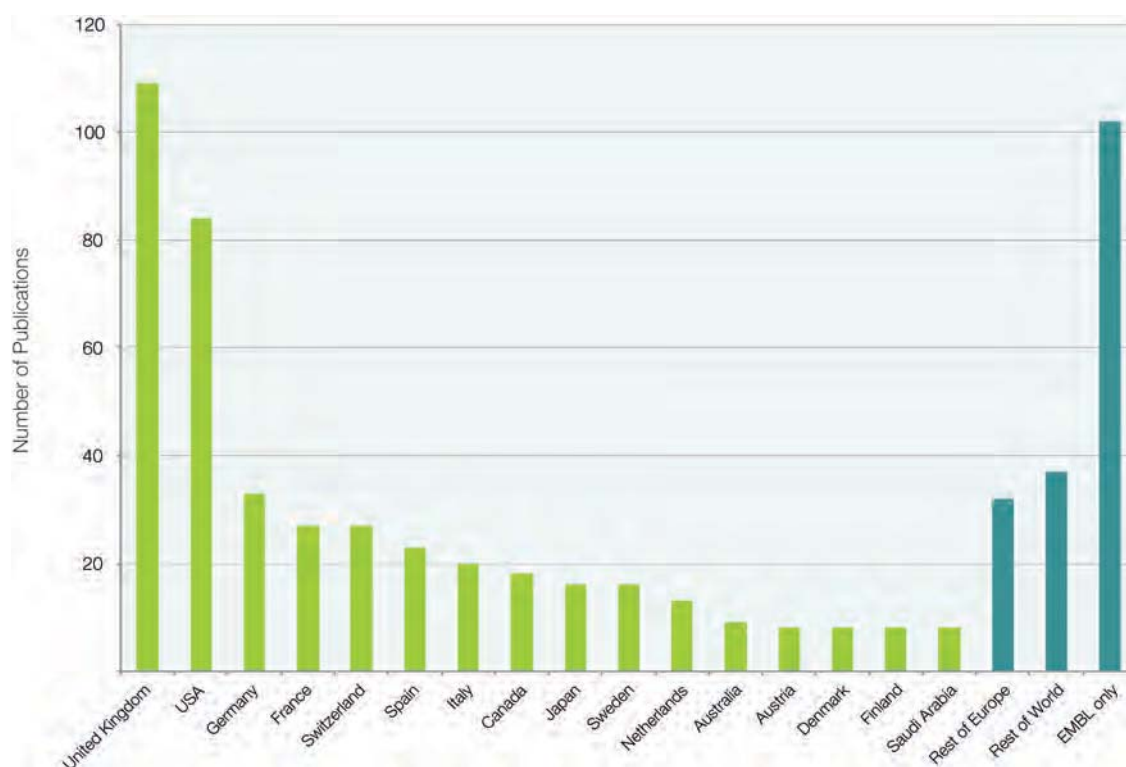
Joint grant funding

In 2014, EMBL-EBI had joint grant funding with researchers and institutes in 65 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 112 grants received, only 17 were exclusively for EMBL. These figures are potentially underestimated, as all partners are not always listed on grants.



Joint publications

Most of our 277 scholarly publications in 2014 were co-authored with colleagues at other institutes throughout the world, including other EMBL outstations. Our most productive partnerships were with people at partner institutes in the United Kingdom, United States, Germany and France, and our collaborations extended well beyond Europe to Malaysia, Mexico and Taiwan.



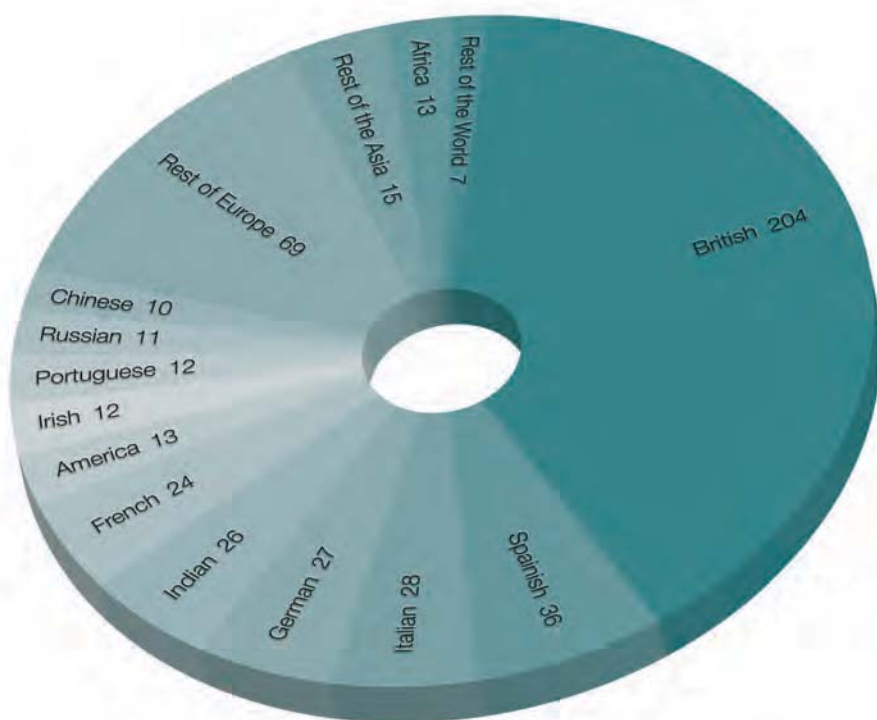


Our staff in 2014

We are proud to report that in 2014, our staff represented 57 nationalities (compare to 55 in 2013). Many of our employees are from the UK, and a substantial number hail from Spain, Italy and Germany.

We had 512 members of staff in 2014, and hosted 91 visitors and 63 trainees. These visitor figures include those who joined us for longer than one month (compare to 76 in 2013)

It is a truly international community comprising people from all over the world, including Jordan, Thailand and Yemen. Our staff in 2014 were American, Australian, Austrian, Belarusian, Brazilian, British, Bulgarian, Cameroonian, Canadian, Chinese, Colombian, Croatian, Cuban, Cypriot, Czech, Danish, Dutch, Egyptian, Estonian, Finnish, French, German, Greek, Hungarian, Indian, Iranian, Irish, Israeli, Italian, Jordanian, Korean, Latvian, Lithuanian, Luxembourgian, Malawian, Malaysian, Mauritanian, New Zealander, Nigerian, Norwegian, Pakistani, Polish, Portuguese, Romanian, Russian, South African, Spanish, Swedish, Swiss, Taiwanese, Thai, Tunisian, Turkish, Ukrainian, Yemeni, Zambian and Zimbabwean.







Scientific Advisory Committees

ArrayExpress and Expression Atlas

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

BioModels Scientific Advisory Committee

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

ChEMBL and ChEBI Advisory Committee

- *Steve Bryant, National Institutes of Health (NIH), United States*
- *Edgar Jacoby, 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*

Ensembl Scientific Advisory Board

- *Anne Ferguson-Smith, Wellcome Trust Senior Investigator, Professor of Genetics in Department of Physiology, Development and Neuroscience, Cambridge University, United Kingdom*
- *Deanna Church, Senior Director of Genomics and Content, Personalis, California, United States*
- *Erich Jarvis, Principal Investigator in Department of*

Neurobiology, Duke University Medical Center, North Carolina, United States

- *Federica Di Palma, Director of Science (Vertebrate and Health Genomics), The Genome Analysis Centre, Norwich, United Kingdom*
- *Felicity Jones, Max Planck Research Group Leader, Friedrich Miescher Laboratory, Tuebingen, Germany*
- *Ian Bird, LHC Computing Project Leader, CERN, Switzerland*
- *Ivo Gut, Director, Centro Nacional de Análisis Genómico, Barcelona, Spain*
- *Jim Reecy, Professor of Animal Science, Iowa State University, Iowa, United States*
- *Mark Diekhans, Technical Project Manager, Center for Biomolecular Science & Engineering, University of California, Santa Cruz, United States*
- *Matt Hurles, Genomic mutation and genetic disease, Wellcome Trust Sanger Institute, Hinxton, United Kingdom*

Ensembl Genomes Scientific Advisory Board

- *Martin Donnelly, University of Liverpool, United Kingdom*
- *Klaus Mayer, Helmholtz Muenchen (HMGU), Germany*
- *Claudine Medigue, Genoscope, France*
- *Allison Miller, University of St. Louis, United States*
- *Rolf Mueller, Helmholtz Institute for Pharmaceutical Research, Saarland, Germany*
- *Chris Rawlings, Rothamsted Research, United Kingdom*
- *Jason Stajich, University of California, Riverside, United States*
- *Denis Tagu, INRA, France*

European Nucleotide Archive Scientific Advisory Board

- *Mark Blaxter, University of Edinburgh, United Kingdom*
- *Antoine Danchin, CNRS, Institut Pasteur, Paris, France*
- *Roderic Guigo, Centre de Regulació Genòmica, Barcelona, Spain*
- *Tim Hubbard, Wellcome Trust Sanger Institute, Hinxton, United Kingdom (Chair)*

- *Jim Ostell, National Centre for Biotechnology Information, United States*
- *Babis Savakis, University of Crete & IMBB-FORTH, Heraklion, Greece*
- *Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin, Germany*
- *Jean Weissenbach, Genoscope, Evry, France*
- *Patrick Wincker, Genoscope, Evry, France*

The Gene Ontology Scientific Advisory Board

- *Philip Bourne, National Institutes of Health, Bethesda, Maryland, United States*
- *Richard Scheuermann, University of Texas Southwestern Medical Center, Dallas, Texas, United States*
- *Michael Schroeder, Technische Universität Dresden, Germany*
- *Barry Smith, SUNY Buffalo, New York, United States*
- *Olga Troyanskaya, Princeton University, Department of Computer Science and Molecular Biology, New Jersey, United States*
- *Michael Tyers, Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Canada*

The International Nucleotide Sequence Database Collaboration (INSDC) International Advisory Committee

- *Antoine Danchin, CNRS, Institut Pasteur, Paris, France*
- *Babis Savakis, University of Crete and IMBB-FORTH, Heraklion, Greece*
- *Jean Weissenbach, Genoscope, Evry, France*

InterPro/Pfam Scientific Advisory Board

- *Philip Bourne, National Institutes of Health, Bethesda, Maryland, United States*
- *Michael Galperin, National Center for Biotechnology Information, Bethesda, Maryland, United States*
- *Erik Sonnhammer, Stockholm University, Sweden (Chair)*
- *Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain*

