EMBL-European Bioinformatics Institute

Annual Scientific Report 2013

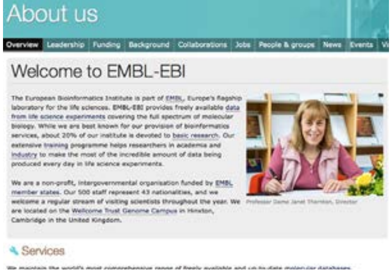
EMBL-EBI

On the cover

Structure 3fof in the Protein Data Bank, determined by Laponogov, I. et al. (2009) Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. Nature Structural & Molecular Biology 16, 667-669.

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



About us

Welcome to EMBL-EBI

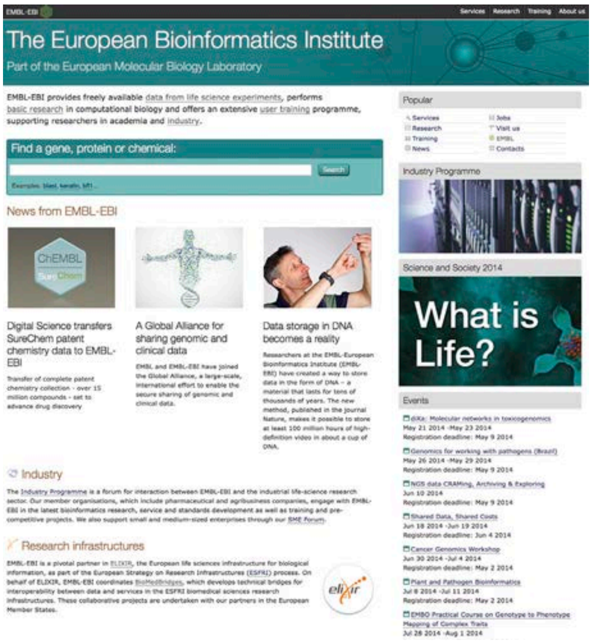
The European Bioinformatics Institute is part of EMBL, Europe's flagship laboratory for the life sciences. EMBL-EBI provides freely available data from life science experiments covering the full spectrum of molecular biology. While we are best known for our provision of bioinformatics services, about 20% of our institute is devoted to basic research. Our extensive training programme helps researchers in academia and industry to make the most of the incredible amount of data being produced every day in life science experiments.

We are a non-profit, intergovernmental organisation funded by EMBL member states. Our 500 staff represent 43 nationalities, and we welcome a regular stream of visiting scientists throughout the year. We are located on the Wellcome Trust Genome Campus in Hinxton, Cambridge in the United Kingdom.

Professor Dame Janet Thornton, Director

Services

We maintain the world's most comprehensive range of freely available and up-to-date molecular databases.



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

News from EMBL-EBI

Digital Science transfers SureChem patent chemistry data to EMBL-EBI

Transfer of complete patent chemistry collection - over 15 million compounds - set to advance drug discovery

A Global Alliance for sharing genomic and clinical data

EMBL and EMBL-EBI have joined the Global Alliance, a large-scale, international effort to enable the secure sharing of genomic and clinical data.

Data storage in DNA becomes a reality

Researchers at the EMBL-European Bioinformatics Institute (EMBL-EBI) have created a way to store data in the form of DNA – a material that lasts for tens of thousands of years. The new method, published in the journal Nature, makes it possible to store at least 100 million hours of high-definition video in about a cup of DNA.

Industry

The Industry Programme is a forum for interaction between EMBL-EBI and the industrial life-science research sector. Our member organisations, which include pharmaceutical and agribusiness companies, engage with EMBL-EBI in the latest bioinformatics research, service and standards development as well as training and pre-competitive projects. We also support small and medium-sized enterprises through our SME Forum.

Research infrastructures

EMBL-EBI is a pivotal partner in ELIXIR, the European life sciences infrastructure for biological information, as part of the European Strategy on Research Infrastructures (ESFRI) process. On behalf of ELIXIR, EMBL-EBI coordinates BioMedBridges, which develops technical bridges for interoperability between data and services in the ESFRI biomedical sciences research infrastructures. These collaborative projects are undertaken with our partners in the European Member States.

Popular

Services	Jobs
Research	Visit us
Training	EMBL
News	Contacts

Industry Programme

Science and Society 2014

What is Life?

Events

diXa: Molecular networks in toxicogenomics
May 21 2014 -May 23 2014
Registration deadline: May 9 2014

Genomics for working with pathogens (Brazil)
May 26 2014 -May 29 2014
Registration deadline: May 9 2014

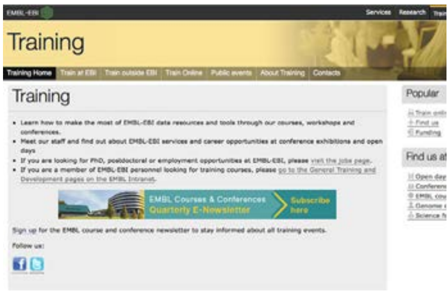
NGS data CRAMing, Archiving & Exploring
Jun 10 2014
Registration deadline: May 9 2014

Shared Data, Shared Costs
Jun 18 2014 -Jun 19 2014
Registration deadline: Jun 4 2014

Cancer Genomics Workshop
Jun 30 2014 -Jul 4 2014
Registration deadline: May 2 2014

Plant and Pathogen Bioinformatics
Jul 8 2014 -Jul 11 2014
Registration deadline: May 2 2014

EMBO Practical Course on Genotype to Phenotype Mapping of Complex Traits
Jul 28 2014 -Aug 1 2014



Training

Training

- Learn how to make the most of EMBL-EBI data resources and tools through our courses, workshops and conferences.
- Meet our staff and find out about EMBL-EBI services and career opportunities at conference exhibitions and open days
- If you are looking for PhD, postdoctoral or employment opportunities at EMBL-EBI, please visit the jobs page.
- If you are a member of EMBL-EBI personnel looking for training courses, please go to the General Training and Development pages on the EMBL Intranet.

EMBL Courses & Conferences Quarterly E-Newsletter Subscribe here

Sign up for the EMBL course and conference newsletter to stay informed about all training events.

Follow us:

Popular

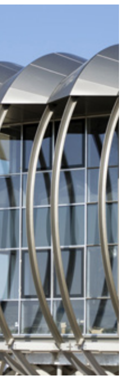
Find us at

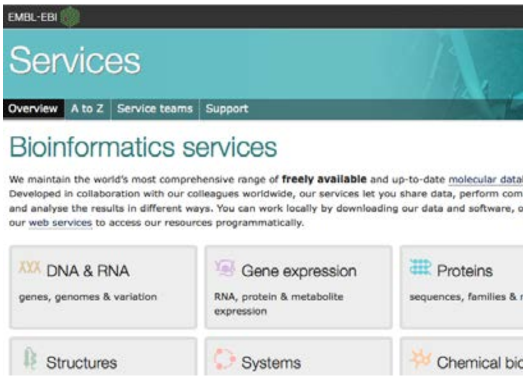
Jobs at EMBL-EBI

About us > Jobs at EMBL-EBI

Want to work with us?

The European Bioinformatics Institute is part of the European Molecular Biology Laboratory, an inter-governmental organisation funded by 20 Member States and one Associate Member State. Our 500+ staff comprise talented scientists, web developers, programmers, project managers and administrators who enjoy a dynamic, creative and supportive international working environment. We are proud to be an equal-opportunity employer, offering a competitive salary and benefits





Services

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular datal
Developed in collaboration with our colleagues worldwide, our services let you share data, perform com
and analyse the results in different ways. You can work locally by downloading our data and software, o
our web services to access our resources programmatically.

DNA & RNA
genes, genomes & variation

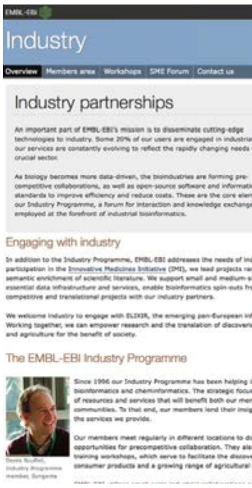
Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & r

Structures

Systems

Chemical bic



Industry

Industry partnerships

An important part of EMBL-EBI's mission is to disseminate cutting-edge
technologies to industry. Some 20% of our users are engaged in industrial
our services are constantly evolving to reflect the rapidly changing needs
crucial sector.

As biology becomes more data-driven, the bioindustries are forming pre-
competitive collaborations, as well as open-source software and informat
standards to improve efficiency and reduce costs. These are the core elem
our Industry Programme, a forum for interaction and knowledge exchange
employed at the forefront of industrial bioinformatics.

Engaging with industry

In addition to the Industry Programme, EMBL-EBI addresses the needs of ind
participation in the Innovative Medicines Initiative (IMI), we lead projects ra
semantic enrichment of scientific literature. We support small and medium-si
essential data infrastructure and services, enable bioinformatics spin-outs fr
competitive and translational projects with our industry partners.

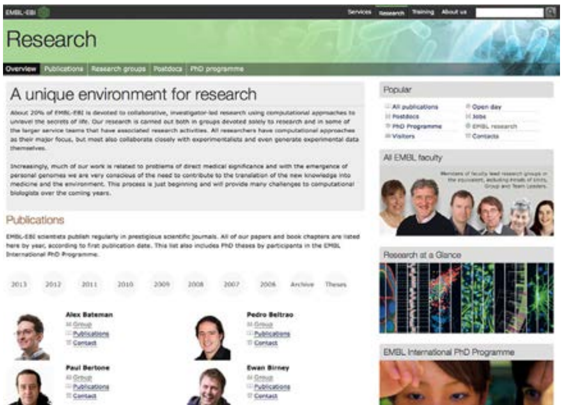
We welcome industry to engage with ELIXIR, the emerging pan-European inf
Working together, we can empower research and the translation of discoveri
and agriculture for the benefit of society.

The EMBL-EBI Industry Programme

Since 1996 our Industry Programme has been helping i
bioinformatics and cheminformatics. The strategic focu
of resources and services that will benefit both our mem
communities. To that end, our members lend their insig
the services we provide.

Our members meet regularly in different locations to di
opportunities for precompetitive collaboration. They als
training workshops, which serve to facilitate the discove
consumer products and a growing range of agricultural

Dennis Rouffet,
Industry Programme
member, Syngenta



Research

A unique environment for research

About 20% of EMBL-EBI is devoted to collaborative, investigator-led research using computational approaches to unravel the secrets of life. Our research is carried out both in groups devoted solely to research and in some of the larger service teams that have associated research activities. All researchers have computational approaches as their major focus, but most also collaborate closely with experimentalists and even generate experimental data themselves.

Increasingly, much of our work is related to problems of direct medical significance and with the emergence of personal genomes we are very conscious of the need to contribute to the translation of the new knowledge into medicine and the environment. This process is just beginning and will provide many challenges to computational biologists over the coming years.

Publications

EMBL-EBI scientists publish regularly in prestigious scientific journals. All of our papers and book chapters are listed here by year, according to first publication date. This list also includes PhD theses by participants in the EMBL International PhD Programme.

2013 2012 2011 2010 2009 2008 2007 2006 Archive Theses

Alex Bateman
Group
Publications
Contact

Pedro Beltrao
Group
Publications
Contact

Paul Bertone
Group
Publications
Contact

Ewan Birney
Group
Publications
Contact

Popular

All publications	Open day
Postdocs	Jobs
PhD Programme	EMBL research
Visitors	Contacts

All EMBL faculty

Members of faculty lead research groups or the equivalent, including Heads of Units, Group and Team Leaders.

Research at a Glance

EMBL International PhD Programme

Foreword

Welcome to EMBL-EBI's 2013 Annual Scientific Report. Here we look back on our major achievements during the year, reflecting on the delivery of our world-class services, research, training, industry collaboration and European coordination of life-science data.

The past year has been one full of exciting changes, both scientifically and organisationally. We unveiled a new website that helps users explore our resources more seamlessly, saw the publication of ground-breaking work in data storage and synthetic biology, joined the Global Alliance for Genomics and Health, built important new relationships with our partners in industry and celebrated the launch of ELIXIR. It really is astonishing to think how things have changed in a single year.

Our new South building opened on the Genome Campus in October, thanks to funding from the United Kingdom's Large Facilities Capital Fund and the efforts of almost every team at the institute. We are very lucky to have expanded into such a beautiful space, which is the home of the ELIXIR Hub, and pleased to welcome a constant stream of visitors and trainees since October.

We are extremely grateful to our funders, especially our EMBL member states, for supporting us through another year of austerity. We were able to 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.

We are in many ways defined by our relationships, and are proud to work with our colleagues throughout the world to define standards for new data, exchange information between resources, develop and evaluate tools for data analysis and share curation of the complex information we make accessible to the public. Our scientists are highly collaborative, and their outstanding work in many different aspects of research and increasingly translation remains at the forefront of computational biology. In addition to our network of academic collaborators, the links we have built with industry remain strong and are evolving in very interesting ways.

If you have read our annual scientific reports in previous years, you may have noticed that the format of the printed version has changed. The printed version contains high-level overviews of key achievements, and we encourage you to explore the work of all EMBL-EBI teams in the online version (www.ebi.ac.uk/about/brochures). Whichever version you choose, we hope you will enjoy reading about what we were up to in 2013.

Sincerely,

JThornton.

Janet Thornton, Director

Rolf Apweiler, Joint Associate Director

Ewan Birney, Joint Associate Director

Major achievements 2013

New beginnings

On a crisp autumn day in 2013, we celebrated the timely opening of our new South building on the Genome Campus, which was built thanks to funding from the United Kingdom's Large Facilities Capital Fund. The new space is light and busy, with a feeling of great energy and community as our employees, visitors and trainees circulate through an open and inviting space. To mark the occasion we held an opening ceremony, featuring guests of honour Rt Hon David Willetts MP, Jackie Hunter, CEO of the Biotechnology and Biological Sciences Research Council (BBSRC), Patrick Vallance, President of Pharmaceuticals R&D at GSK, as well as many of our long-time supporters from industry and partner organisations.

The Rt Hon David Willetts, UK Minister for Universities and Science, at the launch of EMBL-EBI's new South building on the Genome Campus.

The ceremony took place despite hurricane warnings, almost immediately after every team at EMBL-EBI had relocated, some moving into the new space. By the end of the year, all teams had settled in and the technical bumps smoothed out. A big thanks goes to the many teams who contributed to the success of the new building, from project management to technical implementation and graphic design.

Another spectacular unveiling in 2013 was the new EMBL-EBI website, which has a more powerful search, a pleasing new look and feel and offers our users a consistent experience when exploring between our many different data resources. Our user-experience specialists played a pivotal role in this challenging, institute-wide effort, and had a lasting impact on the way we think about connecting with the increasingly diverse scientists who use our bioinformatics services. All of the services we host are expected to adopt the new website guidelines by spring 2015.

New faces

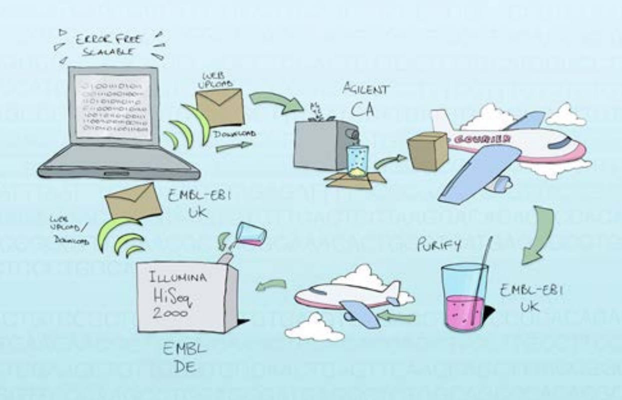
EMBL-EBI has been the driver behind ELIXIR, the infrastructure for biological information in Europe, since its inception. We were very pleased to celebrate the launch of ELIXIR in December 2013, following the arrival of its first Director, Niklas Blomberg, in May. ELIXIR is in good hands as Niklas works with the ELIXIR member states and their 'node' centres of excellence to establish a scientific programme and technical coordination activities.

We also welcomed a new Head of Technical Services in 2013: Steven Newhouse, founding director of the European Grid Infrastructure. Steven's new role at EMBL-EBI involves developing a strategy to convert the diverse needs of the institute's users into solid technical solutions, for cloud and other approaches to handling big data. He will coordinate the efforts of the systems and web teams, who are such a vital part of the institution. Steven is working to expand our collaborations with other EIROFORUM members, and maintains close technical ties with ELIXIR.

New collaborations with industry

Our Industry Programme welcomed pharmaceutical company Bristol Myers Squibb as a new member in 2013. The addition of this large, United States-based company was the result of considerable industry outreach efforts beyond Europe, and we hope to welcome more companies from the United States and Asia to our programme as these efforts continue.

Our Industry Programme offers its members regular opportunities to connect with one another on neutral ground, prioritise topics for workshops and maintain close links with the development of our services. In 2013 we expanded these interactions with new efforts to set up very large-scale, long-term collaborations with companies. To that end, Jason Mundin joined us in September with a remit to facilitate such initiatives, particularly in the area of translational discovery. In a very short space of time Jason has achieved a great deal, with oversight of a major new initiative with GSK and the Sanger Institute that will be launched in 2014.



An illustration showing how EMBL-EBI researchers carried out their work on DNA storage.

New science

One of the most exciting developments in 2013 was the publication of a novel, scalable approach to the long-term archiving of data. Nick Goldman, Ewan Birney, Paul Bertone and others at EMBL-EBI captured the imaginations of people from all walks of life by proposing to use the ‘natural’ storage archive provided by DNA itself. Their ideas continue to gain attention in the popular press. The new method involves translating binary digital files into non-repeating strings of A, T, G and C and – crucially – applying an error-correction algorithm similar to those applied in everyday technologies such as mobile phone transmission. In collaboration with Agilent Technologies in the United States, the group used synthetic biology techniques to manufacture those files as DNA. The resulting material is almost invisible to the naked eye, amounting to what looks like a thin layer of dust at the bottom of a phial. (In fact, DNA is so compact that if this method were used to synthesise all of the world’s video files as DNA, the material would fit into a standard-sized teacup.) Once they had the synthetic DNA in hand, with help from the EMBL Heidelberg sequencing facility they were able to read back the following files with almost 100% accuracy: all of Shakespeare’s sonnets, a segment from Martin Luther King’s “I Have a Dream” speech, a photograph of EMBL-EBI, a PDF of Watson and Crick’s seminal paper describing the structure of DNA and a text file of the code itself. This revolutionary approach to low-energy, long-term data storage opens up a world of possibilities, and the group will continue its collaboration.

Another stand-out research achievement in 2013 was the launch of the Sanger-EBI Single Cell Genomics Centre, which features technology and methods that give a whole-genome view of variability and expression at the single-cell level. The centre’s first paper on single-cell RNA sequencing (Brennecke et al., 2013) was published in its inaugural year.

In 2013 the Thornton group released EC-BLAST: software that compares enzyme reactions and produces a hit list of the most similar reactions. After five years under construction by a highly interdisciplinary team, the tool offers scientists a BLAST-like way to compare enzymes strictly according to their function. This is particularly useful for researchers who are working on novel enzymes for ‘green’ biotech.

Our scientists continue to be involved with developing the new methods desperately needed for handling and interpreting the flood of sequence and other data, applying them to understand more about gene regulation and its evolution, RNA functions and the small molecules that are essential for life.

Digital Science, a MacMillan company, donated the SureChem database of patented chemical compounds to EMBL-EBI in 2013.

Extending our resources

We were proud to accept the donation of SureChem, a large collection of patented chemical structures, from Digital Science, a MacMillan company. This collection (now dubbed SureChEMBL) grows the ChEMBL bioactive entity resource from around 1 million chemical compounds to approximately 15 million, many of which are of commercial or clinical relevance. The public availability of these structures, extracted from the patent literature and paired with bioactivity data where it is available, makes a big difference to researchers working in drug discovery, enabling the swift lookup of a new compound’s novelty.

The scientific literature provides the essential context for the data we host at EMBL-EBI, and considerable efforts are devoted to mining it on a large scale. In our first full year of leading the development of Europe PubMed Central, our Literature Services team expanded the resource considerably. Over 13 million articles in Europe PMC have been cited at least once, representing the largest public-domain citation network in the world.

The technical basis of gene-expression studies has traditionally been microarray experiments, but this has been expanding in recent years to include RNAseq experiments. In 2013 our Functional Genomics teams incorporated a vast amount of RNAseq data into the ArrayExpress archive, and released the new baseline Expression Atlas. The Atlas now encompasses high-throughput-sequencing-based expression data on gene expression levels in healthy,

Major achievements 2013

untreated conditions. This is just the beginning of our ambitious efforts to build the expression fingerprint for each different cell type and capture it in a fully scalable Expression Atlas.

Our service strategy centres on enabling discovery through the integration of all manner of data pertaining to the processes of life. Our interactions with industry partners led to the launch in 2013 of a new Resource Description Framework (RDF) platform that supports Semantic Web technologies. RDF provides easy links between related but differently structured information, enabling the meaningful and intuitive sharing of molecular data amongst different applications. Our users can now make a single query to retrieve all relevant data from UniProt, ChEMBL, the Expression Atlas, Reactome, BioSamples and BioModels.

New genomes

A vastly improved version of the bread wheat genome was made accessible through Ensembl Plants at the end of 2013. These data from the Chromosome Survey Sequence, generated under the auspices of the International Wheat Genomes Sequencing Consortium, represent the most complete version of the wheat genome to date. The resource also provides access to evolutionary trees showing the relationship of the genes encoded in each of the three wheat genomes to each other, as well as to other cereals and plants. These data are an invaluable resource to molecular researchers working in agriculture.

The 1000 Genomes Project, which has collected data on more than 2500 genomes, will release its final integrated variant set in 2014. The rich data produced in the project has been an invaluable resource, with extensive uptake by EMBL-EBI scientists during 2013. For example, the GEUVADIS consortium, of which EMBL-EBI is a member, drew on the datasets to produce the largest-ever dataset linking human genomes to gene activity at the level of RNA. Their work adds a functional interpretation to the most important catalogue of human genomes, and the GEUVADIS data are now freely available though ArrayExpress.

Another area of intense activity in 2013 was development of the new European Variation Archive, which incorporates over 40 million smaller substitutions and indels from large population surveys such as the 1000 Genomes Project and Genome of the Netherlands. This integrating service, managed by the Variation team, is set to launch in 2014.