Literature Services Scientific Advisory Committee

- *Tim Hubbard, King's College London, United Kingdom (Chair)*
- *Larry Hunter, University of Colorado Health Sciences Center, United States*
- *Theo Bloom, British Medical Journal, London, United Kingdom*
- *Terry Attwood, University of Manchester, United Kingdom*
- *Wolfram Horstmann, Göttingen State and University Library, Göttingen, Germany*
- *Martin Fenner, Hannover Medical School Cancer Centre, Germany and Public Library of Science*

Reactome Scientific Advisory Committee

- *Julie Ahringer, University of Cambridge, United Kingdom*
- *Russ Altman, Stanford University, United States*
- *Gary Bader, University of Toronto, Canada*
- *Richard Belew, University of California San Diego, United States*
- *John Overington, EMBL-EBI*
- *Edda Klipp, Max Planck Institute for Molecular Genetics, Germany*
- *Adrian Krainer, Cold Spring Harbor Laboratory, 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, United States*
- *Bill Pearson, University of Virginia, United States*
- *Brian Shoichet, University of California, San Francisco, United States*



Scientific Advisory Committees

Technical Services Cluster Scientific Advisory Committee

- Ewan Birney, Joint Associate Director, EMBL-EBI
- Rolf Apweiler, Joint Associate Director, EMBL-EBI
- Rupert Lueck, Head of IT Services, EMBL, Heidelberg, Germany
- Henning Hermjakob, Proteomics Services Team Leader, EMBL-EBI
- Alvis Brazma, Functional Genomics Senior Team Leader, EMBL-EBI
- Julio Saez-Rodriguez, Research Group Leader, EMBL-EBI
- Tony Cox, Head of Sequencing Informatics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom
- Nick Goldman, Research Group Leader, EMBL-EBI
- Ugis Sarkans, Functional Genomics Technical Team Leader, EMBL-EBI
- Johanna McEntyre, Literature Services Team Leader, EMBL-EBI

Training Programme Scientific Advisory Committee

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

UniProt: The Universal Protein Resource Scientific Advisory Committee

- Patricia Babbitt, University of California, San Francisco, United States
- Helen Berman, Rutgers University, New Jersey, United States
- Judith Blake, The Jackson Laboratory, Maine, 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
- David Searls, Independent Consultant, Philadelphia, United States
- Mathias Uhlén Royal Institute of Technology (KTH), Stockholm, Sweden (Chair)
- Timothy Wells, Medicines for Malaria Venture, Geneva, Switzerland

World-wide Protein Data Bank (wwPDB) Advisory Committee

- Soichi Wakatsuki, Stanford University, United States (Chair)
- Paul Adams, Lawrence Berkeley Laboratory, United States
- Manju Bansal, Indian Institute of Science, India
- 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, United States
- Helen Saibil, Birkbeck College London, United Kingdom
- Edward N. Baker, University of Auckland, New Zealand (Ex Officio)
- R. Andrew Byrd, National Institutes of Health, United States (Ex Officio)

The Protein Data Bank in Europe (PDBe) Scientific Advisory Committee

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

EMDataBank Advisory Committee

- *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, United States*





Major database collaborations

EM Data Bank

- *The National Centre for Macromolecular Imaging (NCMI), Houston, Texas, United States*
- *Protein Data Bank in Europe (PDBe), European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Research Collaboratory for Structural Bioinformatics (RCSB), Piscataway, New Jersey, United States*

Ensembl

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Wellcome Trust Sanger Institute, Hinxton, United Kingdom*

Ensembl Plants

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Cold Spring Harbor Laboratory, New York, United States*

Europe PubMed Central

As part of PubMedCentral International, the United States National Library of Medicine supports Europe PubMed Central and PMC Canada. Europe PubMed Central is developed by:

- *The European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *University of Manchester (Mimas and NaCTeM), United Kingdom*
- *The British Library, London, United Kingdom*

The IMEx Consortium

- *Database of Interacting Proteins (DIP), University of California Los Angeles, United States*
- *I2D, the Interologous Interaction Database, University of Toronto, Canada*
- *InnateDB: Systems Biology of the Innate Human Response, Dublin, Ireland*
- *MatrixDB Extracellular Interactions Database, University of Lyon, France*
- *MBInfo, National University of Singapore*
- *Molecular Connections, Bangalore, India*
- *Molecular Interaction Database (MINT), University Tor Vergata, Rome, Italy*
- *SIB Swiss Institute of Bioinformatics, Switzerland*
- *UniProt, the Universal Protein Resource, EMBL-EBI and SIB Swiss Institute of Bioinformatics*
- *University College London, United Kingdom*

International Nucleotide Sequence Database Collaboration

- *DNA Data Bank of Japan (DDBJ), Mishima, Japan*
- *European Nucleotide Archive (ENA), European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *GenBank, National Center for Biotechnology Information, United States*

ProteomeXchange Consortium

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Faculty of Life Sciences, University of Manchester, United Kingdom*
- *PeptideAtlas, Seattle, United States*
- *Ghent University, Ghent, Belgium*

PhytoPath

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Rothamsted Research, United Kingdom*

PomBase

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *University of Cambridge, United Kingdom*
- *University College London, United Kingdom*

Reactome: the curated human pathways resource

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Ontario Institute of Cancer Research, Toronto, Canada*
- *New York University, United States*

RNAcentral

The RNAcentral Consortium comprises 36 expert databases, 12 of which are integrated into RNAcentral. These are:

- *The European Nucleotide Archive, EMBL-EBI*
- *gtRNAdb: the Genomic tRNA Database, University of California Santa Cruz, United States*
- *lncRNAdb: the Long Noncoding RNA Database, Garvan Institute, Australia*
- *miRBase, University of Manchester, United Kingdom*
- *Protein Data Bank in Europe, EMBL-EBI*
- *RDP: the Ribosomal Database Project, Michigan State University, United States*
- *RefSeq: NCBI Reference Sequence Database, National Center for Biotechnology Information, Bethesda, United States*
- *Rfam, EMBL-EBI*
- *snoRNA Orthological Gene Database, Frontier Science Research Center, University of Miyazaki, Japan*

- *SRPDB: Signal Recognition Particle Database, University of Texas Health Science Center at San Antonio, United States*
- *tmRNA, Sandia National Laboratories, United States*

UniProt

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *SIB Swiss Institute of Bioinformatics, Lausanne and Geneva, Switzerland*
- *Protein Information Resource (PIR), Washington, D.C., United States*

VectorBase

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Harvard University, Cambridge, Massachusetts, United States*
- *Imperial College London, United Kingdom*
- *Institute of Molecular Biology and Biotechnology, Heraklion, Crete, Greece*
- *University of New Mexico, Albuquerque, United States*
- *University of Notre Dame, South Bend, Indiana, United States*

WormBase

- *California Institute of Technology, United States*
- *Ontario Institute for Cancer Research, Toronto, Canada*
- *Wellcome Trust Sanger Institute, Hinxton, United Kingdom*

wwPDB

- *Biological Magnetic Resonance DataBank (BMRB), Madison, Wisconsin, United States*
- *Protein Data Bank in Europe (PDBe), European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *Protein Data Bank of Japan (PDBj), Osaka, Japan*
- *Research Collaboratory for Structural Bioinformatics*



Publications





Publications in 2014

In 2014, we published 277 papers, most of which were co-authored with colleagues at other institutes throughout the world, including other EMBL outstations. Our most productive partnerships were with people at institutes in the United Kingdom, United States, Germany and France, and our collaborations extended well beyond Europe to Malaysia, Mexico and Taiwan.

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. EMBL-EBI rolled out ORCIDs to all staff in 2013, making this the first of our annual publication lists to be generated based on unique IDs.

001. Ahola V, Lehtonen R, Somervuo P, et al. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in *Lepidoptera*. *Nat Commun* 5:4737 doi:10.1038/ncomms5737
002. Alam-Faruque Y, Hill DP, Dimmer EC, et al. (2014) Representing kidney development using the gene ontology. *PLoS ONE* 9(6): doi:10.1371/journal.pone.0099864
003. Almouzni G, Altucci L, Amati B, et al. (2014) Relationship between genome and epigenome - challenges and requirements for future research. *BMC Genomics* 15:487-487. doi:10.1186/1471-2164-15-487
004. Alpi E, Griss J, da Silva AW, et al. (2014) Analysis of the tryptic search space in UniProt databases. *Proteomics* doi:10.1002/pmic.201400227
005. Altenhoff AM, Skunca N, Glover N, et al. (2014) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* 43(D1):D240-D249. doi:10.1093/nar/gku1158
006. Andersson R, Gebhard C, Miguel-Escalada I, et al. (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455-461. doi:10.1038/nature12787
007. Andreeva A, Howorth D, Chothia C, et al. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42(Database issue):D310-4. doi:10.1093/nar/gkt1242
008. Aspesi A, Pavesi E, Robotti E, et al. (2014) Dissecting the transcriptional phenotype of ribosomal protein deficiency: implications for Diamond-Blackfan Anemia. *Gene* 545:282-289. doi:10.1016/j.gene.2014.04.077
009. Audain E, Sanchez A, Vizcaino JA, et al. (2014) A survey of molecular descriptors used in mass spectrometry based proteomics. *Curr Top Med Chem* 14(3):388-397. doi:10.2174/1568026613666131204113537
010. Avery VM, Bashyam S, Burrows JN, et al. (2014) Screening and hit evaluation of a chemical library against blood-stage *Plasmodium falciparum*. *Malar. J.* 13: doi:10.1186/1475-2875-13-190
011. Ayub Q, Moutsianas L, Chen Y, et al. (2014) Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *Am. J. Hum. Genet.* 94(2):176-185. doi:10.1016/j.ajhg.2013.12.010
012. Baker ME, Nelson DR, Studer RA (2014) Origin of the response to adrenal and sex steroids: Roles of promiscuity and co-evolution of enzymes and steroid receptors. *J. Steroid Biochem. Mol. Biol.* doi:10.1016/j.jsbmb.2014.10.020
013. Ballester B, Medina-Rivera A, Schmidt D, et al. (2014) Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife* 3:e02626-e02626. doi:10.7554/elifelife.02626
014. Ballester PJ, Schreyer A, Blundell TL (2014) Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model* 54(3):944-955. doi:10.1021/ci500091r
015. Bansal M, Yang J, Karan C, et al. (2014) A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* : doi:10.1038/nbt.3052
016. Barboza P, Vaillant L, Le Strat Y, et al. (2014) Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS ONE* 9(3):e90536-e90536. doi:10.1371/journal.pone.0090536
017. Barrera A, Alastruey-Izquierdo A, Martin MJ, et al. (2014) Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets. *PLoS Comput. Biol.* 10(7): doi:10.1371/journal.pcbi.1003733
018. Beer R, Herbst K, Ignatiadis N, et al. (2014) Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Mol Biosyst* 10(7):1709-1718. doi:10.1039/c3mb70594c
019. Beerenwinkel N, Schwarz RF, Gerstung M, et al. (2014) Cancer evolution: mathematical models and computational inference. *Syst. Biol.* doi:10.1093/sysbio/syu081
020. Behjati S, Huch M, van Boxtel R, et al. (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 513(7518):422-425. doi:10.1038/nature13448
021. Beisken S, Earll M, Portwood D, et al. (2014) MassCascade: Visual programming for LC-MS data processing in metabolomics. *Mol. Inform.* 33(4):307-310. doi:10.1002/minf.201400016
022. Beisken S, Eiden M, Salek RM (2014) Getting the right answers: understanding metabolomics challenges. *Expert Rev. Mol. Diagn.* 1-13. doi:10.1586/14737159.2015.974562
023. Bento AP, Gaulton A, Hersey A, et al. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42(Database issue):D1083-90. doi:10.1093/nar/gkt1031
024. Berman HM, Burley SK, Kleywegt GJ, et al. (2014) Response to On prompt update of literature references in the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* 70(Pt 10): doi:10.1107/s1399004714020513
025. Berthelot C, Brunet F, Chalopin D, et al. (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 5:3657-3657. doi:10.1038/ncomms4657
026. Billis K, Billini M, Tripp HJ, et al. (2014) Comparative Transcriptomics between *Synechococcus* PCC 7942 and *Synechocystis* PCC 6803 Provide Insights into Mechanisms of Stress Acclimation. *PLoS ONE* 9(10): doi:10.1371/journal.pone.0109738
027. Birney E, Pritchard JK (2014) Archaic humans: Four makes a party. *Nature* 505(7481):32-34. doi:10.1038/nature12847