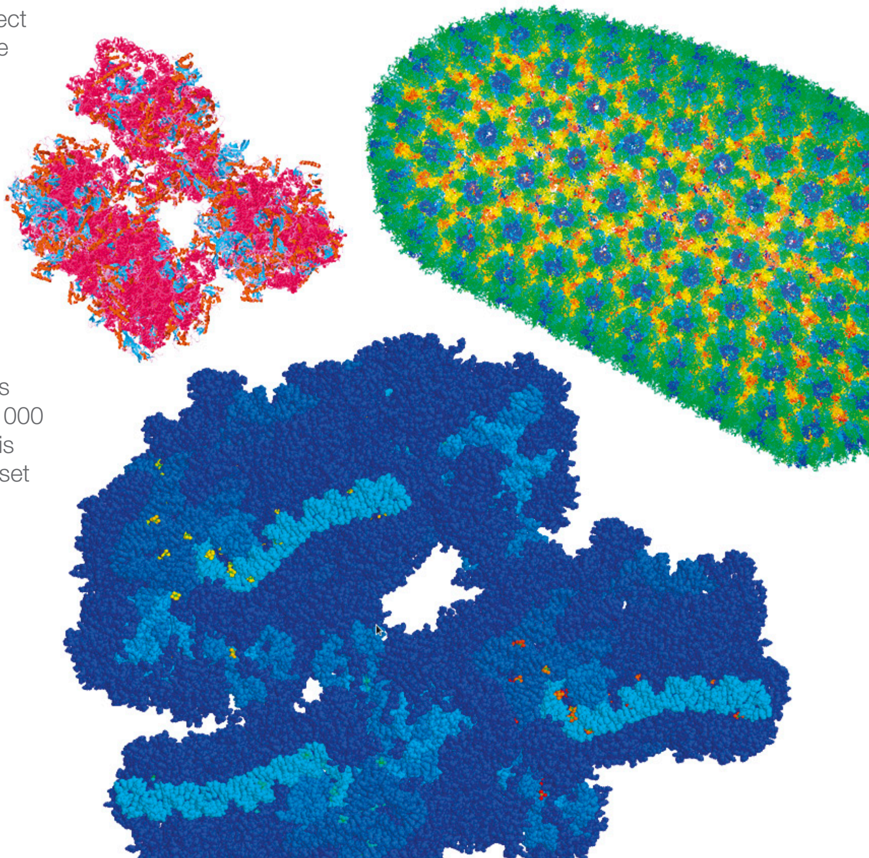
Data supernova

In 2013 the volume of data produced by life science researchers and hosted by EMBL-EBI once again doubled. We now have approximately 40 petabytes of storage (compare to 18 petabytes in 2013) to accommodate these data, including backups and workspace. But the rate of growth continues to outstrip the falling price and rising capacity of commercially available storage technology, while our funding landscape remains fairly level.

EMBL-EBI has not been lacking in creative solutions for this challenge. The Cochrane team and Birney group launched CRAM, an innovative data-compression software and toolkit that dramatically reduces the amount of space it takes to store a full genome sequence. Amongst the earliest adopters of the software was the Wellcome Trust Sanger Institute, Europe's largest sequence data producer. Illumina, the dominant sequencing platform provider, will provide CRAM as an output option from their sequencing machines from early 2014.

BLUEPRINT, an EU-funded 'high impact project', is working to generate at least 100 reference epigenomes of blood-cell components from healthy individuals and those suffering from leukaemia. In 2013 the project generated reference epigenomes for monocytes and neutrophils, and the Vertebrate Genomics team oversaw three releases of these data via the secure European Genome-phenome Archive (EGA).

The first 'big' macromolecular structures – those that are too big to fit in a traditional Protein Data Bank



The Protein Data Bank's new, more flexible system now accepts a wider variety of macromolecular data, including very large files.

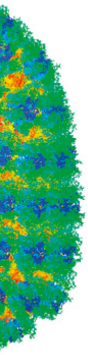
(PDB) entry – were deposited into the PDB and released into the public domain in 2013. This marks a turning point for the PDB as it becomes increasingly accessible to non-specialists. In the past year the archives saw their 10 000th NMR structure and the 2000th entry to the Electron Microscopy Data Bank.

Training

Our goal is to empower researchers worldwide, and these efforts are greatly facilitated by the training we offer. With new data types and ever-advancing data integration, demand for bioinformatics training intensified considerably, with our IT Training Room fully occupied by courses.

But our face-to-face training activities, the majority of which are led by researchers, extend well beyond a single room. In 2013 our Training team brought together specialists from all over the institute, actively involving more than 140 members of staff in 250 events worldwide, reaching thousands of people in 28 countries.

The uptake of online learning platforms has skyrocketed in recent years, and this was certainly the case for Train online. Usage more than doubled over 2012 levels, as biologists working in all areas learned how to make the most of our resources in their own time and at their own pace.



Building bridges

As part of EMBL we are fortunate to be involved in countless scientific endeavours in the member states, and indeed around the world. Our scientists frequently visit collaborating organisations, review grants and advise developments in research institutes around Europe. These relationships are invaluable to us because they give us detailed insights to the needs and future directions of our collaborators. Conversely, the bridges we build enable our member states to tap into a rich seam of knowledge and to inform the evolution of our resources.

The ELIXIR pilot actions provide excellent examples of different countries and institutes working to resolve common technical challenges. As ELIXIR entered its construction phase in 2013, EMBL-EBI teams worked with the CSC IT Centre for Science in Finland and the Centre for Genomic Regulation (CRG) in Barcelona, Spain on the first round of these actions.

With CSC, our Variation team worked on a new federated authentication and authorisation system for the EGA, delivering a framework for connecting trusted institutional logins to EGA accounts. With the CRG, we undertook to expand the EGA to include data resources from both institutes, creating a practical model for distributed, secure access to biomedical research data. Another effort, led by our Systems team, was the construction of a Lightpath – a dedicated ethernet circuit – over European academic networks between EMBL-EBI and CSC Finland.

EMBL-EBI is central to ELIXIR and operates on its behalf in supporting many other research infrastructures. In 2013 our newly created Samples, Phenotypes and Ontologies team, led by Helen Parkinson, led the way in ensuring the adoption of consistent ontologies at EMBL-EBI and played a part in helping Europe's research infrastructures manage their data in a consistent manner. As part of these efforts, her team implemented a new web portal for Infrafrontier, the infrastructure for mouse model data, and laid the groundwork for data distribution in the International Mouse Phenotyping Consortium. Both of these efforts acknowledge the importance of model species and aim to maximise efficiency and reduce waste through standardisation and the provision of global access to reference data.

One of the most important aspects of any research infrastructure is community building, so that people working in different areas become aware of where their work may intersect. Many of our service teams contribute to BioMedBridges, an EU-funded initiative to bring together the many different life-science research infrastructures around shared data challenges: access, security and interoperability. The project achieved many technical advances in 2013, including a new European service registry in collaboration with ELIXIR that provides central access to life science tools, applications and services. Similarly, the UniChem resource was expanded in collaboration with EU-OPENSCREEN to provide cross-references to more than 65 million chemical structures from 22 source databases.

EMBL and EMBL-EBI joined the Global Alliance for Genomics and Health, a large-scale, international effort to enable the secure sharing of genomic and clinical data, in 2013. The alliance comprises nearly 70 organisations in Asia, Australia, Africa, Europe, North America and South America that are committed to creating a common framework to support data analysis and protect the autonomy and privacy of participating individuals. We sincerely hope that our work as an institute and as contributors to European research infrastructures will help the alliance set standards and encourage the sharing of data that will enable biomedical science to progress to the next level, and deliver health solutions that benefit all of humankind.

Genes, genomes and variation

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

The European Nucleotide Archive (ENA) manages a staggering amount of raw sequence data from all domains of life and feeds them into other resources that build knowledge layers on this foundational information. The Ensembl genome explorer enables and advances genome science by providing high-quality, integrated annotation on genomes within a consistent and accessible infrastructure. Ensembl Genomes offers access to genome-scale data on non-vertebrate species. The European Genome-phenome Archive (EGA), our controlled-access resource for human variation data, holds potentially identifiable genetic, molecular and phenotypic data resulting from biomedical research projects.

In 2013 the ENA saw the first submissions of raw sequence data using the new compression format, CRAM, which has radically improved the efficiency of genome data storage. Ensembl was enriched with new information about the human genome and several model species, and welcomed the genomes of 11 new species, including sheep. Ensembl also expanded its gene variation data for several species, adding phenotype data for model organisms relevant to clinical research. Ensembl Plants, part of Ensembl Genomes, provided access to a much-improved bread-wheat genome, an invaluable resource for agricultural researchers. The European Genome-phenome Archive, meanwhile, was central to two ELIXIR pilot projects to demonstrate a practical approach to the distributed, secure sharing of biomedical research data.

Database of Genomic Variants archive (DGVa)

DGVa is a central archive that receives data from, and distributes data to, a number of resources. The DGVa accepts direct submissions from researchers and accession numbers for data objects included in these are given the prefix ‘e’. The DGVa also exchanges data on a regular basis with dbVar (a peer archive hosted by NCBI in the United States). Data objects accessioned by dbVar have the prefix ‘n’. You can retrieve DGVa data from the data download page, search the DGVa using Biomart, and view the data in a genomic context using Ensembl. The DGVa also supplies data to DGV (Database of Genomic Variants, hosted by The Centre for Applied Genomics in Canada), where further curation and interpretation is carried out.

www.ebi.ac.uk/dgva

European Genome-phenome Archive (EGA)

The EGA contains exclusive data collected from individuals whose consent agreements authorise data release only for specific research use or to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the EGA project. As an example, only members of the EGA team are allowed to process data and only in a secure computing facility. Once processed, all data are encrypted for dissemination and the encryption keys are delivered offline. An independent Ethics Committee audits the EGA protocols and infrastructure. The EGA help desk will answer any requests at ega-helpdesk@ebi.ac.uk.

www.ebi.ac.uk/ega

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 are 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. Our exploitation of these systems will continue to develop. With ongoing focus on data presentation, integration within ENA, integration with resources external to ENA, tools provision and services development, the team’s 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 both automatic and manually curated annotation on selected eukaryotic genomes. Automatic annotation is based on mRNA and protein information. Ensembl provides valuable insights into variation within and between species, and allows users to compare whole genomes to identify conserved elements. It is integrated with several other important molecular resources, for example UniProt, and can be accessed programmatically. Ensembl is developed as a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute.

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 (Kersey et al., 2012) 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

Metagenomics

Our Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data, and provides insights into the functional and 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 (CMs). Our families may be divided into three broad functional classes: non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Rfam uses covariance models which simultaneously model RNA sequence and structure to provide better discrimination than purely sequence based models. Rfam families are created from European Nucleotide Archive sequence data and experimental evidence in the literature. In addition to alignments and CMs, we provide 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.sanger.ac.uk

Genes, genomes and variation

Guy Cochrane

- *Revised and extended the ENA data structures that are used to represent assembled information;*
- *Extended the Webin submission system to support the capture of data into these new structures;*
- *Progressed towards offering an end-to-end service for assembly data, including intuitive interactive submissions, fully automated programmatic submissions capable of operating at scale and a comprehensive discovery and retrieval service (launch date early 2014);*
- *Enhanced the Webin data submissions application and introduced support for new data types;*
- *Launched CRAM v. 2 sequence data compression software, representing an extended CRAM format, a public registry of reference sequences and a substantially enriched Java toolkit;*
- *Improved ENA Advanced Search facilities.*

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

Paul Flicek

- *Added 11 new species to Ensembl, five of which are fully supported and include the economically important sheep (Ovis aries), and a model for evolutionary processes, Astyanax mexicanus;*
- *Issued four major releases of Ensembl, with updates to human, mouse and zebrafish resources;*
- *Further developed cloud-based gene annotation in collaboration with Rupert Lück at EMBL Heidelberg and HelixNebula;*
- *Expanded the number of species with variation data and added phenotypes from the mouse and zebrafish knockout experiments, also enriching Ensembl resources for clinical data;*
- *Redesigned the Ensembl Regulatory Build to take advantage of the machine-learning techniques developed under the ENCODE project (to be released in summer 2014);*
- *Issued three public releases of Blueprint data and established the Blueprint Data Access Committee to allow the community to apply for access to the data deposited in the European Genome-phenome Archive (EGA);*
- *Worked with International Human Epigenome Consortium partners to define metadata and assay standards for epigenomic data, establish protocols for distributing the data generated by these projects, and attach IHEC results to browsers like Ensembl;*
- *Worked with HipSci partners to ensure data archiving and distribution mechanisms are in place as the project moves into its production phase.*

www.ensembl.org
www.blueprint-epigenome.eu
www.hipsci.org

Nematode.
Photo courtesy of Arturo Agostino

Paul Kersey

- *Overhauled the Ensembl Bacteria portal to offer interactive and programmatic access to every annotated bacterial genome submitted to the ENA (9765 in the current release);*
- *Issued five public releases of Ensembl Genomes, and contributed to the regular data releases of VectorBase, WormBase and PomBase;*
- *Added 12 new invertebrate metazoan genomes to Ensembl Genomes, and were heavily involved in producing the primary annotation for the kissing bug* (Rhodnius prolixus) *and the myrioapod rum;*
- *Collaborated on a project to annotate 16 different mosquito species (data to be released in 2014);*
- *Contributed to the annotation of* Brugia malayi, *the causative agent of lymphatic filarisis (elephantiasis);*
- *Added the genomes of 12 fungi, four protists, and four plants to the public release;*
- *Added the sequence and annotation from the chromosomal survey sequence assembly released by the International Wheat Genomes Sequencing Consortium to Ensembl Plants, and provided access to evolutionary trees showing the relationship of the genes encoded in each of the three wheat genomes to each other, and to other species;*
- *Developed a pipeline for the scalable management of variation data in transplant;*
- *Began development of RNAcentral, a new resource for information about non-coding RNAs;*
- *As part of a large-scale international effort, published the second Assemblathon paper comparing methods for genome assemblies and defining the methods by which they can be compared.*

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

Justin Pascall

- *In the context of ELIXIR and the Global Alliance for Genomics and Health, worked on developing a 'gold standard' in secure provision of genomics-based human biomedical research data;*
- *In collaboration with the Web Development team, performed major upgrades to the EGA service including improvements to security and full automation of data distribution and upload;*
- *Conducted an ELIXIR pilot study with CSC IT Centre for Science in Finland on federated authentication and authorization of EGA, and delivered a framework for connecting trusted institutional logins to EGA accounts;*
- *Expanded EGA to include resources developed and hosted at the Centre for Genomic Regulation (CRG) in Barcelona, Spain, in the interests of providing a model for distributed, secure access to biomedical research data;*
- *Grew DGVa to host 118 studies across 31 studies, representing over 14 million structural variant observations and over 15 million genotypes;*
- *Developed the new European Variation Archive integrating resource (to launch in 2014), which incorporates over 40 million smaller substitutions and indels from large population surveys such as the 1000 Genomes Project and Genome of the Netherlands.*

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

Alex Bateman

- *Issued Rfam 11.0, which introduces genome-based alignments for large families and the Rfam Biomart.*

http://rfam.sanger.ac.uk

Sarah Hunter

- *Developed EBI Metagenomics taxonomic and functional analysis pages to allow interactive visualisation of information;*
- *EBI Metagenomics reached 40 public metagenomics projects, comprising 1309 separate samples and a significant number of privately held studies.*

www.ebi.ac.uk/metagenomics

European Nucleotide Archive

Our team builds and maintains the European Nucleotide Archive (ENA), the globally comprehensive data resource that preserves the world's public domain output of sequence and related data and makes it available to the research community.

Our team builds and maintains the European Nucleotide Archive (ENA), the globally comprehensive data resource that preserves the world's public domain output of sequence and related data and makes it available to the research community. As a result of the extension and adaptation of software and services that started within ENA, we provide technology and data repository support functions beyond ENA, including the CRAM sequence-data compression technology and the Webin submissions application that supports several of EMBL-EBI's data resources. We are committed to the utility of ENA content and to achieving the broadest reach of sequencing applications.

As nucleotide sequencing becomes increasingly central to applied areas such as healthcare and environmental sciences, ENA data and services have become a core foundation upon which scientific understanding of biological systems are assembled. Our users span primary ENA data and service users and secondary service providers (e.g., UniProt, Ensembl, Ensembl Genomes, ArrayExpress) that build on ENA content. We focus on data presentation, data integration within ENA and with resources external to ENA, the provision of analytic tools and the development of services.

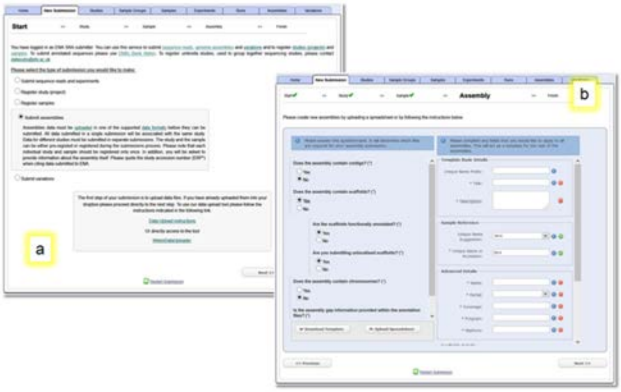
Our technology provision functions include the public distribution of software, hosting of data submissions applications, support for users of the software and the building of user and developer communities around these technologies.

Major achievements

In 2013 we focused on responding to very rapid growth of genome assembly in our user community. This continuing growth is in part due to the increasing viability of whole-genome shotgun-sequencing methods in the analysis of pathogen populations, and in part to advances in assembly methods for more complex genomes. Over the year we revised and extended the ENA data structures that are used to represent assembled information, respecting the richness of description of a full clone-based assembly approach where required but supporting simpler information from one-step assembly processes. We extended the Webin submission system to support the capture of data into these new structures (see Figure 1) and carried out content curation with our global partners to assure consistency. Building on 2012 work in which we deployed search and browser coverage of assembly data, we made substantial progress towards offering an end-to-end service for assembly data. In early 2014 we will have launched all components and will offer intuitive interactive submissions, fully automated programmatic submissions capable of operating at scale and a comprehensive discovery and retrieval service.

We substantially enhanced the Webin data submissions application. Webin supports both small-scale interactive use (through its intuitive web application interface) and high-throughput programmatic use (through its RESTful API). We continued to develop Webin functionalities that are based on workflows in which users upload data and other structured information into their secure user area and subsequently control transactions on these data. These workflows were first deployed for submissions of raw read data and have proven successful and popular. Our approach combines the implementation of both new submissions functionalities and existing ones that lie outside the Webin framework. We introduced support for new data types, such as tabular-derived data for ENA (e.g. functional assignments to clustered environmental reads), assembly data files (AGP) for ENA and, through collaborative development with the Paschall team, variation data (VCF) for EGA.

The CRAM sequence data compression software was further developed, and usage increased throughout 2013. In June 2013 we launched CRAM v. 2, representing an



Webin web application showing (a) assembly data type selection and (b) a 'wizard' that takes the user through a series of questions about the assemblies he or she is submitting (left), and produces a custom form to collect information (right).

Guy Cochrane

PhD University of East Anglia, 1999.

At EMBL-EBI since 2002. Team Leader since 2009.

extended CRAM format, a public registry of reference sequences and a substantially enriched Java toolkit, compliant with such tools as Picard and GATK. In addition, our Wellcome Trust Sanger Institute (WTSI) collaborators released a library of CRAM utilities in C that provide high-performance access to CRAM functions and support integration with technologies such as SamTools. Amongst CRAM adopters is WTSI – Europe's largest data producer and one of only a handful of major global sequencing players. Since late 2013, all routine read data flowing into ENA from WTSI have been arriving in CRAM format. Illumina, the dominant sequencing platform provider, will provide CRAM as an output option from their sequencing machines from early 2014.

We made a variety of improvements and extensions to ENA Advanced Search facilities. One derived browser product is the ENA Marker Portal (Figure 2), an interface to support marker selection and analysis that allows users to slice ENA marker locus sequences by taxonomy and targeted marker gene. Behind this portal lies a curated set of dictionaries and rules that support the discovery of marker locus sequences from this heterogeneously annotated collection within ENA, as well as a specific application of ENA Advanced Search functions.

Future plans

In 2014 we will provide ongoing content management, submissions support and browser helpdesk functions for ENA and will support our CRAM and Webin technologies. Beyond these operational activities, we will focus our development effort in a number of areas.

We will continue to rationalise submissions functionalities into the Webin framework and will extend support to cover new data types. We will develop a system through which functional annotation can be submitted, validated and updated at scale through both interactive and RESTful interfaces. This work requires rule-based semantic validation routines and extensive dictionary support for annotated fields.

Further development work around CRAM technology will include its extension to broader data types, including support for long-read third-generation sequencing platforms and methylation data derived from kinetic outputs. We will continue to work with the community to explore lossy models and their appropriateness for different experimental applications.

A major area of work will be the re-structuring of curation activities in ENA. Curation of sequence and related data represents the insertion of biological knowledge into database content by experts. While in our traditional model we apply curation to incoming submitted data sets independently, a new model of working will allow us to continue to sustain rapid, ongoing growth in content while retaining impactful scientific value addition. In this new system, data will flow more automatically through validation systems into ENA, and curation will be provided in the form of biological dictionary and rule management (to support validation) as well as post hoc attention to classes of information across multiple submissions (e.g. functional annotation for a given gene family across species). The substantial work required relates to the development and adoption of curator workflows and protocols, validation rules, dictionaries and curation support tools.

Selected publications

Pakseresht, N., Alako, B., Amid, C., et al. (2014) Assembly information services in the European Nucleotide Archive. Nucleic Acids Res. 42: D38-D43.

Hunter, S., Corbett, M., Denise, H., et al. (2013) EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. 42: D600-D606.

Cochrane, G., Alako, B., Amid, C., et al. (2013) Facing growth in the European Nucleotide Archive. Nucleic Acids Res. 41, D30-D35.

Cochrane, G., Cook, C.E. and Birney, E. (2012) The future of DNA sequence archiving. Gigascience 1, 2.

Nakamura, Y., Cochrane, G. and Karsch-Mizrachi, I., on behalf of the International Nucleotide Sequence Database Collaboration. (2013) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 41, D21-D24.

Yilmaz, P., Kottmann, R., Field, D., et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415-420.

Vertebrate genomics

The Vertebrate Genomics team develops the Ensembl genome annotation resources and analysis infrastructure in collaboration with the Wellcome Trust Sanger Institute, and leads data management and analysis activities for several large-scale genomics projects including the 1000 Genomes Project, HipSci and Blueprint as part of the International Human Epigenome Consortium (IHEC).

The resources and services of the Vertebrate Genomics team are made publicly available to ensure widest possible use by the scientific community.

Our research efforts focus on the evolution and genetics of transcriptional regulation, with an ultimate goal of understanding mechanisms and maintenance of cell-type specificity. Our major research collaboration with Duncan Odom's group at the University of Cambridge pioneers methods of comparative regulatory genomics for biological discovery across a wide range of mammalian and vertebrate species. We also integrate analysis techniques to assess the functionality of disease-associated genetic variants in human populations.