028. Bolli N, Avet-Loiseau H, Wedge DC, et al. (2014) Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun.* 5:2997-2997. doi:10.1038/ncomms3997
029. Bolli N, Manes N, McKerrel T, et al. (2014) Characterization of gene mutations and copy number changes in acute myeloid leukemia using a rapid target enrichment protocol. *Haematologica* doi:10.3324/haematol.2014.113381
030. Bolser DM, Kerhornou A, Walts B, et al. (2014) Triticeae Resources In *Ensembl Plants*. *Plant Cell Physiol.* 56(1):e3-e3. doi:10.1093/pcp/pcu183
031. Boroviak T, Loos R, Bertone P, et al. (2014) The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat. Cell Biol.* 16(6):516-528. doi:10.1038/ncb2965
032. Brooksbank C, Bergman MT, Apweiler R, et al. (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res.* 42(Database issue):D18-25. doi:10.1093/nar/gkt1206
033. Camps C, Saini HK, Mole DR, et al. (2014) Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *Mol. Cancer* 13:28-28. doi:10.1186/1476-4598-13-28
034. Carbone L, Harris RA, Gnerre S, et al. (2014) Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513(7517):195-201. doi:10.1038/nature13679
035. Carneiro M, Rubin CJ, Di Palma F, et al. (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345(6200):1074-1079. doi:10.1126/science.1253714
036. Carpendale S, Chen M, Evanko D, et al. (2014) Ontologies in biological data visualization. *IEEE Comput. Graph. Appl.* 34(2):8-15. doi:10.1109/mcg.2014.33
037. Casale FP, Giurato G, Nassa G, et al. (2014) Single-cell states in the estrogen response of breast cancer cell lines. *PLoS ONE* 9(2):e88485-e88485. doi:10.1371/journal.pone.0088485
038. Castelnovo M, Zaugg JB, Guffanti E, et al. (2014) Role of histone modifications and early termination in pervasive transcription and antisense-mediated gene silencing in yeast. *Nucleic Acids Res.* 42(7):4348-4362. doi:10.1093/nar/gku100
039. Cejuela JM, McQuilton P, Ponting L, et al. (2014) tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database (Oxford)* 2014:bau033. doi:10.1093/database/bau033
040. Chambers J, Davies M, Gaulton A, et al. (2014) UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. *J. Cheminform.* 6(1): doi:10.1186/s13321-014-0043-5
041. Chelliah V, Juty N, Ajmera I, et al. (2014) BioModels: ten-year anniversary. *Nucleic Acids Res.* doi:10.1093/nar/gku1181
042. Chen L, Kostadima M, Martens JH, et al. (2014) Transcriptional diversity during lineage commitment of human blood progenitors. *Science* 345(6204): doi:10.1126/science.1251033
043. Chibucos MC, Mungall CJ, Balakrishnan R, et al. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO) Database (Oxford) 2014: doi:10.1093/database/bau075
044. Chipman AD, Ferrier DE, Brena C, et al. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. *PLoS Biol.* 12(11): doi:10.1371/journal.pbio.1002005
045. Choi Y, Kim JK, Yoo JY (2014) NF- κ B and STAT3 synergistically activate the expression of FAT10, a gene counteracting the tumor suppressor p53. *Mol. Oncol.* 8(3):642-655. doi:10.1016/j.molonc.2014.01.007
046. Christophorou MA, Castelo-Branco G, Halley-Stott RP, et al. (2014) Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* 507(7490):104-108. doi:10.1038/nature12942
047. Chung CY, Cook CE, Lin GW, et al. (2014) Reliable protocols for whole-mount fluorescent in situ hybridization (FISH) in the pea aphid *Acyrtosiphon pisum*: a comprehensive survey and analysis. *Insect Sci.* 21(3):265-277. doi:10.1111/1744-7917.12086
048. Coimbatore Narayanan B, Westbrook J, Ghosh S, et al. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.* 42(Database issue):D114-22. doi:10.1093/nar/gkt980
049. Colaco CA, Macdougall A (2014) Mycobacterial chaperonins: the tail wags the dog *FEMS Microbiol. Lett.* 350(1):20-24. doi:10.1111/1574-6968.12276
050. Corpas M, Jimenez R, Carbon SJ, et al. (2014) BioJS: an open source standard for biological visualisation - its status in 2014. *F1000Res* 3:55 doi:10.12688/f1000research.3-55.v1
051. Cortes-Ciriano I, van Westen GJ, Lenselink EB, et al. (2014) Proteochemometric modeling in a Bayesian framework. *J. Cheminform* 6:35 doi:10.1186/1758-2946-6-35
052. Cossetti C, Iraci N, Mercer TR, et al. (2014) Extracellular vesicles from neural stem cells transfer IFN- γ via ifngr1 to activate stat1 signaling in target cells. *Mol. Cell* 56(2):193-204. doi:10.1016/j.molcel.2014.08.020
053. Costello JC, Heiser LM, Georgii E, et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* doi:10.1038/nbt.2877
054. Croft D, Mundo AF, Haw R, et al. (2014) The Reactome pathway knowledgebase *Nucleic Acids Res.* 42(Database issue):D472-7. doi:10.1093/nar/gkt1102
055. Croset S, Overington JP, Rebholz-Schuhmann D (2014) The functional therapeutic chemical classification system. *Bioinformatics* 30(6):876-883. doi:10.1093/bioinformatics/btt628
056. Cubillos FA, Stegle O, Grondin C, et al. (2014) Extensive cis-regulatory variation robust to environmental perturbation in *Arabidopsis*. *Plant Cell* 26(11):4298-4310. doi:10.1105/tpc.114.130310
057. Cunningham F, Amode MR, Barrell D, et al. (2014) Ensembl 2015. *Nucleic Acids Res.* doi:10.1093/nar/gku1010
058. Das D, Murzin AG, Rawlings ND, et al. (2014) Structure and computational analysis of a novel protein with metalloproteinase-like and circularly permuted winged-helix-turn-helix domains reveals a possible role in modified polysaccharide biosynthesis. *BMC Bioinformatics* 15: doi:10.1186/1471-2105-15-75
059. Das D, Murzin AG, Rawlings ND, et al. (2014) Structure and computational analysis of a novel protein with metalloproteinase-like and circularly permuted winged-helix-turn-helix domains reveals a possible role in modified polysaccharide biosynthesis. *BMC Bioinformatics* 15:75-75. doi:10.1186/1471-2105-15-75
060. Deligianni E, Dialynas E, et al. (2014) Non-coding RNA gene families in the genomes of anopheline mosquitoes. *BMC Genomics* 15: doi:10.1186/1471-2164-15-1038
061. Denise H, Moschos SA, Sidders B, et al. (2014) Deep sequencing insights in therapeutic shRNA processing and siRNA target cleavage precision. *Mol. Ther. Nucleic Acids* 3:e145-e145. doi:10.1038/mtna.2013.73
062. Dhifli W, Saidi R, Nguifo EM (2014) Smoothing 3D protein structure motifs through graph mining and amino acid similarities *J. Comput. Biol.* 21(2):162-172. doi:10.1089/cmb.2013.0092
063. Di Camillo B, Eduati F, Nair SK, et al. (2014) Leucine modulates dynamic phosphorylation events in insulin signaling pathway and enhances insulin-dependent glycogen synthesis in human skeletal muscle cells. *BMC Cell Biol.* 15: doi:10.1186/1471-2121-15-9
064. Diawara MR, Hue C, Wilder SP, et al. (2014) Adaptive expression of microRNA-125a in adipose tissue in response to obesity in mice and men. *PLoS ONE* 9(3): doi:10.1371/journal.pone.0091375
065. Dikicioglu D, Wood V, Rutherford KM, et al. (2014) Improving functional annotation for industrial microbes: a case study with *Pichia pastoris*. *Trends Biotechnol.* 32(8):396-399. doi:10.1016/j.tibtech.2014.05.003