Major achievements

Ensembl

Ensembl's four major releases in 2013 included regular updates to our human, mouse and zebrafish resources to include manual annotation for both protein-coding and non-coding genes. We added 11 new species to Ensembl, five of which are fully supported and include the economically important sheep (Ovis aries), and a model for evolutionary processes, Astyanax mexicanus. Cloud-based gene annotation continues to develop in collaboration with Rupert Lück at EMBL Heidelberg and HelixNebula.

Variation and regulation data are available for human and other species. We present genomic variation by incorporating and annotating 'patches' to the reference assembly for human and mouse. In 2013 we expanded the number of species with variation data and added phenotypes from the mouse and zebrafish knockout experiments as well as enriching the resources we have for clinical data. Our Variant Effect Predictor (VEP) continues to be a useful tool for indicating the consequence and severity of genomic variations. In 2013 we entirely redesigned the Ensembl Regulatory Build to take advantage of the machine-learning techniques developed under the ENCODE project. This new build will be released in summer 2014.

Our REST API was accessed over 14 million times in 2013, with its VEP bindings proving the most popular. Ensembl continues to engage directly with its users through a variety of channels, including a helpdesk, on-line video examples and social media such as Twitter (ca. 2000 followers). In addition, team members have carried out nearly 100 training courses both on site in Hinxton and throughout the world.

International Human Epigenome Consortium

We made three public releases of Blueprint data and established the Blueprint Data Access Committee to allow the community to apply for access to the data deposited in the European Genome-phenome Archive (EGA). In the context of Blueprint we are working with other IHEC partners to define metadata and assay standards for epigenomic data, establish protocols for distributing the data generated by these projects, and attach IHEC results to browsers like Ensembl. In 2013 the Ensembl regulatory team worked closely with Blueprint consortium members to provide expertise on epigenomic data integration and to ensure that Blueprint data is used in Ensembl's regulatory annotation process as quickly as possible.

We also lead data coordination in the HipSci project, which aims to create a catalogue of human induced pluripotent stem cells (iPSCs) and define the baseline genomic, epigenomic and proteomic state of these lines. In 2013 we worked with the Wellcome Trust Sanger Institute and other HipSci partners to ensure data archiving and distribution mechanisms are in place as the project moves into its production phase.

Both Blueprint and HipSci produce data sets that complement the reference human variation data we continue to manage and make available as part of the 1000 Genomes Project.

Paul Flicek

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

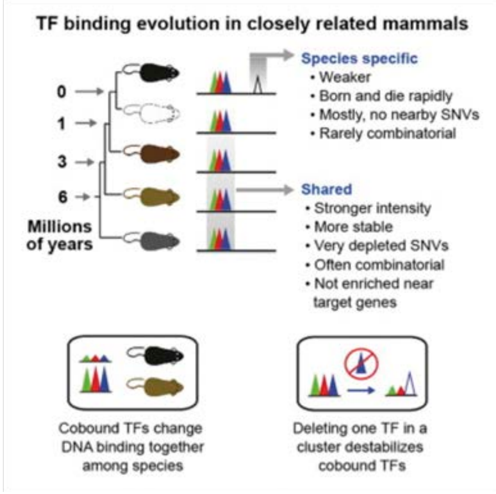
At EMBL-EBI since 2005. Team Leader since 2007, Senior Scientist since 2011.

Research

Our research into the rapid evolution of tissue-specific transcriptional regulation in closely related mammals, a project led by David Thybert, revealed how collections of transcription factors evolve rapidly (Figure 1). As has been shown in fruit flies, the inherent biophysics of transcription factor binding leads to cooperativity, mediated by the bound DNA. Bound regions can be highly stablised in binding intensities across evolution, even in the absence of functional signatures in nearby genes. Team alumna Petra Schwalie used the availability of similar cell lines from multiple primate species to explore how the only known vertebrate insulator, CTCF, is stabilised evolutionarily by its interaction with the co-binding transcription factor YY1. Graham Ritchie, whose ESPOD postdoc is shared with the Zeggini group at the Wellcome Trust Sanger Institute, explored systematic functional annotation of noncoding sequence variants.

Future plans

Genome sequencing and genome annotation is becoming increasingly relevant in clinical applications and ensuring that Ensembl's reference resources will be valuable for this application is an important challenge. High-throughput sequencing enables the sequencing of new species and the creation of new and important datasets for all of genomics. Ensembl will continue to address these needs with increasingly flexible methods of data access, presentation and distribution. The other resources of the Vertebrate Genomics Team are working with specific communities and specific projects to maximise the value of the data that they generate, and part of this effort will include integration of the data into appropriate archives and resources within the team. We will continue to interact closely with Justin Paschall's Variation team, Helen Parksinon's Samples, Phenotypes and Ontologies team, Paul Kersey's Nonvertebrate Genomics team and Guy Cochrane's European Nucleotide Archive team. We anticipate strengthening our collaboration with the Wellcome Trust Sanger Institute and clinical community through our involvement in the DECIPHER, the global alliance for the secure sharing of genomic information, and induced pluripotent Stem Cells projects such as HipSci and the European Bank for Stem Cells.



Interspecies differences in transcription factor (TF) binding are rarely caused by DNA variation in motifs; rather, combinatorial binding is significant for genetic and evolutionary stability. Cobound TFs tend to disappear in concert and were sensitive to genetic knockout of partner TFs.

The 1000 Genomes Project, which has collected data on more than 2500 genomes, will release its final integrated variant set in 2014. This will include short variants, simple large deletions, more complex multi-allelic variations and other structural variation classes such as duplications, inversions and nuclear mitochondrial insertions.

Our research efforts with the Odom group focus on exploring enhancer evolution across the mammalian space, and we will continue our efforts to elucidate the regulatory underpinnings of evolutionary diversity. We will also adapt the tools of classical genetics analysis to dissect the genomic mechanisms underlying transcriptional divergence in closely related mammals.

Selected publications

Flicek, P., Ahmed I, Amode MR, et al. (2013) Ensembl 2013. Nucleic Acids Res. 41, D48-55.

Seitan, V.C., Faure, A.J., Zhan, Y., et al. (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. 23, 2066-2077.

Khurana, E., Fu, Y., Colonna, V., et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342, 1235587.

Stefflova, K., Thybert, D., Wilson, M.D., et al. (2013) Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 154, 530–540.

Wang, Z., Pascual-Anaya, J., Zadissa, A., et al. (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nature Genet. 45, 701-706.

Non-vertebrate genomics

We provide tools supporting the exploration of genome-scale data to communities working in domains as diverse as agriculture, pathogen-mediated disease and the study of model organisms. Our team provides data and analysis for complete bacterial, protist, fungal, plant and invertebrate metazoan genomes, exploiting the power of the Ensembl software suite.

Our major areas of interest include broad-range comparative genomics and the visualisation and interpretation of genomic variation, which is being increasingly studied in species throughout the taxonomy. By collaborating with EMBL-EBI and re-using our established toolset, communities can store, analyse and disseminate data more cheaply and powerfully than if they develop their own tools. Our team helps small communities with little informatics infrastructure analyse, archive and disseminate highly complex and data-generative experiments—the type of work once the sole domain of large, internationally co-ordinated sequencing projects. We also work with international collaborators to provide major genome-centric databases in major areas of reserach. These. include VectorBase (Megy et al., 2012), a resource focused on the annotation of invertebrate vectors; WormBase (Yook et al., 2012), a resource for nematode biology; and PomBase (Wood et al., 2012), a resource focused on the fission yeast Schizosaccharomyces pombe. In the plant domain, we collaborate closely with Gramene in the United States and with a range of European groups in the transPLANT project. We have also launched PhytoPath, a new portal for plant pathogen data and are involved in the development of Microme, a new resource for bacterial metabolic pathways. An integrated view to all these data is provided through the Ensembl Genomes portal.

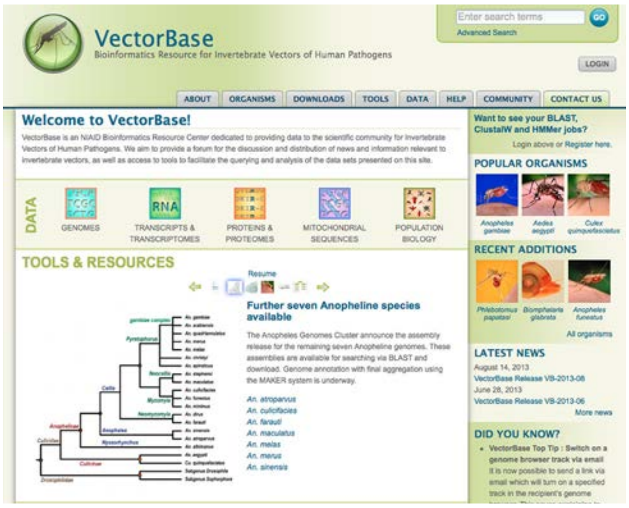
Major achievements

In 2013 we issued five public releases of Ensembl Genomes, and contributed to the regular data releases of VectorBase, WormBase and PomBase. The Ensembl Bacteria portal was completely overhauled and now offers interactive and programmatic access to every annotated bacterial genome submitted to the European Nucleotide Archive; there are 9765 such genomes in the current release.

We added 12 new invertebrate metazoan genomes to Ensembl Genomes, and were heavily involved in producing the primary annotation for two of these: the kissing bug (Rhodnius prolixus) and the myriapod Strigamia maratima. We are also involved in a project to annotate 16 different mosquito species, and expect to publicly release these data in early 2014. The work on Rhodnius and mosquitoes occurred in the context of the VectorBase project, which launched a completely redesigned website early in 2013, designed to the display larger numbers of genomes whose sequence is now available. In the WormBase project, we contributed to the annotation of Brugia malayi, the causative agent of lymphatic filarisis (elephantiasis).

In addition to these metazoan genomes, we added the genomes of 12 fungi, four protists, and four plants to the public release.

Ensembl Plants has been expanding in response to the increased availability of resources for cereal genomes. Most notably, the database now contains the sequence and annotation from the chromosomal survey sequence assembly recently released by the International Wheat Genome Sequencing Consortium. This comprises 10 Gb of assembled sequence, of an estimated total genome size of 16 Gb (by contrast, the human genome sequence is just ~3 Gb in size). The reference sequence is still fragmented and incomplete due to its highly repetitive (and polyploid) nature, which makes assembly difficult. Nonetheless, genic regions are relatively well described and Ensembl Plants provides



Welcome to VectorBase!

VectorBase is an NIAID Bioinformatics Resource Center dedicated to providing data to the scientific community for Invertebrate Vectors of Human Pathogens. We aim to provide a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented on this site.

Want to see your BLAST, ClustalW and HMMer jobs?

Login above or Register here.

POPULAR ORGANISMS

RECENT ADDITIONS

Further seven Anopheline species available

The Anopheles Genomes Cluster announce the assembly release for the remaining seven Anopheline genomes. These assemblies are available for searching via BLAST and download. Genome annotation with final aggregation using the MAKER system is underway.

An. atroparvus
An. culicifacies
An. farauti
An. maculatus
An. melas
An. merus
An. sinensis

LATEST NEWS

August 14, 2013
VectorBase Release VB-2013-08
June 28, 2013
VectorBase Release VB-2013-06

DID YOU KNOW?

- VectorBase Top Tip : Switch on a genome browser track via email
  It is now possible to send a link via email which will turn on a specified track in the recipient's genome

The new VectorBase website, re-implemented to provide easier navigation to larger numbers of genomes. Pictured is a notice of the pre-release of data from the 16 Anopheles genome project, nearing completion as of December 2013.

Paul Kersey

PhD University of Edinburgh 1992. Post-doctoral work at University of Edinburgh and MRC Human Genetics Unit, Edinburgh.

At EMBL-EBI since 1999.

access, on a per-gene basis, to evolutionary trees showing the relationship of each of the three wheat genomes to each other, as well as to other cereals and plants. It also presents genomic and EST alignment data, and the location of inter-homeologous variants.

We are active members of UK consortia that are generating improved assemblies and annotation for both wheat and barley, and continue to develop our resources for both species. In the context of the transPLANT project we have developed a pipeline and archive for the scalable management of variation data, which has now completed testing and is available for external submissions. This is a much-needed infrastructural component in the plant domain, and potentially applicable for a wider range of species.

Our team has also been working on annotating the genomes of other various species including the salmon louse (Lepeophtheirus salmonis), the biting midge (Culicoides sonoerensis), and the diatom (Phaeodactylim tricornutum). We intend to make these data publicly available during 2014. We have also started work (in collaboration with the ENA team) on the development of RNAcentral, a new resource for information about non-coding RNAs, which, in addition to its direct value to researchers working in this area, is also likely to prove highly useful for genome annotation

In 2013 the second Assemblathon paper was published, describing an international effort, to which we have contributed, to compare programs for genome assembly. Just as importantly, the paper defines methods by which such programs can be compared, and paves the way for better critical assessment of future genome assemblies.

Future plans

We anticipate the release of further improvements to the wheat and barley genomes in 2014, and will be developing a new range of comparative views to display genomic relationships among the grasses. We will start work on WormBase-ParaSite, an extension to WormBase focusing on the high-throughput annotation and analysis of the genomes of parasitic worms (including flatworms, an extension of WormBase's current scope). We will also release the 16 Anopheles genomes annotated by VectorBase.

Our team will be identifying and acting on increasing needs for more powerful tools for data mining and data extraction, particularly across multiple species. The new solutions already in development are expected for release in 2015.

Selected publications

Harris, T.W., Baran, J., Bieri, T., et al. (2014). WormBase 2014: new views of curated biology. Nucleic Acids Research 42.

Kersey, P.J., Allen, J.E., Christensen, M., et al. (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Research 42.

Monaco, M.K., Stein, J., Naithani, S., et al. (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Research 42

The i5K Consortium (2013). The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. Journal of Heredity 104 595-600.

Neafsey, D.E., Christophides, G.K., Collins, F.H., et al. (2013). The Evolution of the Anopheles 16 Genomes Project. G3 3:1191-4

Bradnam, K.R., Fass, J.N., Alexandrov, A., et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2013, 2:10

Nayduch, D., Cohnstaedt, L.W., Saski, C. et al. (2013) Studying Culicoides vectors of BTV in the post-genomic era: Resources, bottlenecks to progress and future directions Virus Research.

Variation

Free global access to genetic variation data is critical to advances in knowledge exchange in both research and clinical applications. The Variation team develops and maintains database archive resources for genetic variation in all species, both germline and somatic. We make these data available to the scientific community directly and through other EMBL-EBI services.

Our resources include the Database of Genomic Variants archive (DGVa), a resource of structural variation; the European Variation Archive (EVA), scheduled for launch in spring 2014; and the controlled-access European Genome-phenome Archive (EGA), focused on securely archiving and distributing human biomedical research data.

Major achievements

In its second year, the Variation team achieved considerable progress in the EGA, a resource that plays a central role in two ELIXIR pilot projects. With the creation of the global alliance for the secure sharing of genomic data in 2013, our team's activities came into sharp focus and we endeavoured to produce a 'gold standard' in secure provision of genomics-based human biomedical research data.

The EGA grew considerably during 2013, hosting 415 studies totalling over 800 terabytes of controlled-access, human biomedical research data. The team performed major upgrades to the EGA service, in collaboration with the Web Development team. The service now features fully automated data distribution and upload via a custom Java client, which handles both encryption and data transfer. EGA submission and access became more user friendly as a result, and data security was improved.

To make archived data useful, we ensure that metadata is well structured and rich at the study, sample, phenotype and population level. This allows relevant datasets to be easily discovered and applied to a variety of scientific questions. Recent efforts in the EGA project have sought to provide better curation, structured queries, and summary statistics of available datasets by phenotype, technology and study design. EGA functionality has been improved to allow users to browse and query studies, datasets and samples more intuitively.

We conducted an ELIXIR pilot study with CSC IT Centre for Science in Finland on federated authentication and authorization of EGA. This project delivered a framework for connecting trusted institutional logins to EGA accounts. This simplifies user access and ensures the highest possible security. We also improved the administration of permissions to EGA studies by external Data Access Committees (DACs) by upgrading security and developing tools for hierarchical management of related studies.

The most radical change for EGA in 2013 was the ramping up of another ELIXIR effort: the expansion of EGA to include resources developed and hosted at the Centre for Genomic Regulation (CRG) in Barcelona, Spain. This project is expected to provide a model for distributed, secure access to biomedical research data.

DGVa and EVA are richly annotated archives of genetic variation that provide well-structured data for direct use and incorporation into other resources. Our team ensures that entries include sufficient data and metadata for discoverability, generation of summary statistics and harmonisation across studies. In 2013, together with our peer database dbVar in the United States, DGVa grew substantially to host 118 studies across 31 studies, representing over

14 million structural variant observations and over 15 million genotypes.

We devoted considerable efforts to developing the EVA in 2013, and expect to launch the resource in 2014. EVA already incorporates over 40 million smaller substitutions and indels from large population surveys such as the 1000 Genomes Project and Genome of the Netherlands. EVA continues to receive increasing numbers of rare variant data submissions of clinical interest.

Future plans

In 2014 we will devote considerable effort to the launch of the EVA, which builds on the DGVa structural variation database to include short genetic variation. The purpose of the EVA is to provide users with the most comprehensive view of known genetic variation on a genome-wide scale across human populations. This work will be carried out in collaboration with global partners, including dbSNP and dbVar in the United States.

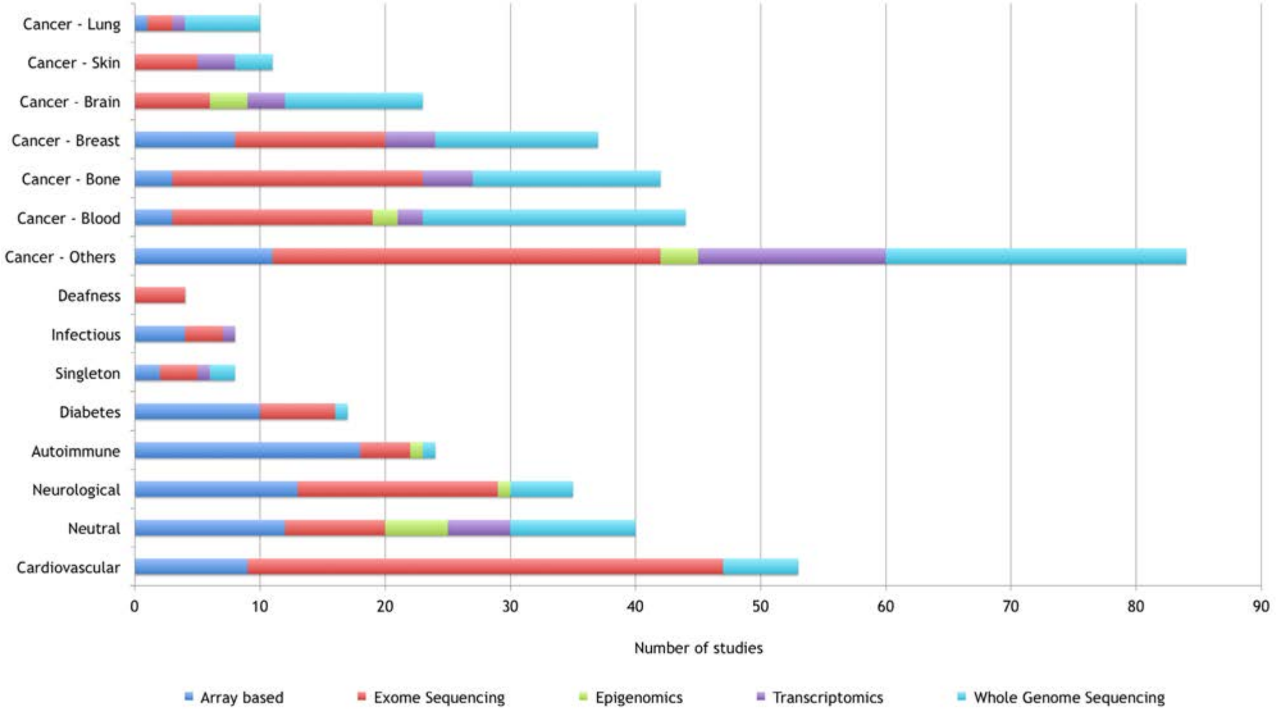
We will continue to improve the EGA in terms of discoverability, ease of submission and data access. A key area of focus over the next two years will be the implementation of additional modes of access beyond simple file download. Infrastructure is now in place to allow genomic-position-based slices of data to be downloaded by users who wish to examine only a specific set of genomic regions. This will lower the barrier of entry for access to these data as labs will not require the necessary large

Justin Pascall

MA 2008, Washington University S. Louis.

Variation Team Leader at EMBL-EBI since 2012.

local disk storage to download entire datasets at once. We will also work on a prototype for easier access using the Embassy Cloud compute facility to provide users with access to EGA data within a secure compute environment hosted at EBI. This will, we hope, remove the need to download large datasets for certain applications.



European Genome-phenome Archive (EGA) studies by technology, assay type and phenotype studied.

Molecular atlas

Our resources that focus on RNA, protein and metabolite expression data are working towards creating a comprehensive, integrated and scalable Atlas of expression. We hope that 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 2013 we released a new baseline Expression Atlas, which now incorporates both microarray and RNAseq data on gene expression in healthy individuals. Many researchers deposit their data directly into ArrayExpress, and we have now made the process simpler with a more intuitive interface. We also played a key role in bringing the proteomics community together around shared data standards, gaining widespread support for the new ProteomeXchange platform.

ArrayExpress

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

www.ebi.ac.uk/arrayexpress

Expression Atlas

The Expression Atlas is an added-value database for the reanalysis of gene expression data from EBI resources such as ArrayExpress. It shows which genes are expressed under different conditions, and how expression differs between conditions. The Expression Atlas currently holds RNA expression data from microarray or RNAseq experiments, and future development will include protein and metabolite expression data.

www.ebi.ac.uk/gxa

PRIDE

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

www.ebi.ac.uk/pride

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

Alzis Brazma

- *Released the new baseline Expression Atlas, which represents high-throughput-sequencing-based expression data on gene expression levels in healthy, untreated conditions;*
- *Contributed to the prototype database for data generated by the Systems Microscopy Network of Excellence;*
- *Co-developed the cellular microscopy phenotype ontology (CMPO) in close collaboration with the Samples, phenotypes and ontologies team.*

www.ebi.ac.uk/gxa
www.ebi.ac.uk/arrayexpress

Henning Hermjakob

- *Released the PRIDE cluster algorithm, used to score peptide identifications.*

www.ebi.ac.uk/pride

Helen Parkinson

- *Delivered the Gene Expression Atlas RDF and BioSamples RDF, along with a related BioConductor package/AtlasRDF that enables enrichment analysis of RDF data;*
- *Developed new infrastructure for the BioSamples database, including a new user interface and programmatic access.*

www.ebi.ac.uk/rdf
www.ebi.ac.uk/gxa
www.ebi.ac.uk/biomodels
www.ebi.ac.uk/biosamples

Ugis Sarkans

- *Redeveloped the ArrayExpress submission system to incorporate the Annotare tool for wizard-style functionality (to replace MIAMExpress in early 2014);*
- *Built an externally accessible release-date management system that submitters can use to update their submissions and manage submission details such as associated publications;*
- *Deployed a generic biological study data management solution for the diXa toxicogenomics data warehouse;*
- *Worked on the basic data infrastructure for EU-AIMS, a five-year, EU-funded project centred on autism-spectrum disorder studies.*

www.ebi.ac.uk/arrayexpress
www.dixa-fp7.eu
www.eu-aims.eu

Christoph Steinbeck

- *Issued a release of MassCascade, a workflow plug-in for processing metabolomics liquid chromatographic mass spectrometry data;*
- *Released a new version of MetaboLights featuring a new design, enriched knowledge base, improved study-upload queuing system and improved download capabilities.*

www.ebi.ac.uk/chebi
www.ebi.ac.uk/metabolights

Explore data on Paramecium species in EMBL-EBI resources, including Array Express and BioSamples.
Photo courtesy of Arturo Agostino

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. We collaborate closely with the Marioni group and others throughout EMBL on research projects that focus on cancer genomics and transcript isoform usage.