Publications in 2014

066. Ding Z, Ni Y, Timmer SW, et al. (2014) Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. *PLoS Genet.* 10(11): doi:10.1371/journal.pgen.1004798
067. Dossetter AG, Ecker G, Lavery H, et al. (2014) 'Big data' in pharmaceutical science: challenges and opportunities. *Future Med Chem* 6(8):857-864. doi:10.4155/fmc.14.45
068. Dutta S, Dimitropoulos D, Feng Z, et al. (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers* 101(6):659-668. doi:10.1002/bip.22434
069. Earl D, Nguyen NK, Hickey G, et al. (2014) Alignathon: A competitive assessment of whole genome alignment methods. *Genome Res.* doi:10.1101/gr.174920.114
070. Eckersley-Maslin MA, Thybert D, Bergmann JH, et al. (2014) Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* 28(4):351-365. doi:10.1016/j.devcel.2014.01.017
071. Egea JA, Henriques D, Cokelaer T, et al. (2014) MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinformatics* 15:136-136. doi:10.1186/1471-2105-15-136
072. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest AR, Kawaji H, et al. (2014) A promoter-level mammalian expression atlas. *Nature* 507(7493):462-470. doi:10.1038/nature13182
073. Faisal A, Peltonen J, Georgii E, et al. (2014) Toward computational cumulative biology by combining models of biological datasets. *PLoS ONE* 9(11): doi:10.1371/journal.pone.0113053
074. Famiglietti ML, Streicher A, Gos A, et al. (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.* 35(8):927-935. doi:10.1002/humu.22594
075. Farrah T, Binz PA, et al. (2014) A standardized framing for reporting protein identifications in mzIdentML 1.2. *Proteomics* 14:2389-2399. doi:10.1002/pmic.201400080
076. Faulconbridge A, Burdett T, Brandizi M, et al. (2014) Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res.* 42(Database issue):D50-2. doi:10.1093/nar/gkt1081
077. Federhen S, Clark K, Barrett T, et al. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand Genomic Sci* 9(3):1275-1277. doi:10.4056/sigs.4851102
078. Feng S, Laketa V, Stein F, et al. (2014) A rapidly reversible chemical dimerizer system to study lipid signaling in living cells. *Angew. Chem. Int. Ed. Engl.* 53(26):6720-6723. doi:10.1002/anie.201402294
079. Finn RD, Bateman A, Clements J, et al. (2014) Pfam: the protein families database. *Nucleic Acids Res.* 42(Database issue):D222-30. doi:10.1093/nar/gkt1223
080. Finn RD, Miller BL, Clements J, et al. (2014) iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res.* 42(Database issue):D364-73. doi:10.1093/nar/gkt1210
081. Flicek P, Amodé M, Barrell D, et al. (2014) Ensembl 2014. *Nucleic Acids Res.* 42(Database issue):D749-55. doi:10.1093/nar/gkt1196
082. Fonseca NA, Marioni J, Brazma A (2014) RNA-Seq gene profiling - a systematic empirical comparison. *PLoS ONE* 9(9): doi:10.1371/journal.pone.0107026
083. Fox SE, Geniza M, Hanumappa M, et al. (2014) De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat *Triticum monococcum*. *PLoS ONE* 9(5):e96855-e96855. doi:10.1371/journal.pone.0096855
084. Franklin CS, Floyd JA, et al. (2014) A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* 19(10):1085-1094. doi:10.1038/mp.2013.187
085. Furnham N, Holliday GL, de Beer TA, et al. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* 42(Database issue):D485-9. doi:10.1093/nar/gkt1243
086. Fusi N, Lippert C, Lawrence ND, et al. (2014) Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat Commun.* 5:4890 doi:10.1038/ncomms5890
087. Gago S, Alastruey-Izquierdo A, Marconi M, et al. (2014) Ribosomal DNA intergenic spacer 1 region is useful when identifying *Candida parapsilosis* spp. complex based on high-resolution melting analysis. *Med. Mycol.* 52(5):472-481. doi:10.1093/mmy/myu009
088. Galardini M, Mengoni A, Biondi EG, et al. (2014) DuctApe: a suite for the analysis and correlation of genomic and OmniLog™ Phenotype Microarray data. *Genomics* 103(1):1-10. doi:10.1016/j.ygeno.2013.11.005
089. Garcia L, Yachdav G, Martin MJ (2014) FeatureViewer, a BioJS component for visualization of position-based annotations in protein sequences. *F1000Res* 3:47-47. doi:10.12688/f1000research.3-47.v2
090. Gerstein MB, Rozowsky J, Yan KK, et al. (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512(7515):445-448. doi:10.1038/nature13424
091. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, et al. (2014) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* doi:10.1093/nar/gku1117
092. Gobbi A, Iorio F, Dawson KJ, et al. (2014) Fast randomization of large genomic datasets while preserving alteration counts. *Bioinformatics* 30(17):i617-i623. doi:10.1093/bioinformatics/btu474
093. Gomez J, Jimenez R (2014) Sequence, a BioJS component for visualising sequences. *F1000Res* 3:52 doi:10.12688/f1000research.3-52.v1
094. Gray KA, Yates B, Seal RL, et al. (2014) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* doi:10.1093/nar/gku1071
095. Greger L, Su J, Rung J, et al. (2014) Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. *PLoS ONE* 9(8):e104567 doi:10.1371/journal.pone.0104567
096. Grison A, Zucchelli S, Urzi A, et al. (2014) Mesencephalic dopaminergic neurons express a repertoire of olfactory receptors and respond to odorant-like molecules. *BMC Genomics* 15: doi:10.1186/1471-2164-15-729
097. Griss J, Jones AR, Sachsenberg T, et al. (2014) The mzTab Data Exchange Format: communicating MS-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell Proteomics* doi:10.1074/mcp.O113.036681
098. Griss J, Perez-Riverol Y, Hermjakob H, et al. (2014) Identifying novel biomarkers through data mining - a realistic scenario? *Proteomics Clin. Appl.* doi:10.1002/prca.201400107
099. Gurdasani D, Carstensen T, Tekola-Ayele F, et al. (2014) The African Genome Variation Project shapes medical genetics in Africa. *Nature* doi:10.1038/nature13997
100. Gutmanas A, Alhroub Y, Battle GM, et al. (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 42(Database issue):D285-91. doi:10.1093/nar/gkt1180
101. Gymrek M, Highnam G, et al. (2014) The landscape of human STR variation. *Genome Res.* 24(11):1894-1904. doi:10.1101/gr.177774.114
102. Gómez J, Brezmes J, Mallol R, et al. (2014) Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. *Anal. Bioanal. Chem.* 406(30):7967-7976. doi:10.1007/s00216-014-8225-6

103. Harris TW, Baran J, Bieri T, et al. (2014) WormBase 2014: new views of curated biology *Nucleic Acids Res.* 42(Database issue):D789-93. doi:10.1093/nar/gkt1063
104. Harrison PW, Jordan GE, Montgomery SH (2014) SWAMP: Sliding Window Alignment Masker for PAML. *Evol. Bioinform. Online* 10:197-204. doi:10.4137/ebo.s18193
105. Hastings J, Frishkoff GA, Smith B, et al. (2014) Interdisciplinary perspectives on the development, integration, and application of cognitive ontologies. *Front. Neuroinform.* 8: doi:10.3389/fninf.2014.00062
106. Hastings J, Haug K, Steinbeck C (2014) Ten recommendations for software engineering in research. *Gigascience* 3(1): doi:10.1186/2047-217x-3-31
107. Hendrickx DM, Aerts HJ, Caiment F, et al. (2014) diXa: a Data Infrastructure for Chemical Safety Assessment. *Bioinformatics* doi:10.1093/bioinformatics/btu827
108. Hiroi N, Swat M, Funahashi A (2014) Assessing uncertainty in model parameters based on sparse and noisy experimental data. *Front. Physiol.* 5:128-128. doi:10.3389/fphys.2014.00128
109. Hoeger B, Diether M, Ballester PJ, et al. (2014) Biochemical evaluation of virtual screening methods reveals a cell-active inhibitor of the cancer-promoting phosphatases of regenerating liver. *Eur J Med Chem.* doi:10.1016/j.ejmech.2014.08.060
110. Holliday GL, Rahman SA, Furnham N, et al. (2014) Exploring the biological and chemical complexity of the ligases. *J. Mol. Biol.* 426(10):2098-2111. doi:10.1016/j.jmb.2014.03.008
111. Hunter S, Corbett M, Denise H, et al. (2014) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* 42(Database issue):D600-6. doi:10.1093/nar/gkt961
112. Huntley RP, Harris MA, Alam-Faruque Y, et al. (2014) A method for increasing expressivity of Gene Ontology annotations using a compositional approach. *BMC Bioinformatics* 15:155-155. doi:10.1186/1471-2105-15-155
113. Huntley RP, Sawford T, Martin MJ, et al. (2014) Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* 3(1):4-4. doi:10.1186/2047-217x-3-4
114. Huntley RP, Sawford T, Mutowo-Muullenet P, et al. (2014) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.* doi:10.1093/nar/gku1113
115. Iantorno S, Gori K, Goldman N, et al. (2014) Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol. Biol.* 1079:59-73. doi:10.1007/978-1-62703-646-7_4
116. Ibarra-Soria X, Levitin MO, Saraiva LR, et al. (2014) The olfactory transcriptomes of mice. *PLoS Genet.* 10(9): doi:10.1371/journal.pgen.1004593
117. International Glossina Genome Initiative (2014) Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. *Science* 344(6182):380-386. doi:10.1126/science.1249656
118. Jaworski J, Govender U, McFarlane C, et al. (2014) A novel RCE1 isoform is required for H-Ras plasma membrane localization and is regulated by USP17. *Biochem. J.* 457(2):289-300. doi:10.1042/bj20131213
119. Jayaseelan KV, Steinbeck C (2014) Building blocks for automated elucidation of metabolites: natural product-likeness for candidate ranking. *BMC Bioinformatics* 15: doi:10.1186/1471-2105-15-234
120. Jiang X, Peery A, Hall AB, et al. (2014) Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. *Genome Biol.* 15(9): doi:10.1186/s13059-014-0459-2
121. Jiang Y, Xie M, Chen W, et al. (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344(6188):1168-1173. doi:10.1126/science.1252806
122. Jones P, Binns D, Chang HY, et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240. doi:10.1093/bioinformatics/btu031
123. Jupe S, Jassal B, Williams M, et al. (2014) A controlled vocabulary for pathway entities and events. *Database (Oxford)* 2014: doi:10.1093/database/bau060
124. Jupp S, Malone J, Bolleman J, et al. (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 30(9):1338-1339. doi:10.1093/bioinformatics/btt765
125. Keildson S, Fadista J, Laderwall C, et al. (2014) Expression of phosphofructokinase in skeletal muscle is influenced by genetic variation and associated with insulin sensitivity. *Diabetes* 63(3):1154-1165. doi:10.2337/db13-1301
126. Kellis M, Wold B, Snyder MP, et al. (2014) Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 111(17):6131-6138. doi:10.1073/pnas.1318948111
127. Kellis M, Wold B, Snyder MP, et al. (2014) Reply to Brunet and Doolittle: Both selected effect and causal role elements can influence human biology and disease. *Proc. Natl. Acad. Sci. U.S.A.* 111(33): doi:10.1073/pnas.1410434111
128. Kersey PJ, Allen JE, Christensen M, et al. (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* 42(Database issue):D546-52. doi:10.1093/nar/gkt979
129. Kibbe WA, Arze C, Felix V, et al. (2014) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* 43(Database issue):D1071-D1078. doi:10.1093/nar/gku1011
130. Kolesnikov N, Hastings E, Keays M, et al. (2014) ArrayExpress update-simplifying data submissions. *Nucleic Acids Res.* 43(D1):D1113-D1116. doi:10.1093/nar/gku1057
131. Koscielny G, Yaikhom G, Iyer V, et al. (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* 42(Database issue):D802-9. doi:10.1093/nar/gkt977
132. Kotelnikova E, Bernardo-Faura M, Silberberg G, et al. (2014) Signaling networks in MS: A systems-based approach to developing new pharmacological therapies. *Mult. Scler.* doi:10.1177/1352458514543339
133. Kruger FA, Gaulton A, Nowotka M, et al. (2014) PPDMs - A resource for mapping small molecule bioactivities from ChEMBL to Pfam-A protein domains. *Bioinformatics* doi:10.1093/bioinformatics/btu711
134. Kultys M, Nicholas L, Schwarz R, et al. (2014) Sequence Bundles: a novel method for visualising, discovering and exploring sequence motifs. *BMC Proceedings* 8(Suppl 2):S8-S8. doi:10.1186/1753-6561-8-S2-S8
135. Laketa V, Zerbakhsh S, Traynor-Kaplan A, et al. (2014) PIP β induces the recycling of receptor tyrosine kinases. *Sci. Signal.* 7(308):ra5-ra5. doi:10.1126/scisignal.2004532
136. Lebedev AV, Westman E, Van Westen GJ, et al. (2014) Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin.* 6:115-125. doi:10.1016/j.nicl.2014.08.023
137. Leprevost Fda V, Barbosa VC, Francisco EL, et al. (2014) On best practices in the development of bioinformatics software. *Front. Genet.* 5:199 doi:10.3389/fgene.2014.00199
138. Lewis TE, Sillitoe I, Andreeva A, et al. (2014) Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Res.* 43(D1):D382-D386. doi:10.1093/nar/gku973
139. Li H, Leung KS, Ballester PJ, et al. (2014) istar: a web platform for large-scale protein-ligand docking. *PLoS ONE* 9(1):e85678-85678. doi:10.1371/journal.pone.0085678



Publications in 2014

140. Lin GW, Cook CE, Miura T, et al. (2014) Posterior localization of ApVas1 positions the preformed germ plasm in the sexual oviparous pea aphid *Acyrtosiphon pisum*. *Evodevo* 5:18-18. doi:10.1186/2041-9139-5-18
141. Lopez R, Cowley A, Li W, et al. (2014) Using EMBL-EBI Services via Web Interface and Programmatically via Web Services. *Curr. Protoc. Bioinformatics* 48:3.12.1-3.12.50. doi:10.1002/0471250953.bi0312s48
142. Lotz C, Lin AJ, Black CM, et al. (2014) Characterization, design, and function of the mitochondrial proteome: from organs to organisms. *J. Proteome Res.* 13(2):433-446. doi:10.1021/pr400539j
143. Louis A, Nguyen NT, Muffato M, et al. (2014) Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Res.* doi:10.1093/nar/gku1112
144. Lucas JM, Muffato M, Roest Crollius H (2014) PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 15(1): doi:10.1186/1471-2105-15-268
145. López-Sánchez I, Valbuena A, Vázquez-Cedeira M, et al. (2014) VRK1 interacts with p53 forming a basal complex that is activated by UV-induced DNA damage. *FEBS Lett.* 588(5):692-700. doi:10.1016/j.febslet.2014.01.040
146. M, Milacic M, et al. (2014) Reactome Knowledgebase: a platform for pathway and network analysis. *Drug Metab. Rev.* 45:45-46
147. Macarthur JA, Morales J, Tully RE, et al. (2014) Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.* 42(Database issue):D873-8. doi:10.1093/nar/gkt1198
148. Mahata B, Zhang X, Kolodziejczyk AA, et al. (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep.* 7(4):1130-1142. doi:10.1016/j.celrep.2014.04.011
149. Malone J, Brown A, Lister AL, et al. (2014) The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. *J. Biomed. Semantics* 5:25 doi:10.1186/2041-1480-5-25
150. Marchini J, 1000 Genomes Project Consortium, et al. (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun.* 5: doi:10.1038/ncomms4934
151. Marguerat S, Lawler K, Brazma A, et al. (2014) Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress. *RNA Biol.* 11(6):702-714. doi:10.4161/ma.29196
152. Marmoset Genome Sequencing and Analysis Consortium (2014) The common marmoset genome provides insight into primate biology and evolution. *Nat. Genet.* 46(8):850-857. doi:10.1038/ng.3042
153. Marsh JA, Teichmann SA (2014) Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *Bioessays* 36(2):209-218. doi:10.1002/bies.201300134
154. Marsh JA, Teichmann SA (2014) Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol.* 12(5):e1001870-e1001870. doi:10.1371/journal.pbio.1001870
155. Marti-Solano M, Birney E, Bril A, et al. (2014) Integrative knowledge management to enhance pharmaceutical R&D. *Nat. Rev. Drug Discov.* 13(4):239-240. doi:10.1038/nrd4290
156. Martinez Cuesta S, Furnham N, Rahman SA, et al. (2014) The evolution of enzyme function in the isomerases. *Curr. Opin. Struct. Biol.* 26C:121-130. doi:10.1016/j.sbi.2014.06.002
157. Masson P, Hulo C, de Castro E, et al. (2014) An integrated ontology resource to explore and study host-virus relationships. *PLoS ONE* 9(9):e108075 doi:10.1371/journal.pone.0108075
158. Mateo JL, van den Berg DL, Haeussler M, et al. (2014) Characterization of the neural stem cell gene regulatory network identifies OLIG2 as a multi-functional regulator of self-renewal. *Genome Res.* doi:10.1101/gr.173.435.114
159. May JW, Steinbeck C (2014) Efficient ring perception for the Chemistry Development Kit. *J. Cheminform.* 6(1):3-3. doi:10.1186/1758-2946-6-3
160. Mayer G, Jones AR, Binz PA, et al. (2014) Controlled vocabularies and ontologies in proteomics: overview, principles and practice *Biochim. Biophys. Acta* 1844(1 Pt A):98-107. doi:10.1016/j.bbapap.2013.02.017
161. McDowall MD, Harris MA, Lock A, et al. (2014) PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res.* doi:10.1093/nar/gku1040
162. McGaugh SE, Gross JB, Aken B, et al. (2014) The cavefish genome reveals candidate genes for eye loss. *Nat Commun.* 5:5307-5307. doi:10.1038/ncomms6307
163. McInerney GJ, Chen M, Freeman R, et al. (2014) Information visualisation for science and policy: engaging users and avoiding bias. *Trends Ecol. Evol. (Amst.)* 29(3):148-157. doi:10.1016/j.tree.2014.01.003
164. Meldal BH, Forner-Martinez O, Costanzo MC, et al. (2014) The complex portal - an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.* doi:10.1093/nar/gku975
165. Meyer P, Cokelaer T, Chandran D, et al. (2014) Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC Syst Biol* 8(1):13-13. doi:10.1186/1752-0509-8-13
166. Mitchell A, Chang HY, Daugherty L, et al. (2014) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* doi:10.1093/nar/gku1243
167. Molina-Morales M, Martínez JG, Martín-Gálvez D, et al. (2014) Cuckoo hosts shift from accepting to rejecting parasitic eggs across their lifetime. *Evolution* 68(10):3020-3029. doi:10.1111/evo.12471
168. Monaco MK, Stein J, Naithani S, et al. (2014) Gramene 2013: comparative plant genomics resources *Nucleic Acids Res.* 42(Database issue):D1193-9. doi:10.1093/nar/gkt1110
169. Morgat A, Axelsen KB, Lombardot T, et al. (2014) Updates in Rhea-a manually curated resource of biochemical reactions. *Nucleic Acids Res.* doi:10.1093/nar/gku961
170. Murray MJ, Bailey S, Raby KL, et al. (2014) Serum levels of mature microRNAs in DICER1-mutated pleuropulmonary blastoma. *Oncogenesis* 3:e87-e87. doi:10.1038/ncsis.2014.1
- Murray MJ, Raby KL, Saini HK, et al. (2014) Solid tumors of childhood display specific serum microRNA profiles. *Cancer Epidemiol. Biomarkers Prev.* doi:10.1158/1055-9965.epi-14-0669
171. Muthukrishnan V, May JW, et al. (2014) Pipeline for automated structure-based classification in the ChEBI ontology. *Abstr. Pap. Am. Chem. S* 247:
172. Nawrocki EP, Burge SW, Bateman A, et al. (2014) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43(D1):D130-D137. doi:10.1093/nar/gku1063
173. Nayduch D, Cohnstaedt LW, Saski C, et al. (2014) Studying *Culicoides* vectors of BTV in the post-genomic era: resources, bottlenecks to progress and future directions. *Virus Res.* 182:43-49. doi:10.1016/j.virusres.2013.12.009
174. Ochoa D, Pazos F (2014) Practical aspects of protein co-evolution. *Front. Cell Dev. Biol.* 2: doi:10.3389/fcell.2014.00014