Major achievements

Our primary focus in 2013 was on developing the baseline Expression Atlas, which extends the resource to represent high-throughput-sequencing-based expression data on gene expression levels in healthy, untreated conditions. The new database, coordinated by Robert Petryszak, was released in November, and an article about it was published in Nucleic Acid Research (Petryszak et al., 2013). The Expression Atlas, a value-added database, provides information about gene, protein and splice-variant expression in different cell types, organs, developmental stages and diseases under various normal and experimental conditions. The resource now comprises selected high-quality microarray and RNA-sequencing experiments from ArrayExpress that have been processed using standardised microarray and RNA-sequencing analysis methods.

We contributed to the development of the BioSamples database (Falconbridge et al., 2013) and we continued the development of the prototype database for data generated by the Systems Microscopy Network of Excellence. We also co-developed the cellular microscopy phenotype ontology (CMPO) in close collaboration with the Samples, phenotypes and ontologies team.

In 2013 we organised and participated in 30 training events, including Bioinformatics Roadshows and on-site courses. These included the EMBO practical course on the analysis of high-throughput sequencing data, which was the most popular and oversubscribed EMBL-EBI training event in 2013.

We completed the analysis of human gene-expression datasets in order to establish the transcript composition (Gonzales-Porta et al., 2013). We showed that, in a given condition, most protein-coding genes have one major transcript expressed at significantly higher level than others. We also found that in human tissues, the major transcripts contribute almost 85% to the total mRNA from protein-coding loci, and that often the same major transcript is expressed in many tissues.

In 2013 we completed our analysis of differential transcript expression and fusion genes in the RNAseq data generated by the EU-funded GEUVADIS consortium (Lappalainen et al., 2013). In addition, as part of our involvement in the EurcanPlatform consortium, we participated in the identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing (Majewski et al., 2013). In the CAGEKID consortium, we completed differential gene and transcript expression and fusion gene analysis in 50 pairs of cancer and normal tissues in kidney cancer.

Future plans

One of the major tasks in 2014 will be streamlining gene-expression submissions through ArrayExpress to Expression Atlas. Automating the handling of ArrayExpress submissions to a greater extent will allow us to devote more time to the effective selection, curation, quality control and analysis of the most valuable and relevant microarray and RNA-seq data sets, both mRNA and smallRNA. We will also work on improving the searchability and presentation of the data in Expression Atlas, including ontology-driven search, faceting and biological data visualisation.

Our research will focus on large-scale data integration and systems biology. With our colleagues at the International Cancer Genome Consortium, we will investigate the impact of cancer genomes on functional changes in cancer development and explore fusion genes and their role in cancer development.

Alvis Brazma

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

At EMBL-EBI since 1997.

Selected publications

Petryszak, R., Burdett, T., Fiorelli, B., et al. (2013) Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 42, D926-D932.

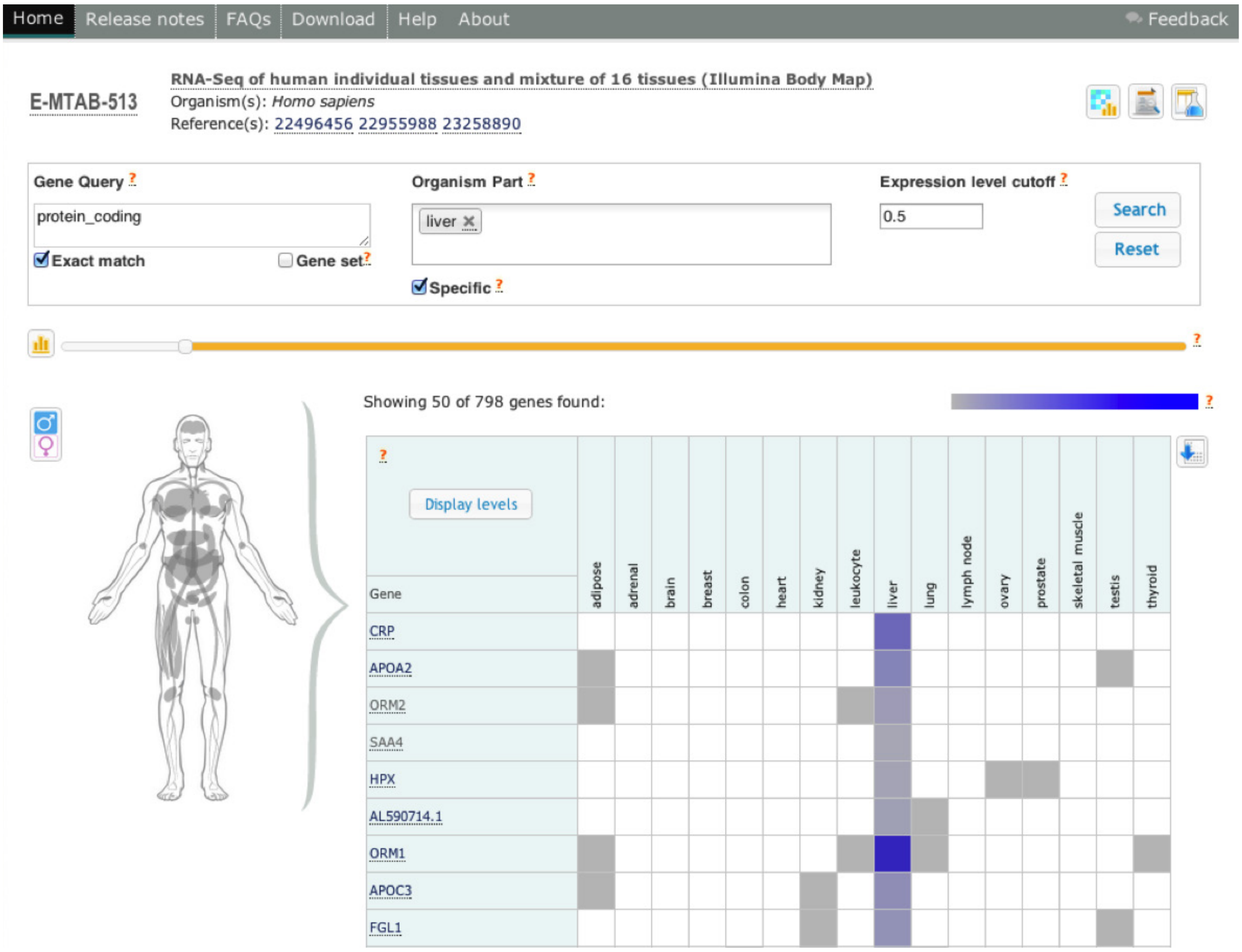
Faulconbridge, A., Burdett, T., Brandizi, M., (2013) Updates to the BioSamples database at the European Bioinformatics Institute. Nucleic Acids Res. 42, D50-D52.

Lappalainen, T., Sammeth, M., Friedländer, M.R., et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.

Gonzàlez-Porta, M., Frankish, A., Rung, J., et al. (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 14, R70

Majewski, I.J., Mittempergher, L., Davidson, N.M., et al. (2013) Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. J. Pathol. 230, 270-276.

Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. Nat. Reviews Genet. 14, 89-99.



E-MTAB-513

RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)
Organism(s): *Homo sapiens*
Reference(s): 22496456 22955988 23258890

Gene Query ?: protein_coding
☑ Exact match
□ Gene set ?

Organism Part ?: liver ✕
☑ Specific ?

Expression level cutoff ?: 0.5

Showing 50 of 798 genes found:

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
CRP									🟦							
APOA2	▨								🟦						▨	
ORM2	▨							▨	🟦							
SAA4									🟦							
HPX									🟦			▨	▨			
AL590714.1									🟦	▨						
ORM1	▨							▨	🟦	▨						▨
APOC3	▨						▨		🟦							
FGL1							▨		🟦						▨	

Prototype of the baseline Expression Atlas.

Functional genomics development

Our team develops software for ArrayExpress, a core EMBL-EBI resource. A new project we started in 2013 is the BioStudies database, a resource for biological datasets that do not have a dedicated home within EMBL-EBI services.

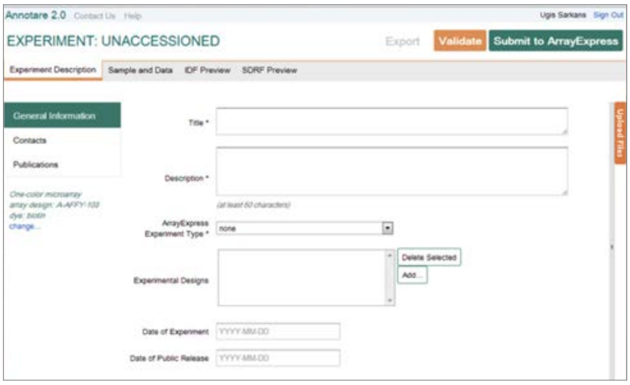
We also contribute to the development of the EBI BioSamples database, which centralises biological sample data. We build and maintain data management tools, user interfaces, programmatic interfaces, and annotation and data submission systems. We also collaborate on a number of European ‘multi-omics’ and medical informatics projects in a data-management capacity.

Major achievements

In 2013 the user interface of ArrayExpress underwent a significant redesign, adopting the EMBL-EBI web guidelines. These include uniform search-box behaviour, ‘cleaner’ results filtering and a common search-result presentation paradigm. We continued to improve the BAM file-generation system for sequencing-based experiments.

We devoted major efforts to redeveloping our submission system. With the new tool, Annotare, we have attempted to strike the right balance between providing wizard-style functionality for ease of use, and the spreadsheet-like interface to enable larger submissions. Annotare will replace our old submission system, MIAMExpress, in early 2014. We also built an externally accessible release-date management system that submitters can use to update their submissions; this can be extended to manage other details of a submission, such as associated publications.

We continued to collaborate with the Samples, Phenotypes and Ontologies team on the basic infrastructure components of the BioSamples database. In 2013 we finished an RDBMS-based implementation, together with tools for data conversion and identifier mapping. Like ArrayExpress, the BioSamples user interface was adjusted to follow the EMBL-EBI website guidelines. A range of scalability issues across all the software components was tackled, and now BioSamples database is ready to deal with tens of millions of samples.



Annotare: a new submission tool for expression data.

Our work on the diXa toxicogenomics data warehouse resulted in a generic biological study data management solution that will be repurposed for the needs of the BioStudies database.

We integrate genomics and imaging data for EU-AIMS, a five-year, EU-funded project centred on autism-spectrum disorder studies. In 2013 we worked on the basic data infrastructure, consulting with project partners to build the data acquisition and task-tracking system.

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

ArrayExpress

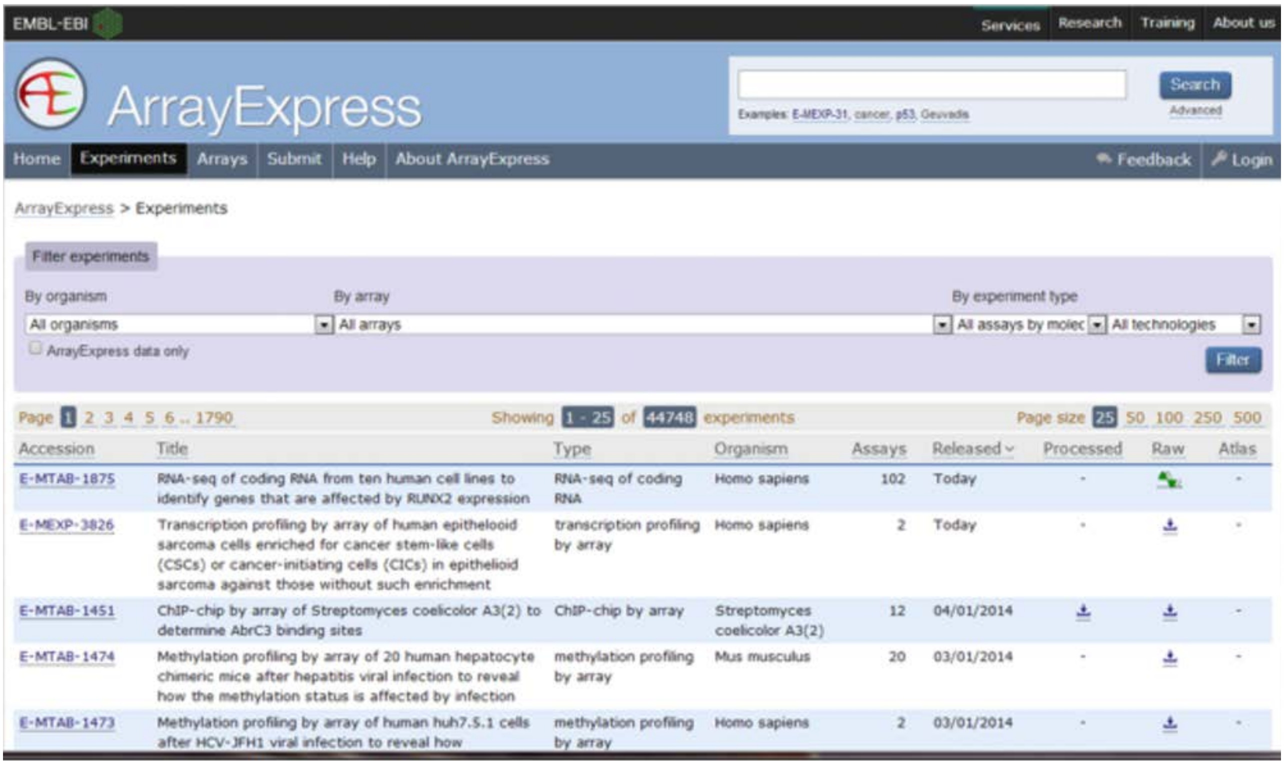
In 2014 we will further minimise the time spent by curators on various data processing tasks, for example by giving simple tools to data submitters to supplement and modify their submissions. After the new user interface has run for a year, we will perform another round of improvements, concentrating on better support for sequencing data. We will deploy the new data acquisition tool Annotare to production and integrate it into ArrayExpress data pipelines.

BioStudies

Our BioStudies resource will evolve from our work on the diXa project. In 2014 we will provide a lightweight structure for metadata describing studies and links to the actual data, and ensure that we can populate the BioStudies database via this format. We will also establish a simple user interface for searching and browsing study data. We are also exploring possibilities of positioning BioStudies database as a target system for submitting supplementary data associated with publications.

Medical informatics

In EU-AIMS we will work on integrating genomic and imaging data. We are getting involved in the European Medical Information Framework (EMIF), a new medical informatics project that aims to develop an information framework for access to diverse medical and research data sources. In this project we will deal with genomics data analysis and visualisation.



All organisms ▾ All arrays ▾ All assays by molec ▾ All technologies ▾

□ ArrayExpress data only

Page 1 2 3 4 5 6 .. 1790 Showing 1 - 25 of 44748 experiments Page size 25 50 100 250 500

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-1875	RNA-seq of coding RNA from ten human cell lines to identify genes that are affected by RUNX2 expression	RNA-seq of coding RNA	Homo sapiens	102	Today	-	⬇	-
E-MEXP-3826	Transcription profiling by array of human epithelioid sarcoma cells enriched for cancer stem-like cells (CSCs) or cancer-initiating cells (CICs) in epithelioid sarcoma against those without such enrichment	transcription profiling by array	Homo sapiens	2	Today	-	⬇	-
E-MTAB-1451	ChIP-chip by array of Streptomyces coelicolor A3(2) to determine AbrC3 binding sites	ChIP-chip by array	Streptomyces coelicolor A3(2)	12	04/01/2014	⬇	⬇	-
E-MTAB-1474	Methylation profiling by array of 20 human hepatocyte chimeric mice after hepatitis viral infection to reveal how the methylation status is affected by infection	methylation profiling by array	Mus musculus	20	03/01/2014	-	⬇	-
E-MTAB-1473	Methylation profiling by array of human huh7.5.1 cells after HCV-JFH1 viral infection to reveal how	methylation profiling by array	Homo sapiens	2	03/01/2014	-	⬇	-

ArrayExpress in 2013: new layout follows EMBL-EBI web guidelines.

Proteins and protein families

UniProt, the unified resource of protein sequence and functional information, is closely integrated with Ensembl and Ensembl Genomes and in 2013 incorporated variation data from the 1000 Genomes Project. The universal protein resource is also working to incorporate proteomics data in their knowledgebase.

Other integration in the proteins area included the addition of a DNA search interface in the Pfam resource for protein families, and search functions for proteins and chemical structure in the Enzyme Portal. In 2013 considerable progress was made towards the creation of a 'gold standard' dataset to help users identify all experimental data for a given protein from a particular strain of a given organism.

The high-performance InterProScan tool was the most heavily used at EMBL-EBI in 2013, and its new release in 2013 was faster, more reliable and easier to search, allowing users to optimise their query for response time or throughput.

UniProt

UniProt is a collaboration among EMBL-EBI, the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (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 crossreferences and querying interfaces.

The work of our team 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 Metagenomic and Environmental Sequences (UniMES) database serves researchers who are exploring the rapidly expanding area of metagenomics, which encompasses both health and environmental data;*
- *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

Enzyme Portal

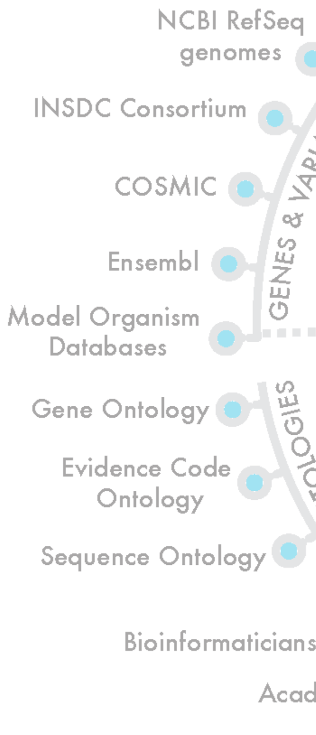
The Enzyme Portal provides integrated enzyme-related data for all EBI enzyme resources as well as the underlying functional and genomic data.

www.ebi.ac.uk/enzymeportal

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.

www.merops.sanger.ac.uk



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.sanger.ac.uk

Treefam

TreeFam (Tree families database) is a database of phylogenetic trees of animal genes. It is a curated resource that aims to provide reliable information about ortholog and paralog assignments, as well as an evolutionary history of various gene families.TreeFam is also an ortholog database. It fits a gene tree into the universal species tree and finds historical duplications, speciations and loss events.