175. Ochoa R, Davies M, Papadatos G, et al. (2014) myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics* 30(2):298-300. doi:10.1093/bioinformatics/btt666
176. Orchard S (2014) Data standardization and sharing-the work of the HUPO-PSI. *Biochim. Biophys. Acta* 1844(1 Pt A):82-87. doi:10.1016/j.bbapap.2013.03.011
177. Orchard S, Albar JP, Binz PA, et al. (2014) Meeting New Challenges: The 2014 HUPO-PSI/COSMOS Workshop: 13-15 April 2014, Frankfurt, Germany. *Proteomics* 14(21-22):2363-2368. doi:10.1002/pmic.201470164
178. Orchard S, Ammari M, Aranda B, et al. (2014) The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42(Database issue):D358-63. doi:10.1093/nar/gkt1115
179. Ottolini B, Hornsby MJ, Abujaber R, et al. (2014) Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the β -defensin-2 gene. *Genome Biol. Evol.* 6(11):3025-3038. doi:10.1093/gbe/evu236
180. Pakseresht N, Alako B, Amid C, et al. (2014) Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res.* 42(Database issue):D38-43. doi:10.1093/nar/gkt1082
181. Pande S, Merker H, Bohl K, et al. (2014) Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J.* 8(5):953-962. doi:10.1038/ismej.2013.211
182. Panoutsopoulou K, Hatzikotoulas K, Xifara DK, et al. (2014) Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun.* 5: doi:10.1038/ncomms6345
183. Papadatos G, Overington JP (2014) The ChEMBL database: a taster for medicinal chemists. *Future Med. Chem.* 6(4):361-364. doi:10.4155/fmc.14.8
184. Papadatos G, van Westen GJ, Croset S, et al. (2014) A document classifier for medicinal chemistry publications trained on the ChEMBL corpus. *J. Cheminform.* 6(1):40 doi:10.1186/s13321-014-0040-8
185. Papatheodorou I, Petrovs R, Thornton JM (2014) Comparison of the mammalian insulin signalling pathway to invertebrates in the context of FOXO-mediated ageing. *Bioinformatics* 30(21):2999-3003. doi:10.1093/bioinformatics/btu493
186. Parks SL, Goldman N (2014) Maximum likelihood inference of small trees in the presence of long branches. *Syst. Biol.* 63(5):798-811. doi:10.1093/sysbio/syu044
187. Patwardhan A, Ashton A, Brandt R, et al. (2014) A 3D cellular context for the macromolecular world. *Nat. Struct. Mol. Biol.* 21(10):841-845. doi:10.1038/nsmb.2897
188. Pavelin K, Pundir S, Cham JA (2014) Ten simple rules for running interactive workshops. *PLoS Comput. Biol.* 10(2):e1003485-e1003485. doi:10.1371/journal.pcbi.1003485
189. Peat JR, Dean W, Clark SJ, et al. (2014) Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. *Cell Rep.* 9(6):1990-2000. doi:10.1016/j.celrep.2014.11.034
190. Peng X, Alföldi J, Gori K, et al. (2014) The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. *Nat. Biotechnol.* doi:10.1038/nbt.3079
191. Perez-Riverol Y, Alpi E, Wang R, et al. (2014) Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics* doi:10.1002/pmic.201400302
192. Perez-Riverol Y, Carvalho PC (2014) Editorial: Genomics and proteomics behind drug design. *Curr. Top. Med. Chem.* 14(3):343-343. doi:10.2174/1568026613666131204101110
193. Perez-Riverol Y, Wang R, Hermjakob H, et al. (2014) Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. *Biochim. Biophys. Acta* 1844(1 Pt A):63-76. doi:10.1016/j.bbapap.2013.02.032
194. Perica T, Kondo Y, Tiwari SP, et al. (2014) Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* 346(6216): doi:10.1126/science.1254346
195. Petryszak R, Burdett T, Fiorelli B, et al. (2014) Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 42(Database issue):D926-32. doi:10.1093/nar/gkt1270
196. Pettit JB, Tomer R, Achim K, et al. (2014) Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput. Biol.* 10(9): doi:10.1371/journal.pcbi.1003824
197. Piccirillo SG, Spiteri I, Sottoriva A, et al. (2014) Contributions to drug resistance in glioblastoma derived from malignant cells in the sub-ependymal zone. *Cancer Res.* 75(194): doi:10.1158/0008-5472.CAN-13-3131
198. Poultney CS, Samocha K, Kou Y, et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515(7526):209-215. doi:10.1038/nature13772
199. 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. doi:10.1093/database/bau016
200. Purwantini E, Torto-Alalibo T, Lomax J, et al. (2014) Genetic resources for methane production from biomass described with the Gene Ontology. *Front. Microbiol.* 5: doi:10.3389/fmicb.2014.00634
201. Rahman SA, Cuesta SM, Furnham N, et al. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods* 11(2):171-174. doi:10.1038/nmeth.2803
202. Rahman SA, Singh Y, Kohli S, et al. (2014) Comparative analyses of nonpathogenic, opportunistic, and totally pathogenic mycobacteria reveal genomic and biochemical variabilities and highlight the survival attributes of *Mycobacterium tuberculosis*. *MBio* 5(6):e02020-e02020. doi:10.1128/mbio.02020-14
203. Ramos EM, Din-Lovinescu C, Berg JS, et al. (2014) Characterizing genetic variants for clinical action. *Am. J. Med. Genet. C Semin. Med. Genet.* 166C(1):93-104. doi:10.1002/ajmg.c.31386
204. Rasmussen MD, Hubisz MJ, Gronau I, et al. (2014) Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* 10(5): doi:10.1371/journal.pgen.1004342
205. Rawlings ND, Barrett AJ, Bateman A (2014) Using the MEROPS database for proteolytic enzymes and their inhibitors and substrates. *Curr. Protoc. Bioinformatics* 48:1.25.1-1.25.33. doi:10.1002/0471250953.bi0125s48
206. Rawlings ND, Waller M, Barrett AJ, et al. (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 42(Database issue):D503-9. doi:10.1093/nar/gkt953
207. Reid AJ, Blake DP, Ansari HR, et al. (2014) Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome Res.* 24(10):1676-1685. doi:10.1101/gr.168955.113
208. Riesen M, Feyst I, Rattanavirotkul N, et al. (2014) MDL-1, a growth- and tumor-suppressor, slows aging and prevents germline hyperplasia and hypertrophy in *C. elegans*. *Aging (Albany NY)* 6(2):98-117
209. Rinschen MM, Wu XW, König T, et al. (2014) Phosphoproteomic analysis reveals regulatory mechanisms at the kidney filtration barrier. *J. Am. Soc. Nephrol.* 25(7):1509-1522. doi:10.1681/ASN.2013070760



Publications in 2014

210. Ritchie GR, Dunham I, Zeggini E, et al. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods* 11(3):294-296. doi:10.1038/nmeth.2832
211. Ritchie GR, Flicek P (2014) Computational approaches to interpreting genomic sequence variation. *Genome Med.* 6(10): doi:10.1186/s13073-014-0087-1
212. Robinson J, Halliwell JA, Hayhurst JD, et al. (2014) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* doi:10.1093/nar/gku1161
213. Rueedi R, Ledda M, Nicholls AW, et al. (2014) Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet.* 10(2):e1004132-e1004132. doi:10.1371/journal.pgen.1004132
214. Ruggeri B, Sarkans U, Schumann G, et al. (2014) Biomarkers in autism spectrum disorder: the old and the new. *Psychopharmacology (Berl.)* 231(6):1201-1216. doi:10.1007/s00213-013-3290-7
215. Ryll A, Bucher J, Bonin A, et al. (2014) A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. *BioSystems* doi:10.1016/j.biosystems.2014.07.002
216. SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32(9):903-914. doi:10.1038/nbt.2957
217. Scelo G, Riazalhosseini Y, Greger L, et al. (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* 5: doi:10.1038/ncomms6135
218. Schmitt BM, Rudolph KL, Karagianni P, et al. (2014) High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface *Genome Res.* 24(11):1797-1807. doi:10.1101/gr.176784.114
219. Schreiber F, Patricio M, Muffato M, et al. (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* 42(Database issue):D922-5. doi:10.1093/nar/gkt1055
220. Schubert M, Iorio F (2014) Exploiting combinatorial patterns in cancer genomic data for personalized therapy and new target discovery. *Pharmacogenomics* 15(16):1943-1946. doi:10.2217/pgs.14.157
221. Schwarz RF, Trinh A, Sipos B, et al. (2014) Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* 10(4):e1003535-e1003535. doi:10.1371/journal.pcbi.1003535
222. Sen S, Young J, Berrisford JM, et al. (2014) Small molecule annotation for the Protein Data Bank. *Database (Oxford)* 2014: doi:10.1093/database/bau116
223. Shameer S, Logan-Klumpler FJ, Vinson F, et al. (2014) TrypanoCyc: a community-led biochemical pathways database for *Trypanosoma brucei*. *Nucleic Acids Res.* doi:10.1093/nar/gku944
224. Sheydina A, Eberhardt RY, Rigden DJ, et al. (2014) Structural genomics analysis of uncharacterized protein families overrepresented in human gut bacteria identifies a novel glycoside hydrolase. *BMC Bioinformatics* 15:112-112. doi:10.1186/1471-2105-15-112
225. Shin SY, Fauman EB, Petersen AK, et al. (2014) An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46(6):543-550. doi:10.1038/ng.2982
226. Shirai H, Prades C, Vita R, et al. (2014) Antibody informatics for drug discovery *Biochim. Biophys. Acta* doi:10.1016/j.bbapap.2014.07.006
227. Sillitoe I, Lewis TE, Cuff A, et al. (2014) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* doi:10.1093/nar/gku947
228. Silvester N, Alako B, Amid C, et al. (2014) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.* doi:10.1093/nar/gku1129
229. Singh Y, Kohli S, Sowpati DT, et al. (2014) Gene cooption in *Mycobacteria* and search for virulence attributes: Comparative proteomic analyses of *Mycobacterium tuberculosis*, *Mycobacterium indicus pranii* and other mycobacteria. *Int. J. Med. Microbiol.* 304(5-6):742-748. doi:10.1016/j.ijmm.2014.05.006
230. Smallwood SA, Lee HJ, Angermueller C, et al. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11(8):817-820. doi:10.1038/nmeth.3035
231. Soler JJ, Avilés JM, Martín-Gálvez D, et al. (2014) Eavesdropping cuckoos: further insights on great spotted cuckoo preference by magpie nests and egg colour. *Oecologia* 175(1):105-115. doi:10.1007/s00442-014-2901-2
232. Soler M, Ruiz-Raya F, Carra LG, et al. (2014) A long-term experimental study demonstrates the costs of begging that were not found over the short term. *PLoS ONE* 9(11): doi:10.1371/journal.pone.0111929
233. Sonnhhammer E, Gabaldón T, Wilter Sousa da Silva A, et al. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics* 30(21):2993-2998. doi:10.1093/bioinformatics/btu492
234. Spivakov M, Auer TO, Peravali R, et al. (2014) Genomic and phenotypic characterization of a wild medaka population: towards the establishment of an isogenic population genetic resource in fish. *G3 (Bethesda)* 4(3):433-445. doi:10.1534/g3.113.008722
235. Studer RA, Oppendoes FR, Nicolaes GA, et al. (2014) Understanding the functional difference between growth arrest-specific protein 6 and protein S: an evolutionary approach. *Open Biol* 4(10): doi:10.1098/rsob.140121
236. Takashima Y, Guo G, Loos R, et al. (2014) Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* 158(6):1254-1269. doi:10.1016/j.cell.2014.08.029
237. Tamuri AU, Goldman N, Dos Reis M (2014) A penalized likelihood method for estimating the distribution of selection coefficients from phylogenetic data. *Genetics* doi:10.1534/genetics.114.162263
238. Ternent T, Csordas A, Qi D, et al. (2014) Standardization and guidelines: how to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* doi:10.1002/pmic.201400120
239. Timpson NJ, Walter K, Min JL, et al. (2014) A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat Commun.* 5: doi:10.1038/ncomms5871
240. Tommaso PD, Bussotti G, Kemena C, et al. (2014) SARA-Coffee web server, a tool for the computation of RNA sequence and structure multiple alignments. *Nucleic Acids Res.* 42(Web Server issue):W356-60. doi:10.1093/nar/gku459
241. Torto-Alalibo T, Purwantini E, Lomax J, et al. (2014) Genetic resources for advanced biofuel production described with the Gene Ontology. *Front. Microbiol.* 5:528-528. doi:10.3389/fmicb.2014.00528
242. Trame CB, Chang Y, Axelrod HL, et al. (2014) New mini-zinc structures provide a minimal scaffold for members of this metalloprotease superfamily. *BMC Bioinformatics* 15(1):1-1. doi:10.1186/1471-2105-15-1
243. Truszkowski A, Daniel M, Kuhn H, et al. (2014) A molecular fragment cheminformatics roadmap for mesoscopic simulation. *J. Cheminform.* 6(1): doi:10.1186/s13321-014-0045-3
244. Tullet JM, Araiz C, Sanders MJ, et al. (2014) DAF-16/FoxO directly regulates an atypical AMP-activated protein kinase gamma isoform to mediate the effects of insulin/IGF-1 signaling on aging in *Caenorhabditis elegans*. *PLoS Genet.* 10(2):e1004109-e1004109. doi:10.1371/journal.pgen.1004109

245. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42(Database issue):D191-8. doi:10.1093/nar/gkt1140
246. Urban M, Pant R, Raghunath A, et al. (2014) The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Res.* doi:10.1093/nar/gku1165
247. Uslu VV, Petretich M, Ruf S, et al. (2014) Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat. Genet.* 46(7):753-758. doi:10.1038/ng.2971
248. Vaga S, Bernardo-Faura M, Cokelaer T, et al. (2014) Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast. *Mol. Syst. Biol.* 10(12): doi:10.15252/msb.20145112
249. Vázquez-Castellanos JF, García-López R, Pérez-Brocá V, et al. (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15:37
250. van Westen GJ, Bender A, Overington JP (2014) Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling. *J. Chem. Biol.* 7(4):119-123. doi:10.1007/s12154-014-0112-2
251. van Westen GJ, Gaulton A, Overington JP (2014) Chemical, target, and bioactive properties of allosteric modulation. *PLoS Comput. Biol.* 10(4):e1003559-e1003559. doi:10.1371/journal.pcbi.1003559
252. Venkata C, Forster MJ, Howe PW, et al. (2014) The potential utility of predicted one bond carbon-proton coupling constants in the structure elucidation of small organic molecules by NMR spectroscopy. *PLoS ONE* 9(11): doi:10.1371/journal.pone.0111576
253. Viereck M, Gaulton A, Digles D, et al. (2014) Transporter taxonomy – a comparison of different transport protein classification schemes. *Drug Discov. Today Technol.* 12:e37-46. doi:10.1016/j.ddtec.2014.03.004
254. Villar D, Flicek P, Odom DT (2014) Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat. Rev. Genet.* 15(4):221-233. doi:10.1038/nrg3481
255. Villaveces JM, Jimenez RC, Habermann BH (2014) KEGGViewer, a BioJS component to visualize KEGG Pathways. *F1000Res* 3:43-43. doi:10.12688/f1000research.3-43.v1
256. Villaveces JM, Jimenez RC, Habermann BH (2014) PsiquicGraph, a BioJS component to visualize molecular interactions from PSICQUIC servers. *F1000Res* 3:44 doi:10.12688/f1000research.3-44.v1
257. Villavicencio-Díaz TN, Rodríguez-Ulloa A, Guirola-Cruz O, et al. (2014) Bioinformatics tools for the functional interpretation of quantitative proteomics results. *Curr. Top. Med. Chem.* 14(3):435-449. doi:10.2174/1568026613666131204105110
258. Vizcaino JA, Deutsch EW, Wang R, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32(3):223-226. doi:10.1038/nbt.2839
259. Vuister GW, Fogh RH, Hendrickx PM, et al. (2014) An overview of tools for the validation of protein NMR structures. *J. Biomol. NMR* 58(4):259-285. doi:10.1007/s10858-013-9750-x
260. Wagih O, Parts L (2014) Gitter: a robust and accurate method for quantification of colony sizes from plate images. *G3 (Bethesda)* 4(3):547-552. doi:10.1534/g3.113.009431
261. Walzer M, Pernas LE, Nasso S, et al. (2014) qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments *Mol. Cell Proteomics* 13:1905-1913. doi:10.1074/mcp.M113.035907
262. Wang R, Perez-Riverol Y, Hermjakob H, et al. (2014) Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics* doi:10.1002/pmic.201400377
263. Welch L, Lewitter F, Schwartz R, et al. (2014) Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput. Biol.* 10(3):e1003496-e1003496. doi:10.1371/journal.pcbi.1003496
264. Welter D, MacArthur J, Morales J, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42(Database issue):D1001-6. doi:10.1093/nar/gkt1229
265. Westbrook JD, Shao C, Feng Z, et al. (2014) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* doi:10.1093/bioinformatics/btu789
266. Wimalaratne SM, Grenon P, Hermjakob H, et al. (2014) BioModels linked dataset. *BMC Syst. Biol.* 8(1): doi:10.1186/s12918-014-0091-5
267. Wong ES, Thybert D, Schmitt BM, et al. (2014) Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* 25:167-178. doi:10.1101/gr.177840.114
268. Wright MW (2014) A short guide to long non-coding RNA gene nomenclature. *Hum. Genomics* 8:7-7. doi:10.1186/1479-7364-8-7
269. Xu QW, Griss J, Wang R, et al. (2014) jmzTab: a Java interface to the mzTab data standard. *Proteomics* 14(11):1328-1332. doi:10.1002/pmic.201300560
270. Yachdav G, Kloppmann E, Kajan L, et al. (2014) PredictProtein - an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* 42(Web Server issue):W337-43. doi:10.1093/nar/gku366
271. Yates A, Beal K, Keenan S, et al. (2014) The Ensembl REST API: Ensembl data for Any language. *Bioinformatics* doi:10.1093/bioinformatics/btu613
272. Yue F, Cheng Y, Breschi A, et al. (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515(7527):355-364. doi:10.1038/nature13992
273. Yue Q, Wu K, Qiu D, et al. (2014) A formal re-description of the cockroach *Hebardina concinna* anchored on DNA barcodes confirms wing polymorphism and identifies morphological characters for field identification. *PLoS ONE* 9(9):e106789 doi:10.1371/journal.pone.0106789
274. Zdravil B, Chichester C, Zander Balderud L, et al. (2014) Transporter assays and assay ontologies: useful tools for drug discovery. *Drug Discov. Today Technol.* 12:e47-54. doi:10.1016/j.ddtec.2014.03.005
275. Zerbino DR, Johnson N, Juettemann T, et al. (2014) WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* 30(7):1008-1009. doi:10.1093/bioinformatics/btt737
276. Zhou Y, Park SY, Su J, et al. (2014) TCF7L2 is a master regulator of insulin production and processing. *Hum. Mol. Genet.* 23(24):6419-6431. doi:10.1093/hmg/ddu359