www.treefam.org

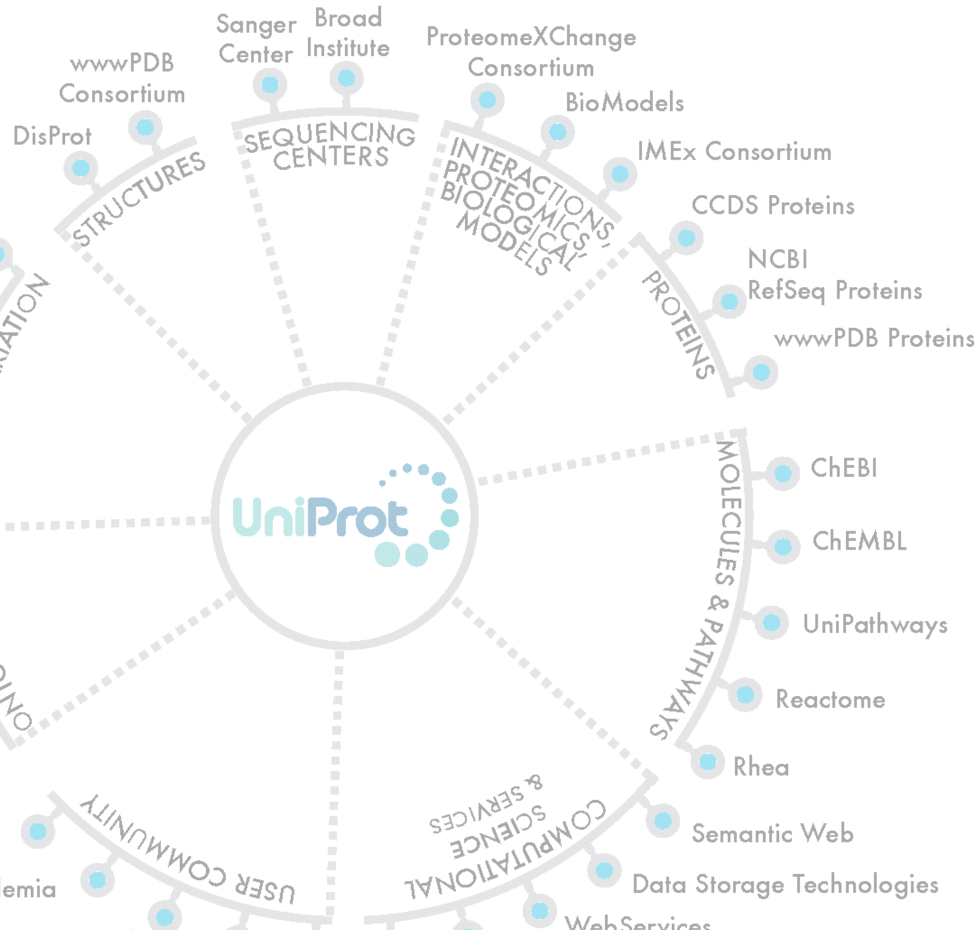
InterPro

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



Proteins and protein families

Alex Bateman

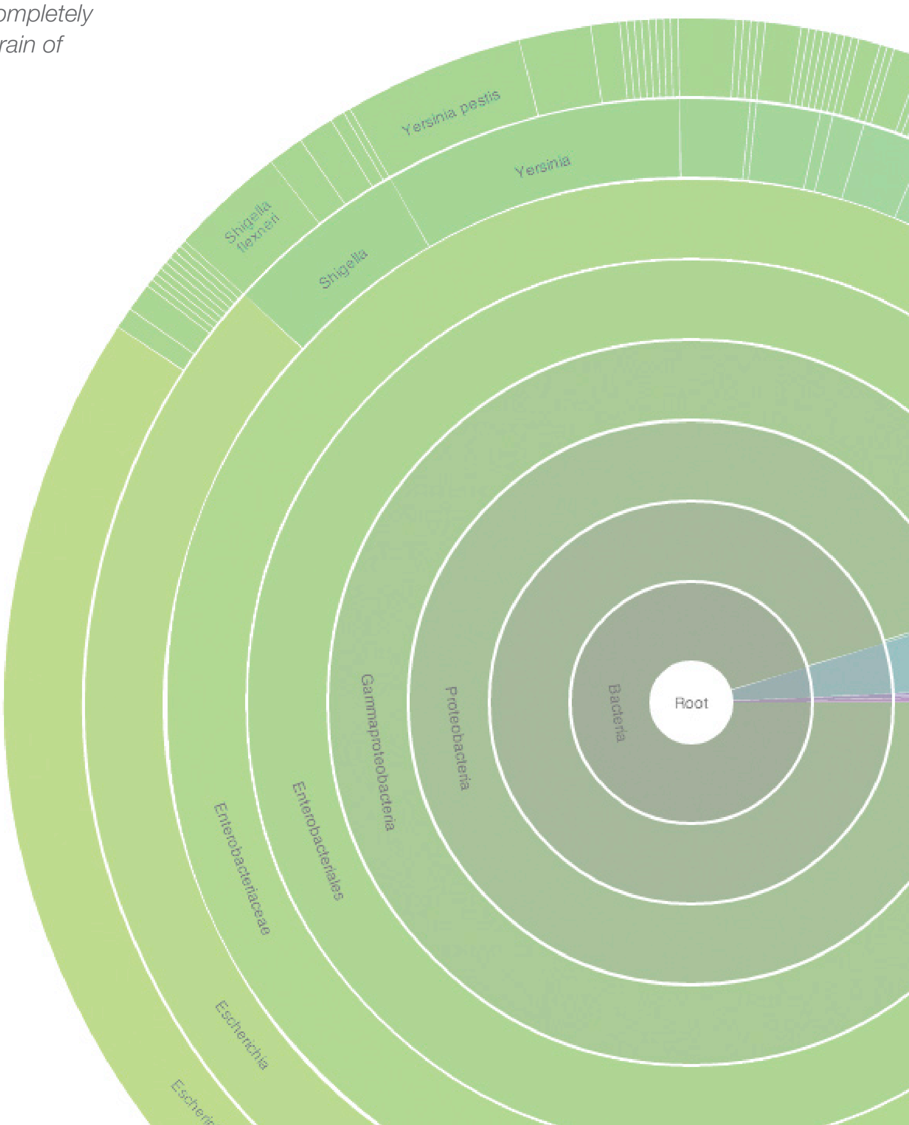
- *Maintained Pfam sequence coverage of the UniProt Knowledgebase at nearly 80%, despite a great increase in the size of the underlying sequence database;*
- *Increased the sequence coverage of human sequences in Pfam by almost 5%;*
- *Provided a new interactive DNA search interface for Pfam;*
- *Issued TreeFam release 9, the first release at EMBL-EBI, providing an HMM-based sequence search that places a user-provided protein sequence into a TreeFam gene tree and provides rapid orthology prediction;*
- *Presented a new approach to visualising gene trees and alternative displays that focuses on showing homology information from a species-tree point of view;*
- *Developed MEROPS to include a community annotation project in which acknowledged experts are invited to contribute summaries for peptidases;*
- *Enabled the filtering of peptidase lists from a completely sequenced bacterial genome for a particular strain of the organism.*

http://pfam.sanger.ac.uk
www.treefam.org
http://merops.sanger.ac.uk

Sarah Hunter

- *Released InterProScan5, which features improved running speed, searchability, reliability and stability;*
- *Reconfigured InterProScan5 to allow users to optimise for response time or throughput, and moved the source code to be hosted at Google Code;*
- *Increased InterPro coverage of UniProtKB proteins by performing significant data curation and integration efforts;*
- *Moved the main InterPro website to the London Data Centres, backed by a re-written data warehouse structure, which implified the web application code necessary to access the data.*

www.ebi.ac.uk/interpro
www.ebi.ac.uk/Tools/pfa/iprscan5



Maria Martin

- *Developed new interfaces and tools for analysing and accessing UniProt data, focusing on user interaction with the website;*
- *Provided a non-redundant collection of 594 well-annotated reference proteomes, representing the taxonomic diversity in UniProtKB;*
- *In collaboration with Ensembl, incorporated variation data from the 1000 Genomes Project;*
- *In collaboration with PRIDE, analysed proteomics data for incorporation into UniProtKB;*
- *Extended Protein2GO, the common annotation tool for the GO Consortium, to include new functionalities;*
- *Developed an annotation-scoring system that represents the quality and depth of information present in UniProtKB records.*

www.uniprot.org

Claire O'Donovan

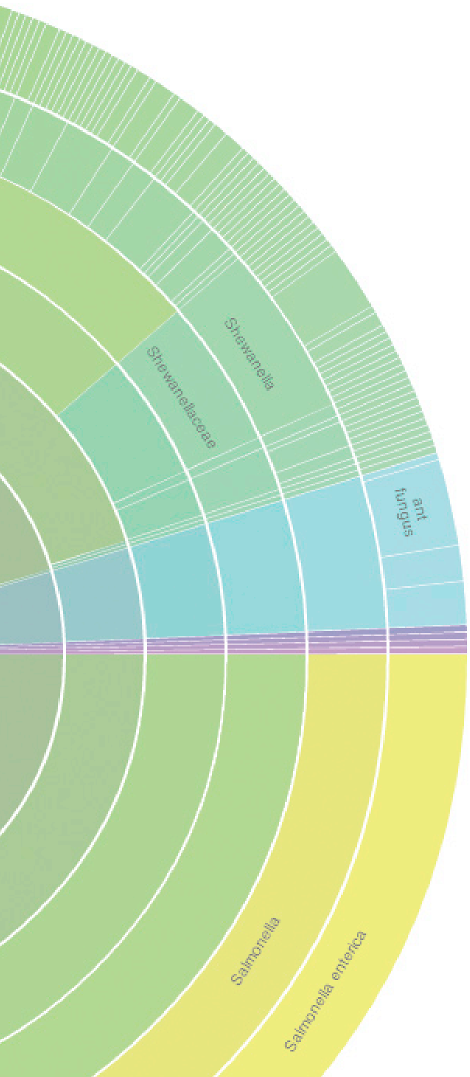
- *As part of the Consensus CDS (CCDS) project, worked on an authoritative complete proteome set for Homo sapiens, in part by ensuring a curated and complete synchronisation with the HUGO Gene Nomenclature Committee (HGNC);*
- *Played a major role in establishing minimal standards for genome annotation across the taxonomic range, largely thanks to collaborations arising from the annual NCBI Genome Annotation Workshops;*
- *Made substantial progress on a 'gold standard' dataset to help users identify all experimental data for a given protein from a particular strain of a given organism, with a focus on particular proteomes and protein families;*
- *Continued to provide high-quality annotations for human proteins as part of the GO Consortium Reference Genomes Initiative.*

www.uniprot.org

Christoph Steinbeck

- *Updated the Enzyme Portal web application by adding protein sequence search, chemical structure search and enzyme comparisons;*

www.ebi.ac.uk/enzymeportal



Protein sequence resources

The Protein sequence resources team sets out to classify protein and RNA sequences to build ‘periodic tables’ of these molecules for biologists. The classifications are embodied in several databases that the group develops, including Pfam, Rfam, TreeFam and MEROPS.

The group moved to EMBL-EBI at the end of 2012 and has begun to integrate with other groups and resources throughout the institute. After 15 years of overseeing the development of these resources, Alex Bateman welcomes Rob Finn as his successor in 2014. Rob Finn will also be leading the InterPro and EBI Metagenomics resources, thus bringing all the Protein and RNA Family resources at EMBL-EBI under his leadership.

Major achievements

Pfam is a widely used database of protein families, containing 14 831 manually curated entries in the current release, version 27.0 (Finn et al., 2013). In 2013 we maintained our sequence coverage of the UniProt Knowledgebase (UniProtKB) at nearly 80%, despite a great increase in the size of the underlying sequence database. We increased the sequence coverage of human sequences by almost 5%. We now provide family alignments based on four different representative proteome sequence datasets and a new interactive DNA search interface.

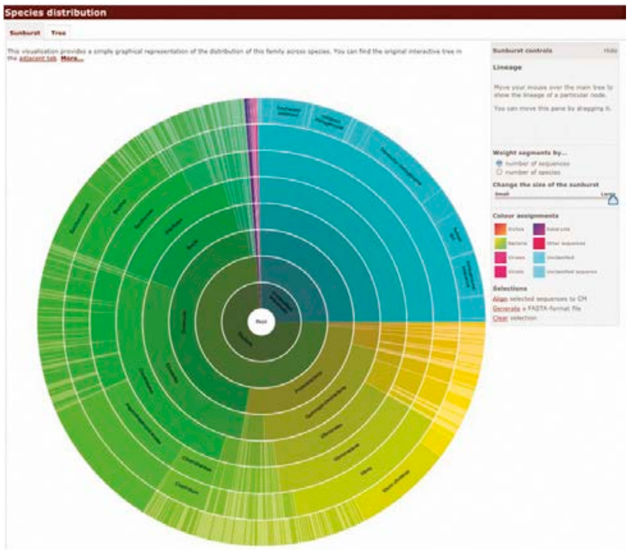
TreeFam is a database of phylogenetic trees inferred from animal genomes. For every TreeFam family we provide homology predictions together with the evolutionary history of the genes. The TreeFam project was resurrected in 2012 and has seen two releases since. The latest release (TreeFam 9) was made available in March 2013 (Schreiber et al., 2014). It has orthology predictions and gene trees for 109 species in 15 736 families, covering 2.2 million sequences. With release 9 we now provide an HMM-based sequence search that places a user-provided protein sequence into a TreeFam gene tree and provides rapid orthology prediction. We also present a new approach to visualising gene trees and alternative displays that focuses on showing homology information from a species-tree point of view. From release 9, the TreeFam website has been hosted at EMBL-EBI.

Peptidases, their substrates and inhibitors are of great relevance to biology, medicine and biotechnology. The MEROPS database aims to fulfil the need for an integrated source of information about these (Rawlings et al., 2013). Recent developments include a community annotation project in which acknowledged experts are invited to contribute summaries for peptidases. Contributors are acknowledged on the relevant web page. It is now possible to filter the list of peptidases from a completely sequenced bacterial genome for a particular strain of the organism.

The Rfam database is a collection of non-coding RNA families, primarily RNAs with a conserved RNA secondary structure, including both RNA genes and mRNA cis-regulatory elements. Each family is represented by a multiple sequence alignment, predicted secondary structure and covariance model. The latest release, Rfam 11.0, introduced genome-based alignments for large families, the introduction of the Rfam Biomart as well as other user interface improvements.

Future plans

In 2014 we will complete the move of all Pfam, Rfam and MEROPS software and hardware infrastructure and websites to EMBL-EBI. We will also coordinate the TreeFam set of gene trees and releases with Ensembl Compara. Rfam will continue to rebuild all its families using the new, faster Infernal 1.1 software. These improvements in speed will open up new application areas, for example identifying non-coding RNAs in metagenomic data sets.



This visualization provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab.

Move your mouse over the main tree to show the lineage of a particular node.

You can move this pane by dragging it.

Weight segments by...

● number of sequences

○ number of species

Change the size of the sunburst

Align selected sequences to CM

Generate a FASTA-format file

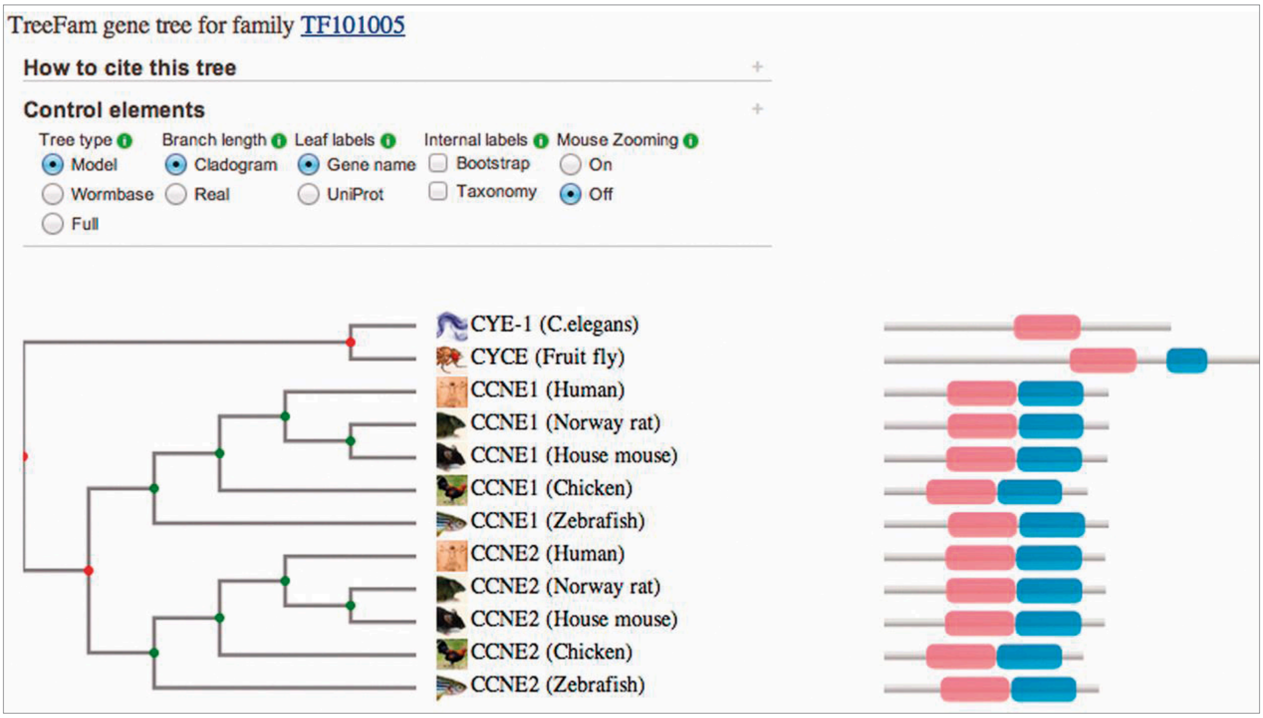
Clear selection

New Rfam visualisation showing the species distribution of the cyclic di-GMP-I riboswitch.

Alex Bateman

PhD 1997, University of Cambridge. Postdoctoral work at the Sanger Centre. Group Leader at Wellcome Trust Sanger Institute 2001-2012.

Head of Protein Sequence Resources at EMBL-EBI since 2012.



TreeFam gene tree for family TF101005

How to cite this tree +

Control elements +

Tree type: ● Model ○ Wormbase ○ Full

Branch length: ● Cladogram ○ Real

Leaf labels: ● Gene name ○ UniProt

Internal labels: □ Bootstrap □ Taxonomy

Mouse Zooming: ○ On ● Off

New TreeFam visualisation showing only model organisms for the Cyclin E family.

Selected publications

Finn, R.D., Bateman, A., Clements, J., et al. (2014) Pfam: the protein families database. Nucleic Acids Res. 42, D222-D230.

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, D922-D925.

Rawlings, N.D., Waller, M., Barrett, A.J. and Bateman, A. (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 42, D503-D509.

Burge SW, Daub J, Eberhardt R, et al. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41, D226-D232.

InterPro

Our team co-ordinates the InterPro and EBI Metagenomics projects at EMBL-EBI. InterPro integrates protein data from 11 major sources, classifying them into families and predicting the presence of domains and functionally important sites.

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.

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, a resource established at EMBL-EBI in 2011, enables metagenomics researchers to submit sequence data and associated descriptive metadata to the public nucleotide archives. Deposited data is subsequently functionally analysed using an InterPro-based pipeline, taxonomically analysed using the QIIME software package, and the results generated are visualised via a web interface.

Major achievements

InterPro

InterProScan5 was officially released in 2013, and is now the main InterPro scanning software hosted by EMBL-EBI. We extensively refactored the software to improve running speed, reliability and stability. We also improved configuration substantially, allowing users to optimise for response time (for small numbers of sequences) or throughput (for large numbers of sequences). We moved the source code to be hosted at Google Code as a first step towards InterProScan becoming a fully open-source software-development project.

The official release of InterProScan 5 in 2013 provided improvements to the following features:

- *In addition to searching all 11 member databases, four different algorithms can now be applied: Phobius (for predicting transmembrane topology and signal peptides), TMHMM (for predicting transmembrane helices in proteins), Coils (for predicting coiled-coils) and SignalPv4 (for predicting signal peptide cleavage sites);*
- *InterPro results can be used to predict potential membership of a protein in a pathway;*
- *A BerkeleyDB-based protein-match look-up service reduces calculation overheads by only searching sequences not already found in UniProtKB;*
- *Output formats now include HTML, GFF3, XML, TSV and SVG;*
- *The software can be run 'out of the box' on any Linux machine with minimal configuration, utilising cluster-queuing technologies;*
- *Scale-up: the software can be run efficiently on a single machine or on clusters with tens of thousands of machines;*
- *Both protein and nucleotide sequences can be processed, with results mapped back to the original sequence.*

The InterPro database now offers improved coverage of UniProtKB proteins, increasing to 81.9% in the latest release (v. 45.0), compared to 80.8% in 2012. This is partly due to significant data curation and integration efforts, which led to an additional 2143 signatures being incorporated into the database in 2013. Focussed curation of InterPro2GO term associations led to 1378 entries being assigned new or updated GO terms; 45% of entries now have at least one term associated. The total number of GO mappings increased by 1488 overall.

In 2013, the main InterPro website and was moved to the London Data Centres, backed by a re-written data warehouse structure. This simplified the web application code necessary to access the data.

EBI Metagenomics

EBI Metagenomics reached 40 public metagenomics projects in 2014, comprising 1309 separate samples and a significant number of privately held studies. The total number of raw nucleotide reads processed by the resource passed 25 billion.

Sarah Hunter

MSc University of Manchester, 1998.
Pharmaceutical and Biotech Industry (Sweden), 1999–2005.

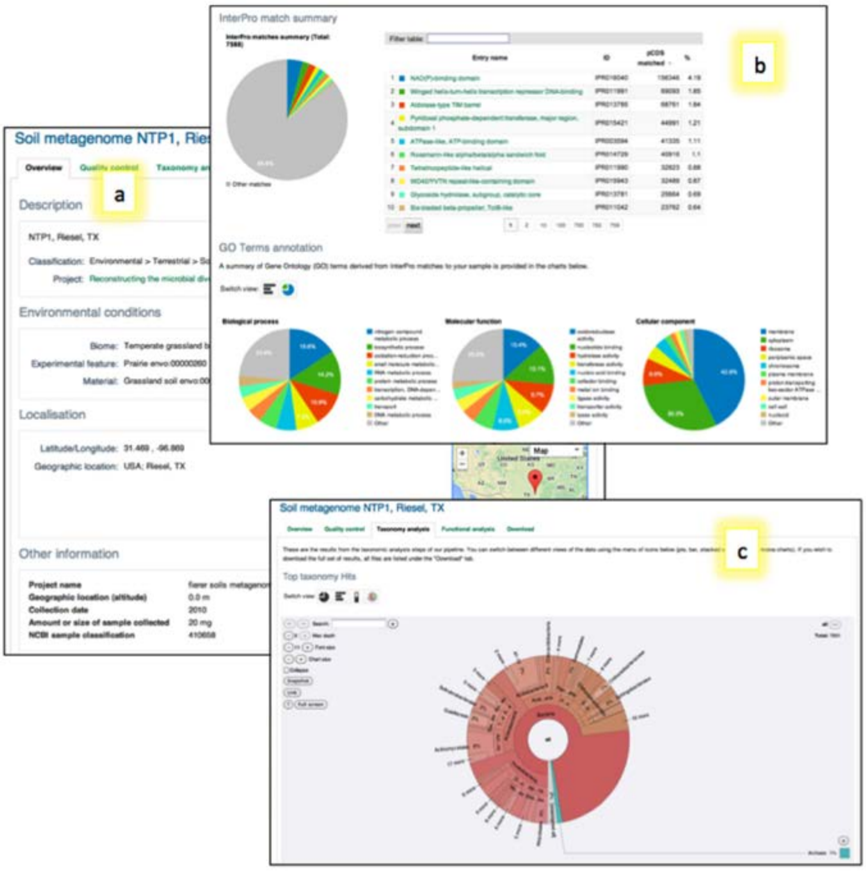
At EMBL-EBI since 2005. Team Leader since 2007.

Our work on the organisation and display of data on the website has made it easier for users to visualise analysis results. We developed EBI Metagenomics taxonomic and functional analysis pages to allow interactive visualisation of information as pie charts, bar charts or with the open source Krona viewer, where appropriate (see Figure).

Future plans

In 2014 InterPro will join the new Protein Families team, which will also include the Pfam database. There will be a period of transition as the two resources start to align strategic goals, particularly in regards to improving communication, data exchange and software. We will also work to make the data contained within the InterPro website more discoverable by expanding our online search tools and improving the navigability of the site. In addition, we will redevelop our internal pipelines to use InterProScan5, thereby streamlining our database production and ensuring data consistency.

For EBI Metagenomics, we will investigate tools for enhancing the current analysis pipeline and for enriching the taxonomic and functional analysis of the different samples. We plan to develop web services, more powerful and precise metadata searching, add a number of visualisation options for analysis results and add information about pathways associated with samples.


Soil metagenome NTP1, Riesel, TX
Overview
Quality control
Taxonomy analysis
a
Description
NTP1, Riesel, TX
Classification: Environmental > Terrestrial > Soil
Project: Reconstructing the microbial diversity
Environmental conditions
Biome: Temperate grassland
Experimental feature: Prairie envo:00000260
Material: Grassland soil envo:00
Localisation
Latitude/Longitude: 31.469 , -96.869
Geographic location: USA; Riesel, TX
Map
Other information
Project name
Geographic location (altitude)
Collection date
Amount or size of sample collected
NCBI sample classification
0.0 m
2010
20 mg
410658
InterPro match summary
b
Filter table:
Entry name
ID
pCDS matched
%
GO Terms annotation
A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.
Switch view:
Biological process
Molecular function
Cellular component
Soil metagenome NTP1, Riesel, TX
Overview
Quality control
Taxonomy analysis
Functional analysis
Download
c
Top taxonomy Hits
Switch view:

Sample page showing the richness of EBI Metagenomics data, contextual information and analysis content, highlighting (a) the Overview tab, with contextual information; (b) the Functional analysis tab and (c) the Taxonomic analysis tab.