Organisation of EMBL-EBI Leadership in 2014

RESEARCH GROUPS

Ewan Birney group

Pedro Beltrao group

Proteomics services
Henning Hermjakob

Paul Bertone group

Anton Enright group

Nick Goldman group

John Marioni group

Julio Saez-Rodriguez

Oliver Stegle group

Sarah Teichmann group

Janet Thornton group

Administration
Mark Green

External relations
Lindsey Crosswell

Janet Thornton
Director

Ewan Birney
Associate director

Rolf Apweiler
Associate director

SERVICE TEAMS

Functional genomics
Alvis Brazma

**Protein Databank in Europe
(PDBe)**
Gerard Kleywegt

Protein resources
Alex Bateman

**Functional genomics
development**
Ugis Sarkans

**PDBe content
& intergration**
Sameer Velankar

UniProt content
Claire O'Donovan

UniProt development
Maria-Jesus Martin

Protein families
Rob Finn

EXTERNAL-FACING ACTIVITIES

Web development
Brendan Vaughan

Web production
Rodrigo Lopez

Systems and networking
Petteri Jokinen

Vertebrate genomics

Paul Flicek

Chemogenomics

John Overington

Literature Services

Johanna
McEntyre

Samples, phenotypes & ontologies

Helen Parkinson

Non-vertebrate genomics

Paul Kersey

Cheminformatics

Christoph Steinbeck

Variation

Justin Paschall

European Nucleotide Archive

Guy Cochrane

Training

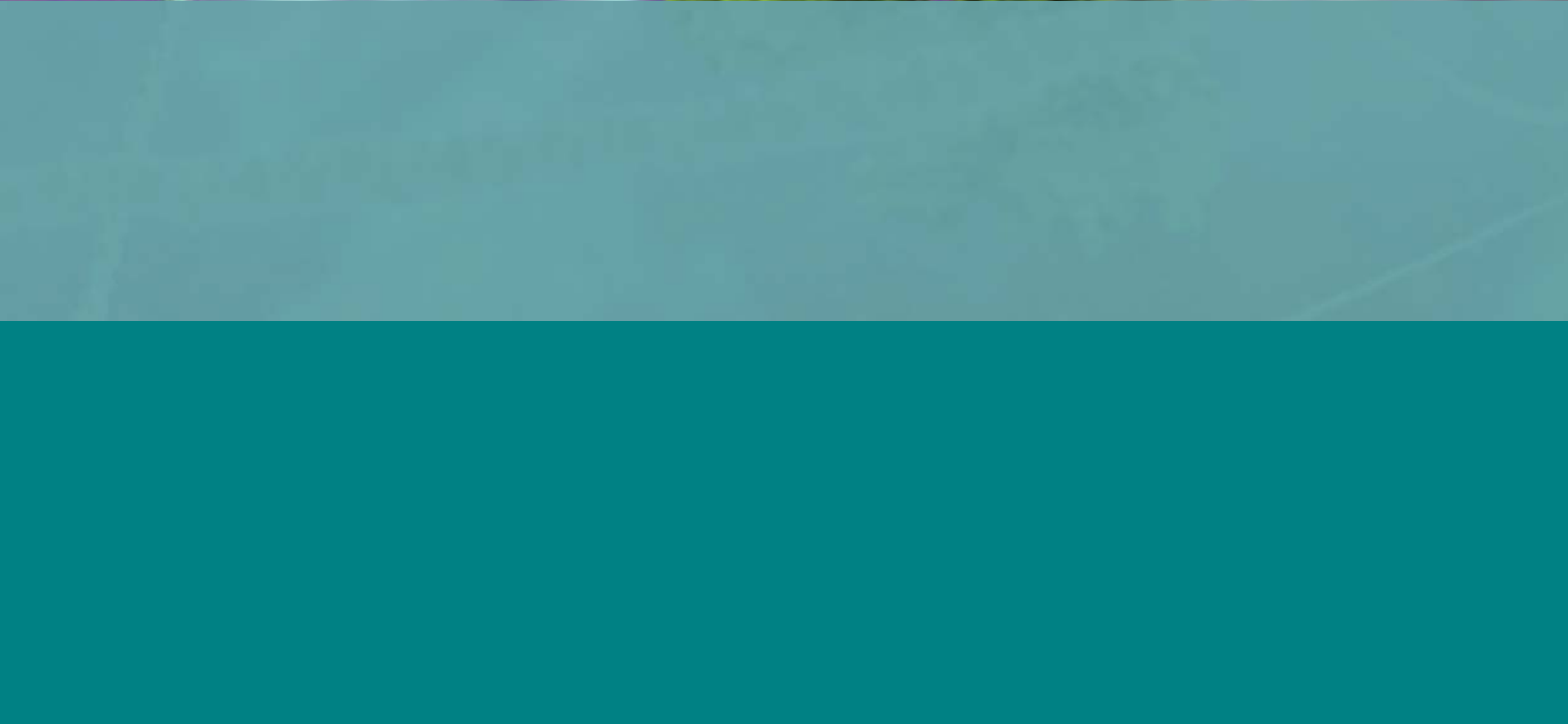
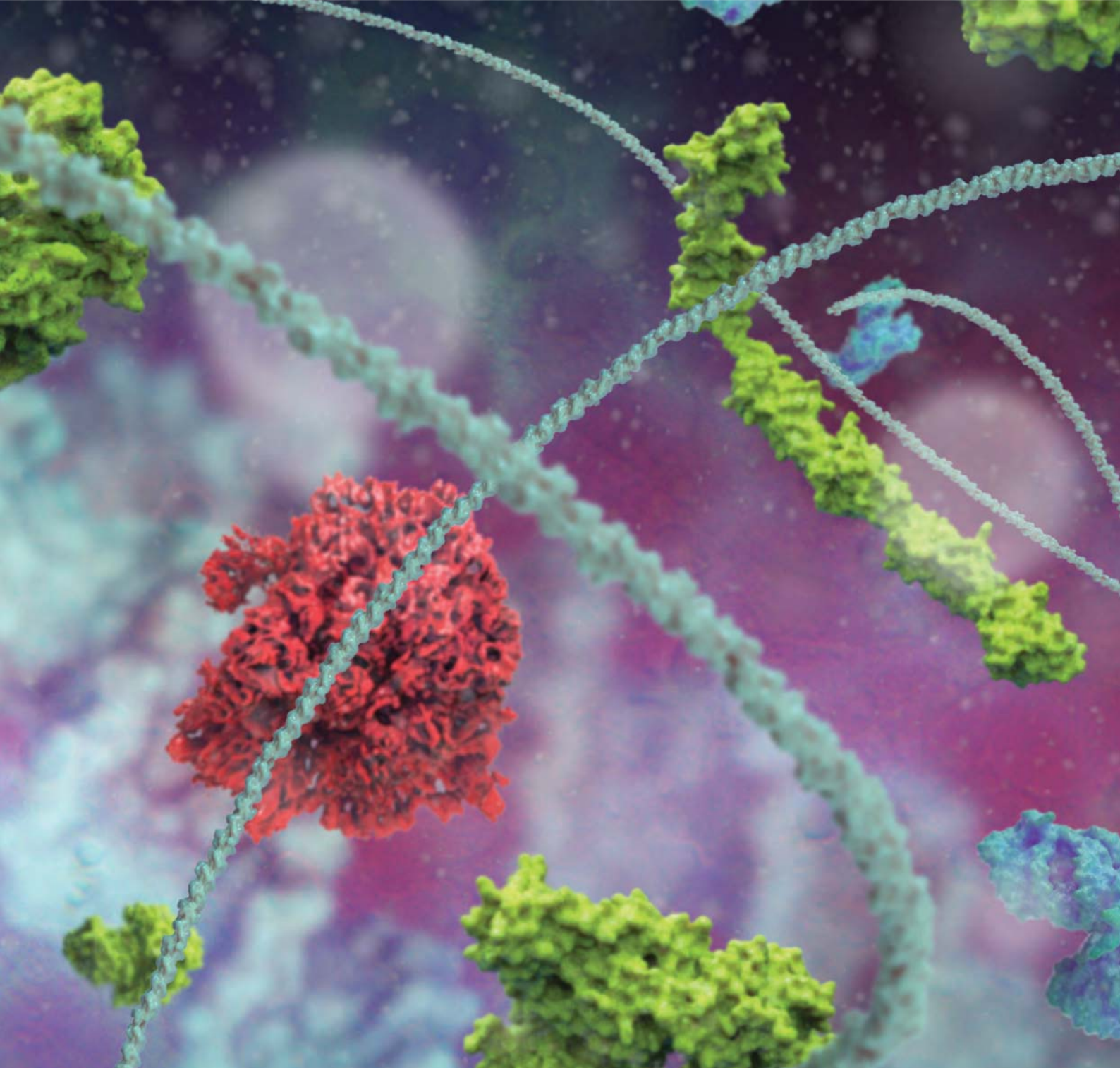
Cath Brooksbank

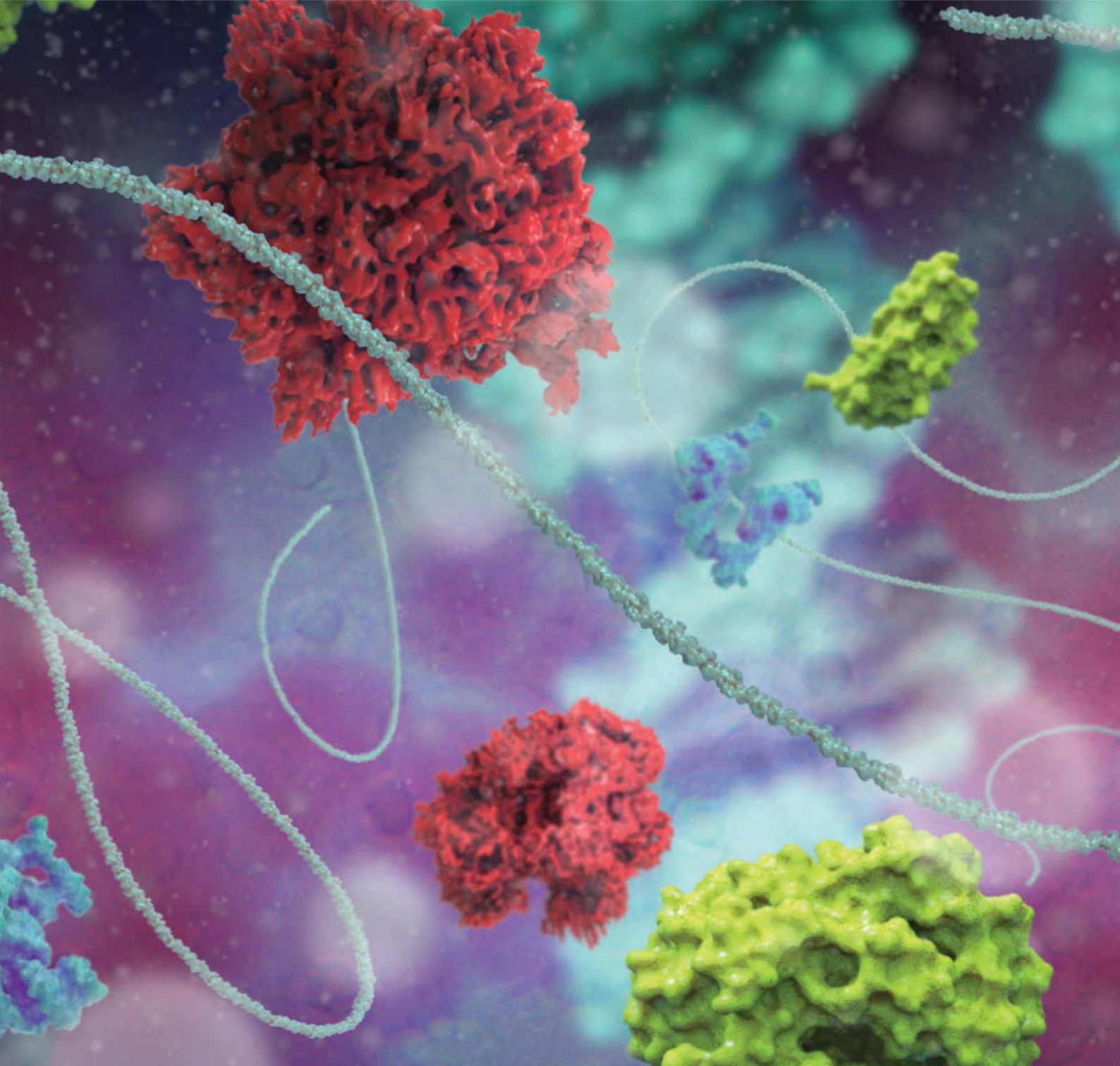
Industry programme

Dominic Clark

Innovation & translation

Jason Mundin, EMBLEM





EMBL-European Bioinformatics Institute
Wellcome Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom

 www.ebi.ac.uk
 +44 (0)1223 494 444
 +44 (0)1223 494 468
 comms@ebi.ac.uk

 @emlebi
 /EMBLEBI
 /EMBLEBI

EMBL member states:

Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, Associate member states: Argentina and Australia

EMBL-EBI is a part of the European Molecular Biology Laboratory

www.ebi.ac.uk/about/brochures