Selected publication

Hunter, S., Corbett, M., Denise, H., et al. (2013) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. DOI: 10.1093/nar/gkt96

UniProt development

Our team develops the software and services for protein information in the UniProt and GO annotation databases. We are also responsible for the development of tools for UniProt and GOA curation and the study of novel, automatic methods for protein annotation.

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, the UniProt Reference Clusters, and the UniProt Metagenomic and Environmental Sequences.

Major achievements

The UniProt website facilitates the search, identification and analysis of gene products. Our team has been working on new web interfaces and functionalities in response to user feedback gathered in a number of website reviews. We implemented a new website in collaboration with our colleagues at the SIB-Swiss Institute of Bioinformatics (SIB), and it will be released in 2014.

In order to integrate user-community annotation efforts with UniProt, the team developed a pipeline for cross-referencing to Wikipedia articles describing human and E. coli genes and their corresponding gene products.

We worked with our user community and the NCBI RefSeq group to provide a collection of non-redundant reference proteomes and to maintain well-annotated model organisms and others of interest for biomedical and biotechnological research. The collaboration with Ensembl and Ensembl Genomes allowed us to complete the proteome sets and include new species with no coding sequence annotations in the INSDC nucleotide databases.

Our team organised and distributed 594 reference proteomes. All of these datasets are now associated with their corresponding genome assembly, and gene-product centric data sets provided for each reference. New species released in 2013 include Felis catus (cat), Meleagris gallopavo (common turkey), Solanum tuberosum (potato), Musa acuminata (banana), Ashbya gossypii (yeast) and Cochliobolus heterostrophus (Southern corn leaf blight fungus), amongst others.

Collaborations with Ensembl and Ensembl Genomes allowed us to create data links between DNA sequences and the functional proteins they encode. We are now distributing variants with consequences at the protein level from the 1000 Genomes Project.

In 2013 we implemented the XML format for UniRule, a system for automatic annotation of a large volume of uncharacterised proteins. We are planning to distribute the annotation rules in this format for those user communities interested in functional annotation of genomes. We extended the Statistical Automatic Annotation System (SAAS) system to increase the prediction of protein names in UniProtKB. We enhanced the UniRule tool with a statistical assessment of existing and new rules. UniRule annotates over 16 million sequences in UniProtKB/TrEMBL.

In 2013 the team began the task of re-engineering QuickGO, the UniProt GO browser, to use Apache Solr for indexing. This will help us better support the significantly increased data set, which currently stands at more than 250M annotations. We also started to develop a new user interface for QuickGO that adheres to the EBI web style guidelines introduced in 2013. We continued to develop the web-based Protein2GO tool, which UniProt curators use to contribute annotations to the GOA project. Protein2GO has also been adopted as the common annotation tool for the GO Consortium, and we continue to extend it to include new functionality as requested by the GO Consortium curators.

Future plans

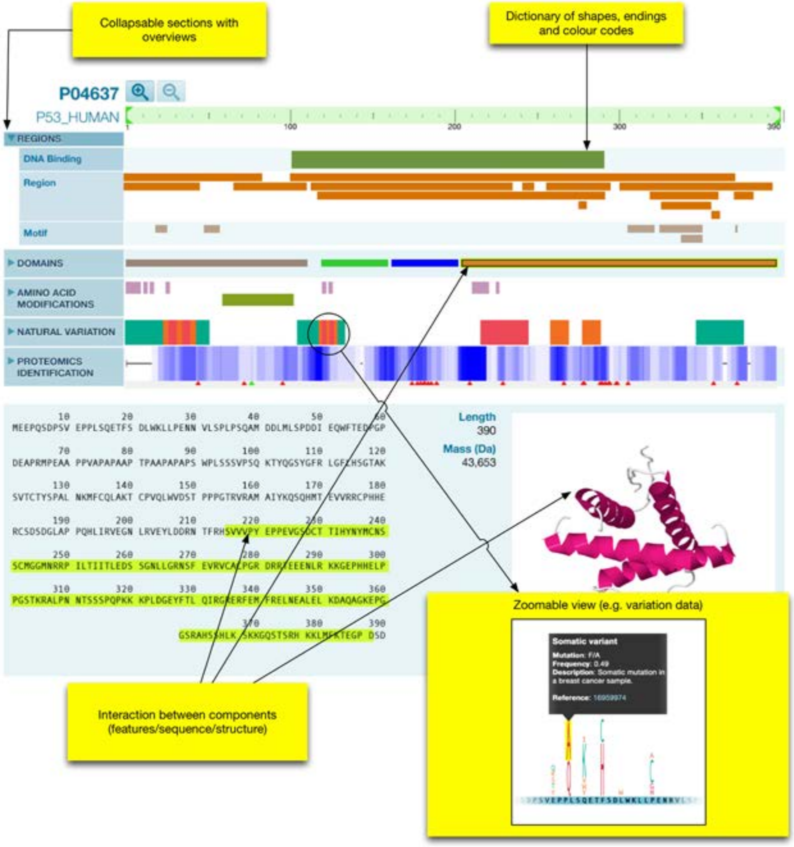
In 2014 we plan to release the new UniProt web site with new interfaces and improved functionality that will facilitate easy access and navigation of UniProt data. We will modify the flat file format to include evidences and methods for annotation, which will help users understand the data sources and the quality of both manual and computational data. We will provide a UniRule XML format for public distribution and explore data-exchange mechanisms with user communities interested in functional prediction. 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 make it easier for our users to explore in-depth the variations and annotations for each specific protein within our resources. 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. In 2014 we also plan to extend the scope of GO annotation to encompass entities other than proteins, in particular RNA and protein complexes.

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.


Collapsable sections with overviews
Dictionary of shapes, endings and colour codes
P04637
P53_HUMAN
100
200
300
390
REGIONS
DNA Binding
Region
Motif
DOMAINS
AMINO ACID MODIFICATIONS
NATURAL VARIATION
PROTEOMICS IDENTIFICATION
Length
390
Mass (Da)
43,653
Interaction between components (features/sequence/structure)
Zoomable view (e.g. variation data)
Somatic variant
Mutation: F/A
Frequency: 0.49
Description: Somatic mutation in a breast cancer sample.
Reference: 16959974
PSVEPPLSQETFSDLWKLLPENNVL

```
        10         20         30         40         50         60
MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP

        70         80         90        100        110        120
DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK

       130        140        150        160        170        180
SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE

       190        200        210        220        230        240
RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS

       250        260        270        280        290        300
SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP

       310        320        330        340        350        360
PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG

                                 370        380        390
                          GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
```

UniProt's Sequence Feature viewer.

Selected publications

The UniProt Consortium (2013). Update on activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 41, D43-D47.

Pedruzzi, I., Rivoire, C., Auchincloss, A.H., et al. (2013). HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 41, D773-D780.

Gene Ontology Consortium, et al. (2013). Gene ontology annotation and resources. Nucleic Acids Res. 41, D530-D636.

Mutowo-Meullenet, P., Huntley, R.P., Dimmer, E.C., et al. (2013). Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. J. Biol. Databases Curation 2013: bas062.

Velankar, S., Dana, J.M., Jacobsen, J., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 41, D483-D489.

Anton, B.P., Chang, Y.C., Brown, P., et al. (2013). The COMBREX Project: design, methodology and initial results. PLoS Biol. 11, e1001638.

Gómez, J., García, L.J., Salazar, G.A.,,et al. (2013). BioJS: An open source JavaScript framework for biological data visualization. Bioinformatics 29, 1103-1104.

Alcántara, R., Onwubiko, J., Cao, H., et al. (2013) The EBI Enzyme Portal. Nucleic Acids Res. 41, D773-D780.

UniProt content

One of the central activities of the UniProt Content team is the biocuration of our databases. Biocuration involves the interpretation and integration of information relevant to biology into a database or resource that enables integration of the scientific literature as well as large datasets. The primary goals of biocuration are accurate and comprehensive representation of biological knowledge, as well as easy access to this data for working scientists and a basis for computational analysis.

UniProt manual curation

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 automatic annotation

UniProt has developed two complementary approaches in order to automatically annotation ofe 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.

UniProt GO annotation (GOA)

The UniProt GO annotation (GOA) program aims to add high-quality 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 (CCDS) project, UniProt is creating an authoritative complete proteome set for Homo sapiens in close collaboration with the RefSeq annotation group (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 38 000 human loci (over 19 000 of these are listed as coding for proteins).

We play a major role in establishing minimal 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.

In 2013 our team is workingmade substantial progress on a ‘gold standard’ dataset to help users identify all experimental data for a given protein from a particular strain of a given organism. This work isWe undertake this workn in collaboration with model organism databases and the Evidence Code Ontology (ECO).

The UniProt GO annotation program provides high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (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.

Future plans

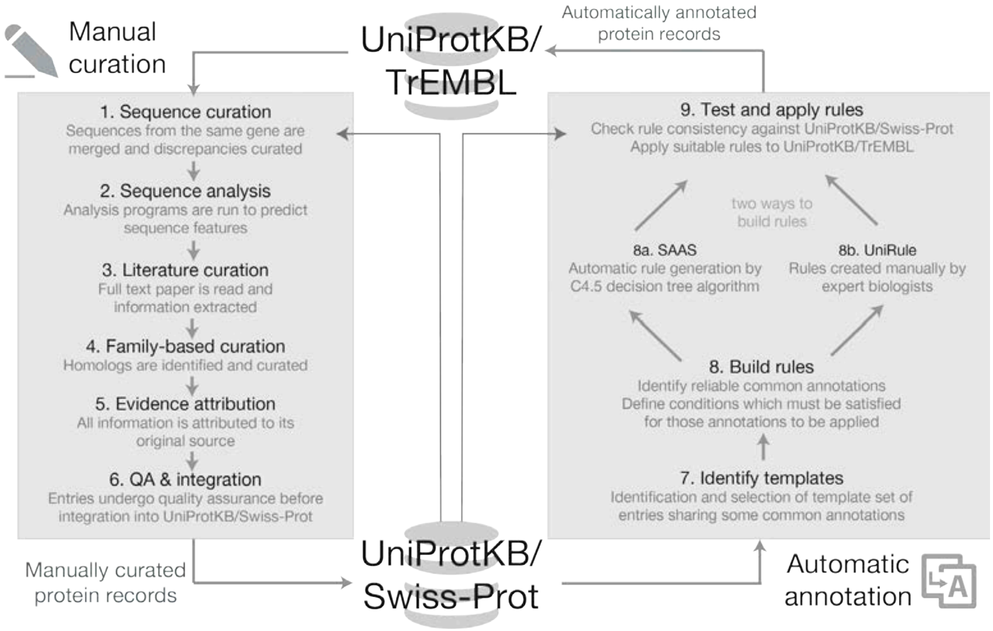
In 2014 we will continue work on a gold-standard dataset across the taxonomic range to fully address the requirements of the biochemical community. We will also continue to expand and refine our Ensembl and Genome Reference Consortium collaborations to ensure that UniProtKB provides the most appropriate gene-centric view of protein space, allowing a cleaner and more logical

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.


Manual curation
UniProtKB/ TrEMBL
Automatically annotated protein records
1. Sequence curation
Sequences from the same gene are merged and discrepancies curated
2. Sequence analysis
Analysis programs are run to predict sequence features
3. Literature curation
Full text paper is read and information extracted
4. Family-based curation
Homologs are identified and curated
5. Evidence attribution
All information is attributed to its original source
6. QA & integration
Entries undergo quality assurance before integration into UniProtKB/Swiss-Prot
9. Test and apply rules
Check rule consistency against UniProtKB/Swiss-Prot
Apply suitable rules to UniProtKB/TrEMBL
two ways to build rules
8a. SAAS
Automatic rule generation by C4.5 decision tree algorithm
8b. UniRule
Rules created manually by expert biologists
8. Build rules
Identify reliable common annotations
Define conditions which must be satisfied for those annotations to be applied
7. Identify templates
Identification and selection of template set of entries sharing some common annotations
Manually curated protein records
UniProtKB/ Swiss-Prot
Automatic annotation

Organisation of data and information in the Universal Protein Resource.

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 2014 we also plan to extend the scope of GO annotation to encompass entities other than proteins, in particular RNA and protein complexes.

Selected publications

Velankar, S., Dana, J.M., Jacobsen, J., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 41 (Database issue), D483-489.

The UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 41 (Database issue), D43-D47.

Alcántara, R., Onwubiko, J., Cao, H., Alcantara, R., et al. (2013) The EBI enzyme portal. Nucleic Acids Res. 41 (Database issue), D773-D780.

Pedruzzi, I., Rivoire, C., Auchincloss, A.H., et al. (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 41 (Database issue), D584-D589.

Mutowo-Meullenet, P., Huntley, R.P., Dimmer, E.C., et al. (2013) Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. Database (Oxford) 2013, bas062.

The UniProt Consortium (2013). Update on activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 41 (Database issue), D43-D47.

Blake J.A.Gene Ontology Consortium, et al. (2013). Gene ontology annotation and resources. Nucleic Acids Res. 41 (Database issue), D530-D636.

Anton B.P., Chang, Y.C., Brown, P., et al. (2013). The COMBREX Project: Design, Methodology and Initial Results. PLloS Biol.ogy 11(8):, e1001638.

Molecular and cellular structures

Understanding the structure of a molecule is key to understanding how it may function. PDBe, Europe's arm of the Worldwide Protein Data Bank collaboration, has a mission to 'bring structure to biology' by making this complex field more accessible to non-specialists.

Protein Data Bank in Europe

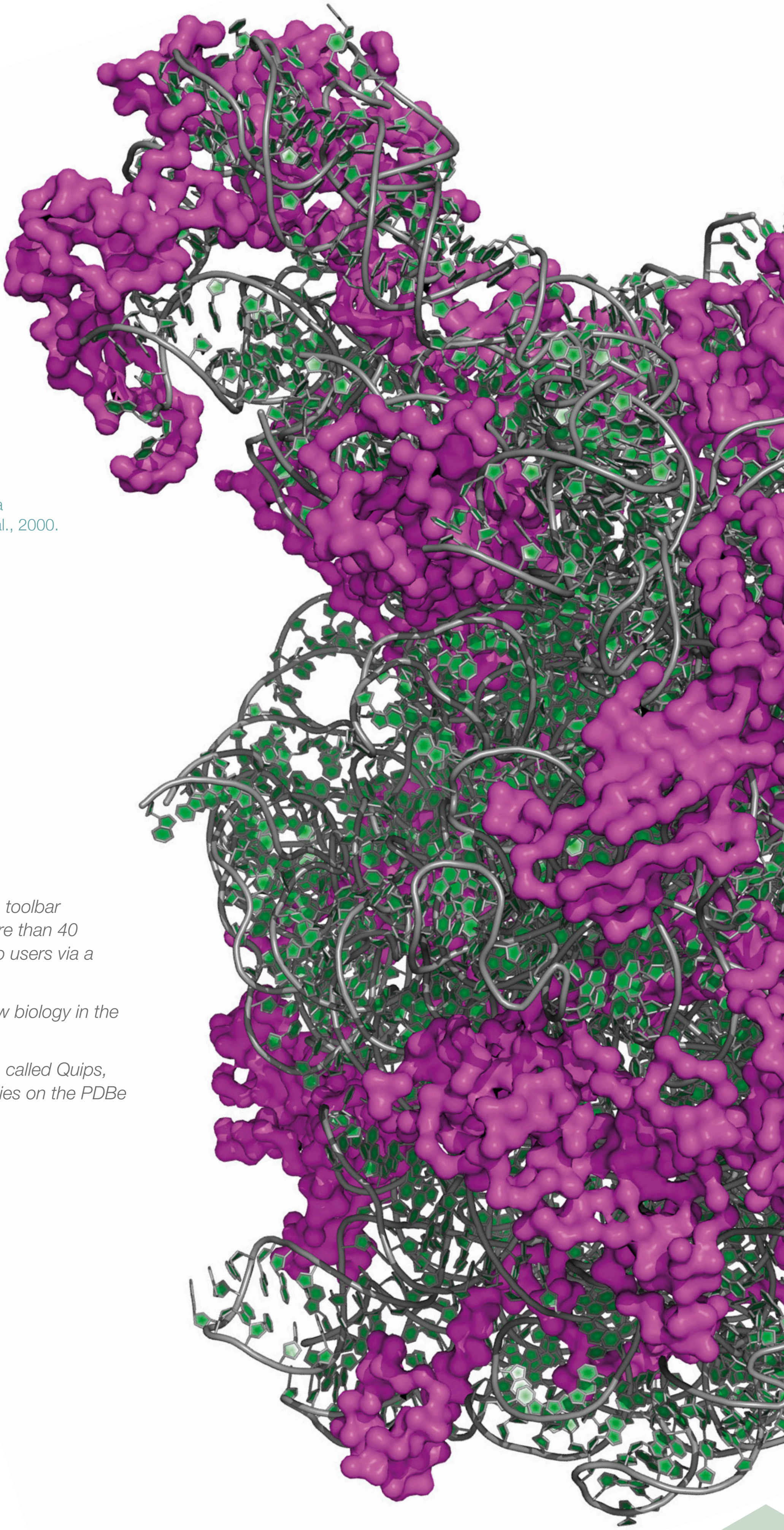
PDBe is the European partner in the Worldwide Protein Data Bank organisation (wwPDB), which maintains the single international archive for biomacromolecular structure data. The other wwPDB partners are the Research Collaboratory for Structural Bioinformatics (RCSB) and Biological Magnetic Resonance Bank (BMRB) in the United States and the Protein Data Bank of Japan (PDBj). PDBe is a deposition and annotation site for the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB).

EMDB is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures. It covers a variety of techniques, including singleparticle analysis, electron tomography and electron (2D) crystallography.

In 2013 the PDB archive welcomed its 10 000th NMR structure and the EMDB acccepted its 2000th entry. Together with all wwPDB partners, we contributed to the successful implementation of a new, more flexible system that now accepts a wider variety of structural data into the Protein Data Bank. During the same period we devoted considerable efforts to improving the navigation and search of PDBe, and we expect to launch the new website in 2014.

Gerard Kleywegt, Tom Oldfield & Sameer Velankar

- *Curated a record 1788 Protein Data Bank (PDB) entries;*
- *Handled the deposition of 268 Electron Microscopy Data Bank (EMDB) entries, representing 49% of worldwide depositions;*
- *In collaboration with partners in the United States and Japan, developed a new common tool for handling the deposition and annotation of structural data on biomacromolecules—obtained using any technique or combination of techniques—at all wwPDB and EMDataBank partner sites;*
- *Carried out extensive internal and external testing of the new tool;*
- *Handled the deposition and release of the first 'big' structures (i.e., too big to fit in a single PDB file or entry), which marked a major step in phasing out the 40-year old PDB format in favour of the much more flexible mmCIF format;*
- *Released the 10,000th NMR structure in the PDB archive and the 2000th EMDB entry;*
- *Put the wwPDB X-ray validation pipeline into production;*
- *Released a stand-alone wwPDB validation server for X-ray crystal structures, which crystallographers can use to assess their structures prior to deposition and publication;*
- *Worked on improving and extending the EMDB data model and on handling 3DEM data in the new deposition and annotation system;*
- *Released a new system for faceted searching and browsing of EMDB as well as several servers that help microscopists analyse their own data;*
- *Improved the validation of NMR structures and data for wwPDB and PDBe;*
- *Made significant progress in implementing prototypes for the new PDBe website, a new query system and a powerful API (for release in 2014);*
- *In collaboration with the Cambridge Crystallographic Data Centre, developed infrastructure to integrate information from the Cambridge Structural Database into the PDB chemical component dictionaries (to be released in 2014);*



Large ribosomal subunit from Haloarcula marismortui, PDB entry 1ffk, by Ban et al., 2000.

- *Released a new version of Biobar, a toolbar for Firefox browsers that makes more than 40 bioinformatics resources available to users via a simple interface;*
- *Introduced a weekly overview of new biology in the PDB archive;*
- *Produced six interactive blog posts, called Quips, and began presenting them as movies on the PDBe YouTube channel.*

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

Protein Data Bank in Europe

The Protein Data Bank in Europe (PDBe) is a resource with integrated structural data 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 Electron Microscopy.

We focus on providing advanced services, integration of structural and other information, and providing ligand-related, validation and experimental data. Our mission is 'Bringing Structure to Biology', and our long-term goal is to make PDBe the logical first stop on any quest for information about 3D molecular and cellular structure.

Major achievements

In 2013 we curated a record 1788 Protein Data Bank (PDB) entries (up from 1676 in 2012, an increase of 7%). The global number of PDB depositions in 2013 was 10,560. PDBe's share of global depositions remained at 17%. In addition, 268 Electron Microscopy Data Bank (EMDB) entries were deposited with PDBe (up from 253 in 2012, an increase of 6%), representing 49% of worldwide depositions.

PDBe and its partners in the United States and Japan developed a new common tool for handling the deposition and annotation of structural data on biomacromolecules—obtained using any technique or combination of techniques—at all wwPDB and EMDataBank partner sites. Our team made major contributions to the system design and software development, and is responsible for the workflow system, validation modules and deposition interface. The new tool underwent extensive internal and external testing in 2013, and Andrea Mattevi (a former member of the PDBe Scientific Advisory Committee) was the first to use it to complete a real deposition of a new structure (PDB entry 4ovj).

In May 2013, the first 'big' structures (i.e., too big to fit in a single PDB file or entry) were deposited as a single entry and released as both split PDB entries and intact mmCIF files. This marked a major step towards phasing out the 40-year old PDB format in favour of the much more flexible mmCIF format. Another milestone in 2013 was the release of the 10,000th NMR structure in the PDB archive. The EMDB archive has almost doubled in size since 2011, and released its 2000th entry in 2013.

PDBe develops advanced software pipelines for the validation of X-ray, NMR and 3DEM structures based on input from community experts on Validation Task Forces. In August 2013 the wwPDB X-ray validation pipeline was put into production, and the validation reports it produces have been well received by depositors, journal editors and structure users. In November, we released a stand-alone wwPDB validation server for X-ray crystal structures, which crystallographers can use to assess their structures prior to deposition and publication.

We worked with the EMDataBank partners and the EM community to improve and extend the EMDB data model and to handle 3DEM data in the new deposition and annotation system. We also released a new system for faceted searching and browsing of EMDB as well as several servers that help microscopists analyse their own data. The group's EM staff published a paper describing improvements to the visualisation and analysis of EM structures on EMDB entry pages (Lagerstedt et al., 2013).

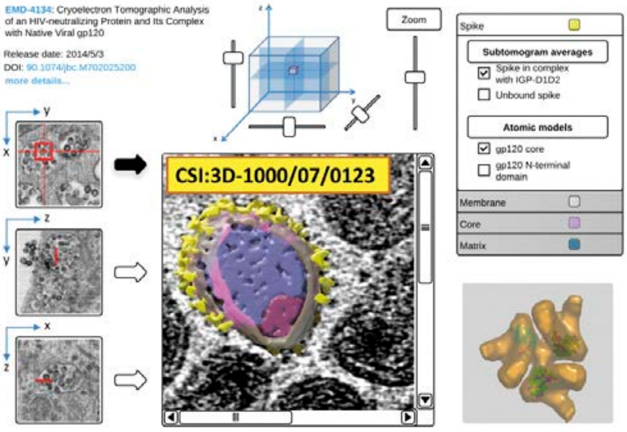
In collaboration with our wwPDB partners, we addressed the handling of NMR data in the new deposition and annotation system. PDBe's NMR experts also focused on improving the validation of NMR structures and data for wwPDB and PDBe. Following the publication of wwPDB's NMR Validation Task Force report, we organised a workshop for major developers of NMR structure-determination software in order to agree on a unified format for NMR restraints.

The PDBe team organises outreach and training activities and participates in EMBL-EBI Bioinformatics Roadshows. In addition to expanding our social media presence in 2013, we ran several webinars and launched our YouTube channel, which now hosts all the video content displayed on the PDBe website. We released six interactive articles on interesting structures ('Quips'), which now include animations of featured structures presented as movies. Usage of our data and services continued to increase steadily, with the average monthly usage of our website doubling between 2011 and 2013. We published 20 papers, welcomed over 30 visitors and participated in two dozen outreach events such as conferences, courses, workshops and roadshows.

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.



EMD-4134: Cryoelectron Tomographic Analysis of an HIV-neutralizing Protein and Its Complex with Native Viral gp120

Release date: 2014/5/3

DOI: 90.1074/jbc.M702025200

more details...

Spike

Subtomogram averages

☑ Spike in complex with IGP-D1D2

☐ Unbound spike

Atomic models

☑ gp120 core

☐ gp120 N-terminal domain

Membrane

Core

Matrix

Providing an overview of macromolecular structures in their 3D cellular context requires data integration, appropriate annotation of structures and good visualisation tools. Here, we show what a volume browser for 3D cellular imaging data could look like, using HIV/SIV as an example.

Future plans

To transform the structural archives into a truly useful resource for biomedical and related disciplines, we will continue to focus on developing advanced services such as PDBePISA, PDBeFold, PDBeMotif and powerful new search and browse facilities. We will also devote considerable efforts to 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 help all users understand the extent to which the data support the structural models and inferences.

In 2014 we will see the fruits of large-scale efforts that got underway in 2012. A completely redesigned PDBe website will feature improved design and functionality guided by the expectations and preferences of our users. We hope the new presentation – particularly of individual entries –will raise the bar and attract many new users. We will also launch a unique resource with comprehensive analysis and validation information and tools for X-ray, NMR and 3DEM structures in the PDB and EMDB archives. With the launch of these new facilities, we will endeavour to raise awareness of the wealth of 3D structural information available from PDBe throughout the life-science community, with particular emphasis on engaging non-expert users of structures.

Rollout of the new system for deposition and annotation of X-ray structures at some of the wwPDB sites will begin in January 2014, with the other sites and techniques (NMR, 3DEM) following shortly after. There will be a transition period in which the old and new systems will operate in parallel, and by the end of 2014 we expect all structure depositions to be handled with the new system.

In 2014 the wwPDB organisation will introduce the new deposition and annotation system, release a new style of validation report for all major techniques, launch the repurposed pdb.org website, include intact 'big' structures in the archive (ending the long-standing archival PDB format), and release the milestone 100 000th PDB entry. During this year of significant change, PDBe will host the annual meeting of the wwPDB Advisory Committee and the inaugural meeting of the wwPDB Hybrid Methods Task Force.

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. We are collaborating actively in this area so we can meet the challenges and seize upon the opportunities afforded by the exciting developments ahead.

Selected publications

Berman, H.M., et al. (2013) How community has shaped the Protein Data Bank. Structure 21, 1485-1491.

Gutmanas, A., et al. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. Acta Crystallogr. D69, 710-721.

Hendrickx, P.M.S., et al. (2013) Vivaldi: visualization and validation of biomacromolecular NMR structures from the PDB. Proteins Struct. Funct. Bioinf. 81, 583–591.

Lagerstedt, I., et al. (2013) Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. J Struct. Biol. 184, 173-181.

Velankar, S., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 41, D483-D489.

PDBe databases and services

The Protein Data Bank in Europe (PDBe) is a key resource at EMBL-EBI and is a partner in the Worldwide Protein Data Bank organisation (wwPDB) along with RCSB and BMRB in the United States and PDBj in Japan.

The PDBe Databases & Services team manages two production systems: the weekly exchange of deposited data and the weekly increment of newly released PDB and EMDB data. These production data systems are managed within multiple Oracle databases and support a large number of integrated web resources to collect data and disseminate information to the wider life-science community.

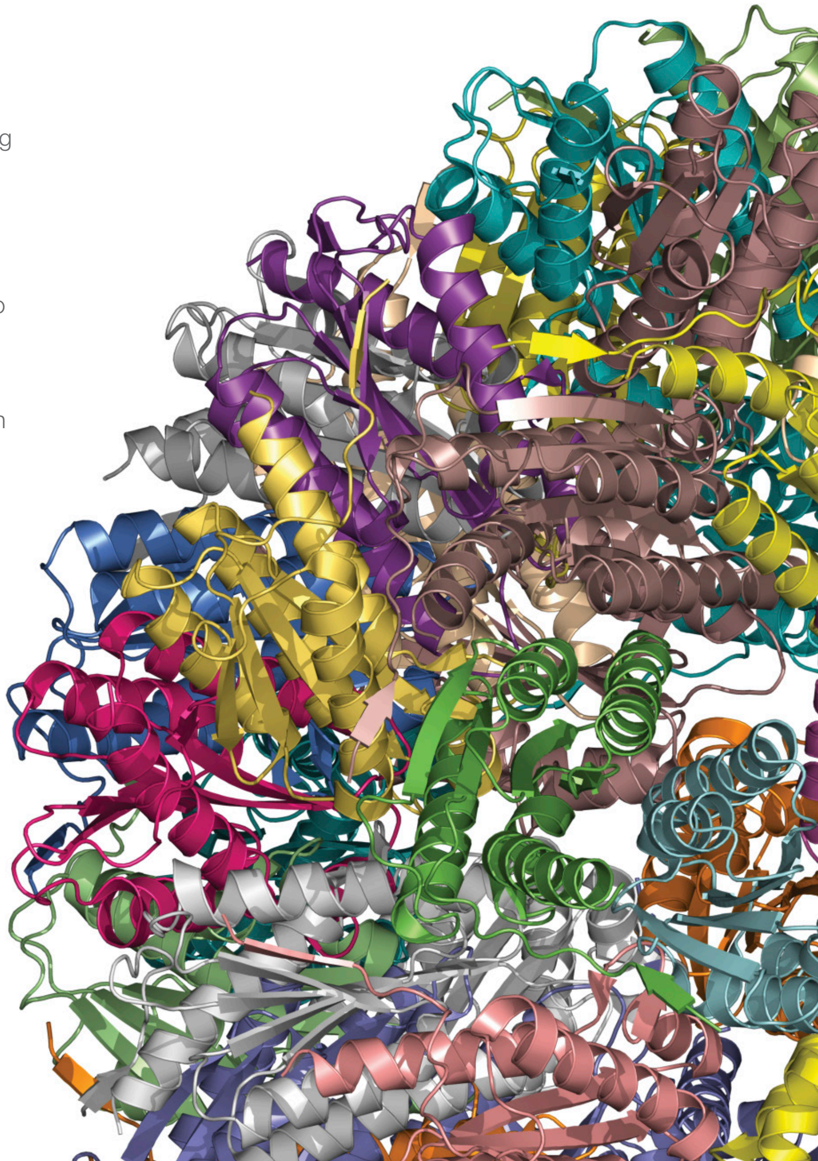
Major achievements

During 2013 the PDBe Databases & Services team completed the migration of services to the London data centres. This work required the creation of the PDBeMotif database infrastructure in London and major modifications to the design of the underlying loading and processing steps for PDBeMotif to support the tier 3 London system. Other systems that support the static webpages, NMR pages and EM infrastructure are now all managed by the London infrastructure. Additionally, our team created a number of new resources to automate the management of systems in London in a seamless manner.

Our team put the SIFTS infrastructure into production during 2013. This required an additional database schema and processing of the data each week, as well as maintaining the old system during the transition period.

The PDBe team put the stand-alone wwPDB X-ray validation server into production. This server allows users to upload a structure model and underlying experimental data and checks the quality before they deposit or publish it. The team made significant contributions to the coordinated development of the new wwPDB deposition and annotation system, and released a beta version of the system in September 2013. This involved the integration of the deposition user interface, workflow system and workflow manager for annotators.

As part of the design of the new PDBe website (to be launched in 2014), we specified and designed a number of new infrastructure components. This included hardware servers, Lucene search systems, an API to collate data and a unified search system with faceted output.



Tom Oldfield

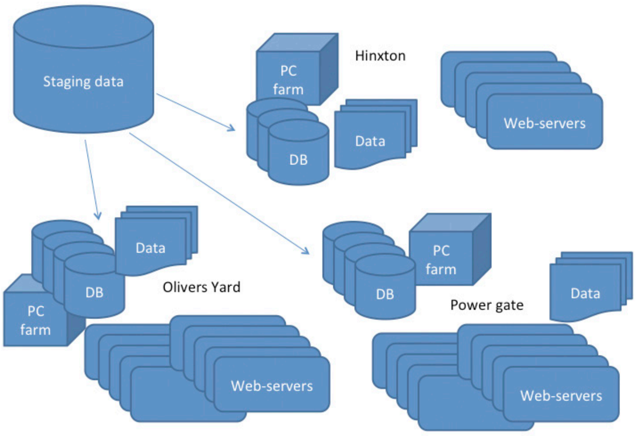
DPhil University of York, 1990. Postdoctoral research at GlaxoSmithKline, 1990-1993. Principal Scientist at Accelrys Inc., 1993-2002.

At EMBL-EBI since 2002. Team Leader since 2010.

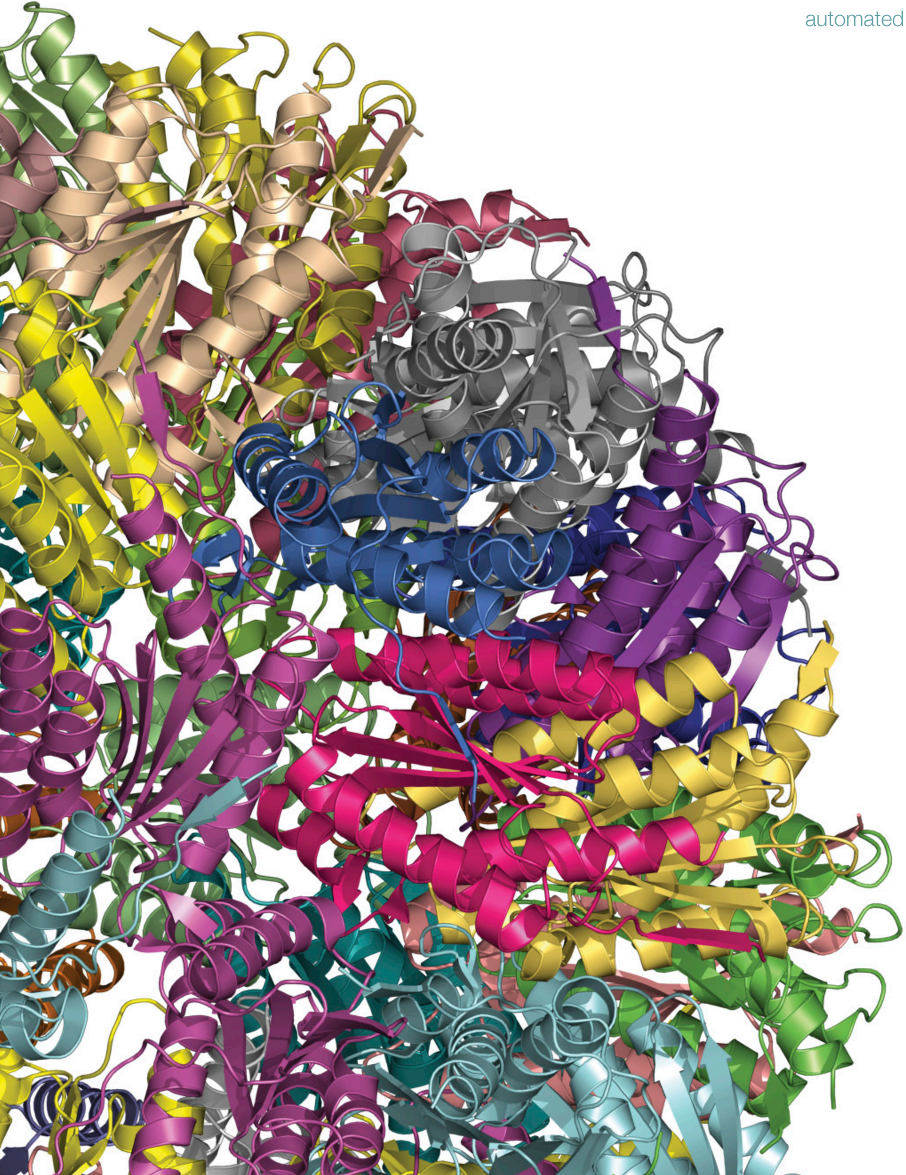
Future plans

The new wwPDB deposition and annotation system will go into production in early 2014. Due to the significant changes in the way depositions will be carried out, with far more data checking taking place before submission, it is expected that a major effort by wwPDB (and PDBe) technical staff will be required to make sure the final improvements to the code base are all working properly.

The redesigned PDBe website will be released in 2014. We will devote considerable efforts to creating an integrated system that allows all the underlying PDBe infrastructure to appear as a single, unified resource that is easily accessible for casual and novice users, while also providing expert knowledge to advanced users efficiently.



PDBe maintains three copies of fail-over production databases that are recreated at 00:00 every Wednesday. This involves rebuilding databases, running hundreds of processes and moving six terabytes of data in a window of 38 hours in time for the weekly release. The data and remote database servers are managed with automated systems.



The ever-larger and increasingly complex structures deposited in the PDB require robust data-handling and archiving systems to manage the weekly release of approximately 200 new entries.

PDBe content and integration

The goal of the Protein Data Bank in Europe (PDBe) is to serve the biomedical community by providing easy access to molecular and cellular structure data. Our team is responsible for ensuring the PDBe web interface serves users well, and for designing new tools to facilitate access to integrated, high-quality structural data.

Major achievements

In 2013 PDBe annotation staff curated a record number of PDB entries (1788 entries, up 7% from 2012), the majority of which were annotated within one working day of being deposited. PDBe staff also annotated 268 EMDB entries; this represents 49% of all EMDB entries deposited in 2013.

We continued our efforts to improve data quality by carrying out an extensive remediation of experimental crystallographic data in the PDB archive, addressing inconsistencies in the labelling of data items in the structure-factor files that were deposited at PDBe. In collaboration with our wwPDB partners, the PDBe team implemented a new annotation policy for entries that contain more coordinate data than can be fit into a single PDB file. These entries are now made available as a single mmCIF file and annotated as a single entry. This removed a long-standing issue related to large macromolecular complexes such as ribosomes that previously had to be represented in multiple PDB entries.

The PDBe team developed a new wwPDB validation pipeline for X-ray crystallographic data and model coordinates. This pipeline was put into production in August 2013 and made available to all wwPDB partner sites. Since then, all newly deposited structures determined by X-ray crystallography have been validated using the new pipeline, and the resulting validation reports sent to the depositors. The pipeline is also available as a stand-alone server, which allows structural biologists to validate their structures before depositing them to the PDB.

We carried out rigorous testing of the new wwPDB deposition and annotation system in 2013. The new system was made available to a small number of crystallographers who deposited structures and provided feedback. The wwPDB development team acted on feedback from internal and external testing, which resulted in significant improvements to the functionality of the new system.

In 2012, we carried out a user-experience study to better understand the use of our website and resources. We also gathered user requirements for a future API for PDB and EMDB data. In 2013, we made significant progress in implementing prototypes for the new PDBe website, a new query system and a powerful API. These prototypes will be tested and released in 2014.

Our team is working closely with the Cambridge Crystallographic Data Centre (CCDC) to develop infrastructure to integrate information from the Cambridge Structural Database (CSD) into the PDB chemical component dictionaries. The information will include atomic coordinates for the small molecule components that are common to PDB and CSD. This will be released in 2014 and updated as new compounds are archived.

We produced six blog posts, called Quips, in 2013. These posts highlight one or more interesting or topical structures, coupled with animations shown in an interactive 3D viewer. We developed software to capture the animations from Quips articles so that we can also present them as movies on the PDBe YouTube channel.

The PDBe team released a new version of Biobar, a toolbar for Firefox browsers that makes more than 40 bioinformatics resources available to users via a simple interface. We also introduced a weekly updated overview of new biology in the PDB archive, which provides users with information about any new proteins, Pfam domains and GO terms in the weekly release of the PDB.

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.

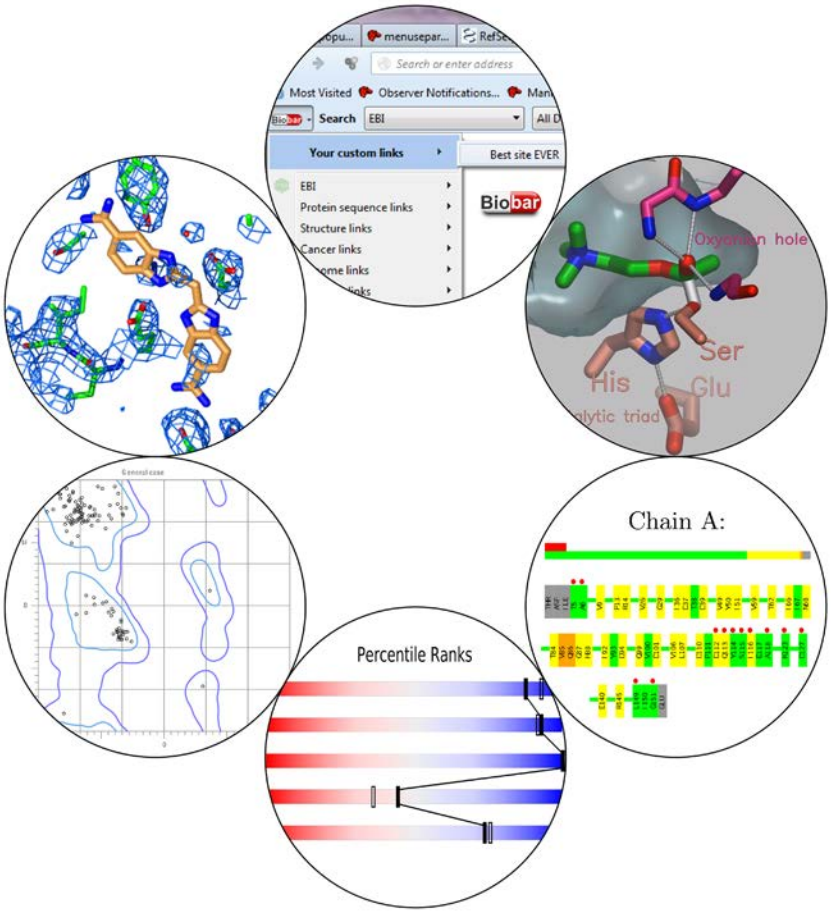
Future plans

We will work with our wwPDB partners to launch the new deposition and annotation system in 2014, which will significantly change the annotation practices at PDBe. We are developing validation pipelines for NMR and EM structures and these will be put into production in 2014 and also made available as stand-alone servers. The wwPDB X-ray validation pipeline produces reports that make it easier for non-experts to understand the quality of a PDB entry. We will integrate this validation information in the new website. We will complete the redesign of the website, query system and API and release them in 2014. We will continue our data-integration efforts and plan to integrate genomic data in the SIFTS infrastructure by developing infrastructure to map gene data and variation information to macromolecular structures.

Selected publications

Velankar, S., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res.41, D483-D489.

Gutmanas A., et al. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. Acta Crystallogr.D69, 710-721.



PDBe Content and Integration team activities and services (clockwise from top): Biobar toolbar for Firefox web browser, providing access to over 40 bioinformatics resources; Quips blog posts with animations highlight interesting structures in the PDB archive; the wwPDB structure validation pipelines developed at PDBe produce reports with a graphical overview of the structural quality of individual protein molecules in a structure; validation reports also contain a “slider plot” summarising the overall quality of the structure; another feature of the reports: quality information derived from Ramachandran plots produced by MolProbity (Duke University); the validation pipeline includes assessment of ligand quality in terms of geometry and fit to the experimental data. For X-ray crystal structures, the EDS software (Uppsala university) is used.

Chemical biology

The ability to explore, modulate and control bioactive entities has huge economic and healthcare impacts, and EMBL-EBI's chemical biology resources help researchers study small molecules and their effects on biological systems.

In 2013, Digital Science donated the SureChem patent database of chemical structures to EMBL-EBI. Now dubbed SureChEMBL, this substantial resource now makes information about millions of structures freely available to the drug discovery community. Other additions to ChEMBL include thousands of USAN compounds and the extension of UniChem to provide structure cross-references to chemical structures in 22 databases. ChEMBL is also now available through EMBL-EBI's new Semantic-Web platform.

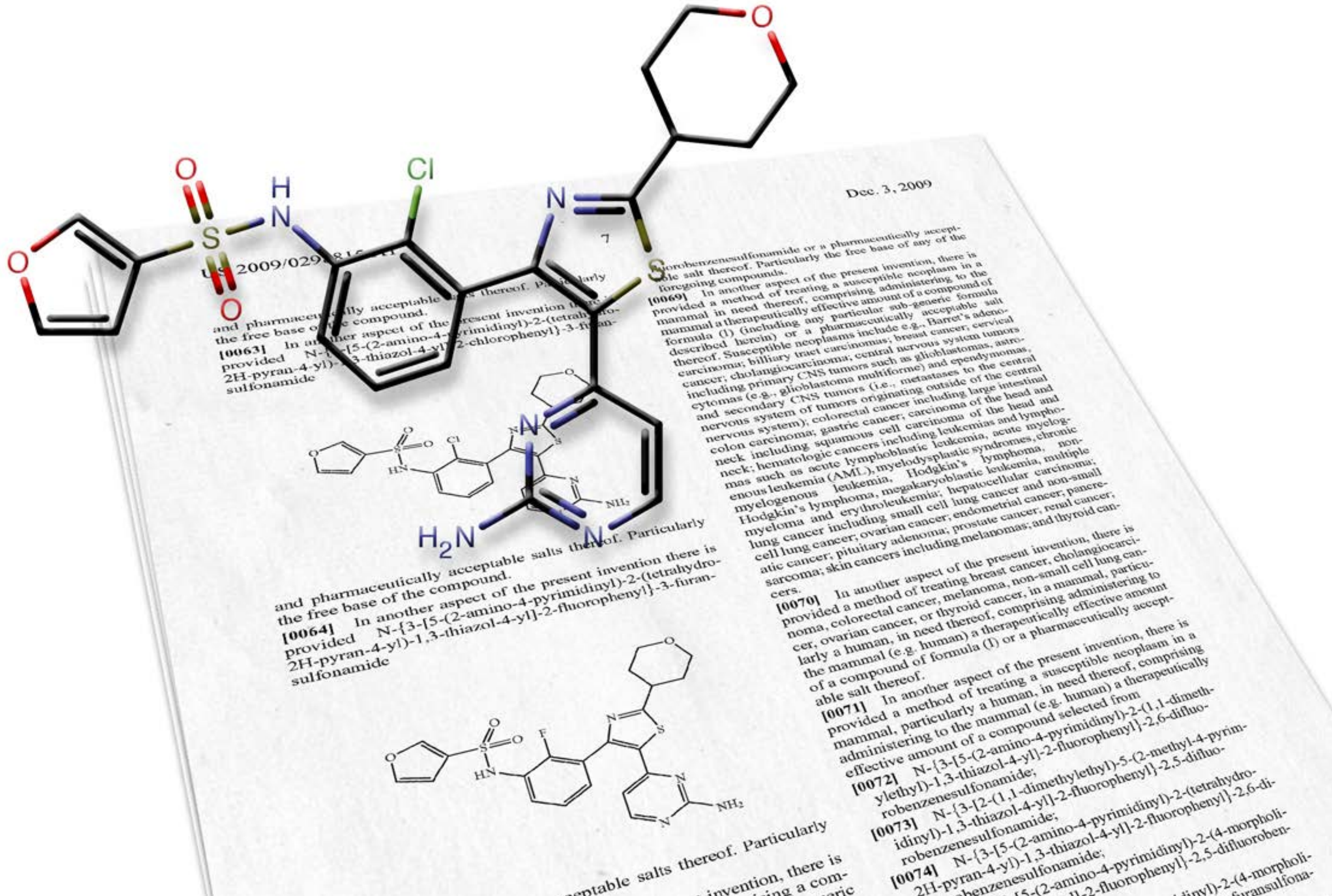
The Cheminformatics and Metabolism team launched Metingear, an application for annotating metabolic reconstructions that enables researchers to analyse the properties of reconstructed networks. They also launched MassCascade, a tool for processing metabolomics liquid chromatography tandem MS data. MassCascade takes a visual approach and simplifies the process of peak extraction and identification into modular steps.

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 GPCR and Kinase SARfari web portals.

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'.



John Overington

- *Transferred the SureChem (now SureChEMBL) patent database of chemical structures gifted from Digital Science to EMBL-EBI;*
- *Grew the ChEMBL database by approximately 100 000 compounds and 2 million bioactivity values, largely thanks to extraction of data from the scientific literature and deposition of bioactivity data for compounds tested for malaria and tuberculosis;*
- *In the context of BioMedBridges, grew the UniChem structure cross-referencing resource to provide cross-references to more than 65 million chemical structures from 22 source databases;*
- *Supplemented the drug-discovery data in ChEMBL with the set of over 10 000 compounds that have been given USAN (United States Adopted Names);*
- *Annotated all the FDA-approved drugs in the database with information about their therapeutic target(s) and mode of action;*
- *Improved the ChEMBL data model to allow for better representation of protein families and complexes;*
- *Produced a Semantic Web version of ChEMBL in the Resource Description Framework (RDF) format;*
- *Produced a fully open-source version of ChEMBL on a virtual machine (called myCHEMBL) that is easy to install for users who do not have access to commercial chemical structure searching or database software.*

www.ebi.ac.uk/chembl

Christoph Steinbeck

- *Introduced new tools into the ChEBI application suite: BiNChE 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;*
- *Developed a Java library for NMR signal processing;*
- *Completed a quantum-mechanics pipeline for NMR spectra prediction;*
- *Significantly improved the Chemistry Development Kit (CDK);*
- *Issued a release of Metingear, a desktop application for creating and curating metabolic reconstructions;*
- *Improved the desktop and command-line versions of the SENECA structure elucidator.*

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

ChEMBL

Our teams activities centre on ChEMBL, EMBL-EBI's drug discovery resource of quantitative small-molecule bioactivity data. The combination of structure–activity relationship (SAR) data from the scientific literature, deposited data from neglected disease high-throughout screens and the patent literature all make ChEMBL an important and enabling resource for scientists working in pharmaceutical R&D.

Drug discovery is more costly than ever, and achieving an appropriate balance of efficacy and safety remains a challenge. Changes in the structure of the pharmaceutical industry over the past decade have led to an increase in drug-discovery activities in academic and small biotechnology organisations, which do not typically have access to large databases of legacy bioactivity data. ChEMBL is a freely available resource that contains curated chemical structures, bioactivity values and their relationship to biological targets. It can be used to bridge this data gap between big pharma and small biotech, for example in terms of identifying new tool compounds or repurposing opportunities.

Our research interests centre on mining ChEMBL for data that can be applied to drug-discovery challenges, and finding new ways to increase the database in a qualitative and quantitative manner.

Major achievements

The biggest change in 2013 was the transfer of the SureChem patent database of chemical structures to EMBL-EBI. Now dubbed SureChEMBL, this significant resource is now freely available to the drug discovery community. As early information on novel chemical entities is published initially in the patent literature, the data in SureChEMBL now provide a useful complement to the peer-reviewed scientific literature data in ChEMBL itself.

During 2013 the content of the ChEMBL database increased 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 EU-OPENSCREEN, is our structure cross-referencing resource, and now provides database cross-references to more than 65 million chemical structures from 22 source databases, both internal and external to EMBL-EBI, and is a core part of chemical structure integration for the BioMedBridges project.

The ChEMBL database continues to be used extensively by researchers: each month, the website is accessed by more than 1500 distinct users and the database download page visited by 250 organisations. The paper about ChEMBL in Nucleic Acids Research (Gaulton et al., 2012) was well received, with 430 citations as of February 2014. We ran several webinars and training courses in 2013, and carried out a roadshow at 14 university chemistry departments in the United Kingdom.

To help researchers understand the key properties of druggable targets and compounds, we supplemented the drug-discovery data in ChEMBL with the set of over 10 000 compounds that have been given USAN (United States Adopted Names). This set comprises compounds in later preclinical or clinical development. We also annotated all the FDA-approved drugs in the database with information about their therapeutic target(s) and mode of action.

We also made a number of changes to the ChEMBL data model to allow for better representation of protein families and complexes. Our data curation activities focused on standardising bioactivity types (e.g. IC50, Ki), standardising units, and identifying possible duplicate data and values outside of a normal range or with incorrect units for the activity type.

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.

In response to user feedback and input from the Innovative Medicines Initiative (IMI)-funded OpenPHACTS project, we produced a Semantic Web version of ChEMBL in the Resource Description Framework (RDF) format. In addition to accessing ChEMBL data via the web interface, database downloads and via REST web services, users can now query bioactivity data through the new RDF platform. This framework is updated for each ChEMBL release. We also produced a fully open-source version of ChEMBL on a virtual machine (called myCHEMBL) that is easy to install for users who do not have access to commercial chemical structure searching or database software.

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. In 2013 we completed ADME SARfari, a comparative genomics system of drug-metabolising systems that enables identification of the absorption, distribution, metabolism, and excretion protein targets with which a compound is likely to interact. ADME SARfari allows comparison of species differences in these protein sequences, their variants and tissue distribution.

We are participating in three EU-funded projects: eTOX, diXa, and HeCaTos, which aim to better curate toxicity data and use it to predict toxicity. We are partners on OpenPHACTS, an IMI project that integrates pharmacological data across diverse resources, and contribute to the BioMedBridges and EU-OpenScreen projects.

Research

Our basic research uses ChEMBL as the primary source for data mining. We also collaborate with several groups on integrating ChEMBL data and methods in other datasets and applications.

Using small-molecule bioactivity data as a functional output, we conducted large-scale analyses to assess conservation of ligand-binding function between related proteins. We began reconstructing assay cascades from ChEMBL data and explored changes in affinity that occur as a molecular-to-phenotypic scale is traversed.

We studied the differential expression of all known, pharmacologically relevant 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.

In order to identify leads for completely novel tuberculosis medicines, we used predictive models that identify the primary target for a set of previously identified compounds with an anti-tuberculosis activity. We are working with collaborators at PLACE to confirm these targets experimentally.

To better understand resistance mechanisms within drug design and agricultural chemistry, we applied integrated sequence and molecular models followed by molecular dynamics simulations and in vitro experiments. We found physicochemical and structural properties of allosteric regulators that may offer opportunities to tackle classically difficult drug targets. These trends allow us to predict allosteric interactions with a good reliability.

Future plans

In 2014 our primary goal is to broaden ChEMBL's utility by adding additional annotation, for example on diseases and targets. We will expand our use of ontologies to increase indexing of ChEMBL data, particularly for complex and high-value endpoints such as ADMET, and in vivo pharmacology assays. We will continue development of technologies that enable us to build curation and data submission interfaces in a flexible and extendable way. We will also use text-mining methodologies to identify journal articles that enhance our coverage of chemical space.

Selected publications

Bento, A.P., Gaulton, A., Hersey, A., et al. (2014) The ChEMBL bioactivity database: an update. Nucl. Acids Res. 42, D1083-D1090.

Van Westen, G. and Overington, J.P. (2013) A ligand's-eye view of protein similarity. Nat. Methods 10, 116-117.

Ochoa, R., Davies, M., Papadatos, G., et al. (2014) myChEMBL: a virtual machine implementation of open data and cheminformatics tools. Bioinformatics 30, 298-300.

Cheminformatics and metabolism

Our team provides the information on metabolism: small molecules and their interplay with biological systems. We develop and maintain MetaboLights, a metabolomics reference database and archive; ChEBI, the database and ontology of chemical entities of biological interest; and the Enzyme Portal, which comprises EMBL-EBI enzyme resources including Rhea and IntEnz.

We develop methods to decipher, organise and publish the small molecule metabolic content of organisms. 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 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 chem- and bioinformatics. The chemical diversity of the metabolome and a lack of accepted reporting standards currently make analysis challenging and time-consuming. Typical mass spectrometry-based studies, for instance, generate complex data where the signals of interest are obscured by systematic and random noise. Proper data pre-processing and consequent peak detection and extraction is essential for compound identification. Part of our research comprises the development and implementation of methods to analyse spectroscopic data in metabolomics.

Major achievements

Quantum mechanics allows the ab initio calculation of NMR parameters, which can be used to simulate NMR spectra. In 2013 we concluded the design of a pipeline for NMR simulation and tuned it to run on the Hinxton cluster. We assessed its performance by simulating NMR spectra for a small set of Natural Products (NP), and presented the results at the first European Conference on Natural Products.

In 2013 we released and published a desktop application, Metingear, for annotation metabolic reconstructions with chemical structure. Metingear bridges the gap between manual and automated curation and provides tools for analysing the properties of reconstructed networks. It is already being used to reconstruct and compare metabolic networks of axilla bacteria. Metingear provides accurate and efficient handling of chemical structure information, and now allows on-demand processing of metabolic data sets.

We contributed our work on Metingear to the open source Chemistry Development Kit (CDK) library, which is widely utilised by groups at EMBL-EBI.

We also released MassCascade, a tool for metabolomics liquid chromatography tandem mass spectrometry data processing, and made the core library available as a plug-in for an open-source workflow platform. This visual approach to data processing breaks down the long and complex process of peak extraction and identification into modular steps, following the visual programming paradigm. The integration of mass spectrometry functionality into a well-accepted workflow platform with a strong data analysis user base enables scientists to use our plug-in in synergy with existing tools.

In 2013 we generated statistical models for the prediction of NMR one-bond proton-carbon coupling constants to aid in computer assisted structure elucidation. The models are based on quantitative structure property descriptors calculated from mined literature data. Predictions were benchmarked against one-bond proton-carbon values computed via a quantum-mechanical approach in NWChem.

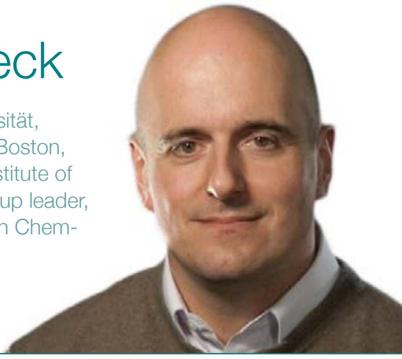
We reviewed existing information-extraction tools for complex chemical documents (e.g. Optical Structure Recognition Application), and evaluated their performance against a manually curated reference set of full-text articles. We also developed a streamlined and optimised desktop application for automated extraction and manual curation of full-text documents.

In 2013 we enhanced the Enzyme Portal web site with two new search types: protein sequence search and chemical structure search, which use NCBI-BLAST and ChEBI structure search, respectively. Users can now compare two enzymes side by side, with highlighted differences.

We continued to develop Rhea, which is incorporated in the Enzyme Portal, to handle polymers and macromolecules properly, with stoichiometry checks.

This past year we extended MetaboLights to offer more reference data and visualisations around spectral information and pathways. Our improved study-upload queueing system and simplified download mechanism made sharing primary metabolomics research data easier.

Christoph Steinbeck

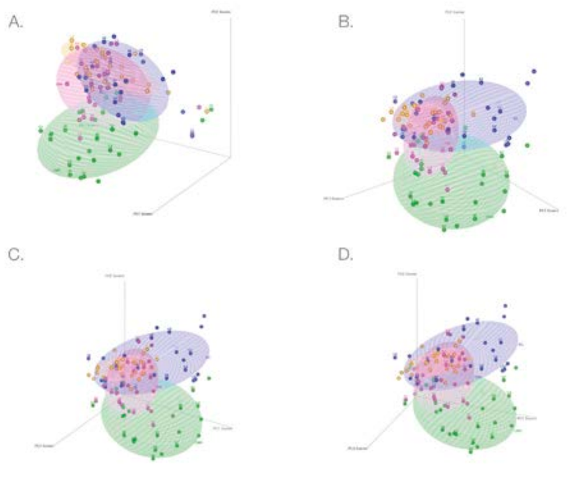


PhD Rheinische Friedrich-Wilhelm-Universität, Bonn, 1995. Postdoc at Tufts University, Boston, 1996-1997. Group leader, Max Planck Institute of Chemical Ecology, Jena, 1997-2002. Group leader, Cologne University 2002-2007. Lecturer in Cheminformatics, University of Tübingen, 2007.

At EMBL-EBI since 2008.

We improved our ISAcreator Metabolite Annotation Plugin to fully support Java7 for both OpenJDK and standard Oracle Java.

ChEBI saw strong data growth in 2013, with a 20% increase over 2012 in fully annotated entries and additional cross-references to important resources such as Wikipedia and the Enzyme Portal.



PCA plot from metabolomics LC-MS data for four genotypes of tomato (Solanum lycopersicum). Metabolomics snapshots were studied at different stages of ripening (days indicated by labels). The plot reveals distinctly different ripening trajectories for each genotype.

Future plans

Our central theme of research is 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 various 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 these projects with an emphasis on applicability in our service foci. Here, our focus is on the extension of the ChEBI database towards greater usability for metabolism and natural products research, wider scope of the Rhea database by proper handling of polymers and macromolecules as reaction participants, improving the curation efficiency (which requires work on curation tools and narrow AI in the area of on-the-fly information retrieval) and the extension and establishment of our metabolomics database, MetaboLights. We will further enrich MetaboLights with more curated knowledge such as reference spectra, pathways, protocols and references to a larger number of existing resources. New online data analysis capabilities will strengthen MetaboLights position as an important research tool for the metabolomics community.

Selected publications

May, J.W., James, G. and Steinbeck, C. (2013) Metingear: A development environment for annotating genome-scale metabolic models. Bioinformatics 29, 17.

Beisken, S., Meinl, T., Wiswedel, B., et al. (2013) KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinform. 14, 257.

Beisken, S., Earll, M., Portwood, D., et al. (2013) MassCascade: Visual programming for LC-MS data processing in metabolomics. Bioinformatics (submitted)

May, J.W. and Steinbeck, C. (2014) Efficient ring perception for the Chemistry Development Kit J. Cheminform. 6, 3.

De Matos, P., Cham, J.A., Cao, H., et al. (2013) The Enzyme Portal: A case study in applying user-centred design methods in bioinformatics. BMC Bioinform. 14, 103.

Salek, R.M., Haug, K., Conesa, P., et al. (2013) The MetaboLights repository: curation challenges in metabolomics. Database 2013, 10.1093/database/bat029.

Hastings, J., de Matos, P., Dekker, A., et al. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 41, D456-D463.

Tudose, I., Hastings, J., Muthukrishnan, V., et al. (2013) OntoQuery: easy-to-use web-based OWL querying. Bioinformatics 29, 2955-2957.

Molecular systems

The genes and gene products encoded by genomes act in co-ordinated systems, often containing protein, small molecule and oligonucleotide or oligosaccharide components. Our Molecular Systems resources are focused on systems biology, spanning enzymes and their mechanisms, protein—protein interactions and networks, molecular pathways and approaches to quantitatively model entire complex biological systems.

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.

http://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://ebi.ac.uk/reactome

BioModels

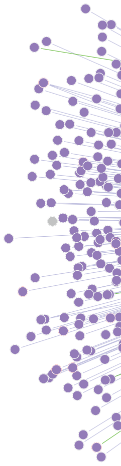
BioModels Database is a repository of peer-reviewed, published, computational models, primarily from the field of systems biology and of wide biological application. BioModels allows biologists to store, search and retrieve mathematical models covering a wide range of diverse systems. In addition, the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

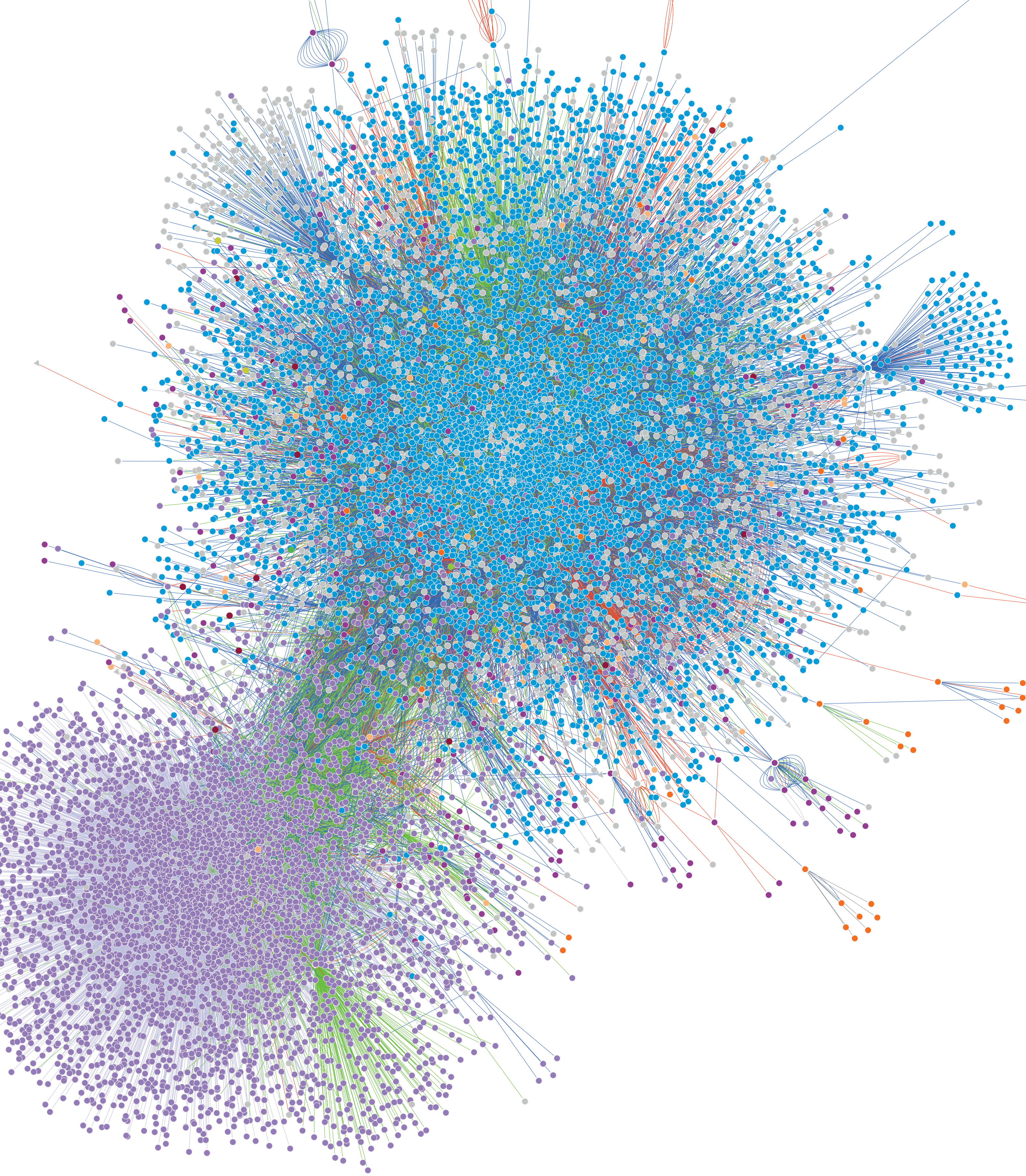
http://ebi.ac.uk/biomodels

Henning Hermjakob

- *The MINT database, a long-standing IMEx partner, joined forces with IntAct and is now sharing the IntAct curation platform;*
- *Released a complete redevelopment of the Reactome web interface, providing tight integration with other resources;*
- *Contributed to a large-scale collaborative effort for the semi-automated generation of systems biology models, which produced more than 142 000 models covering 1852 species, and released the models in the BioModels database;*
- *Expanded the IMEx collaboration for the globally co-ordinated curation of molecular interaction data;*
- *Released a complete redevelopment of the Reactome web interface, providing tight integration with other resources including IntAct, Expression Atlas, PDBe, ChEBI and Rhea;*
- *Contributed to a large-scale effort for the semi-automated generation of systems biology models based on KEGG pathways, Biocarta, MetaCyc and SABIO-RK data, releasing over 142 000 models through BioModels Database;*
- *Contributed to BioJS, a JavaScript-based library of reusable components for presenting biological data.*

www.reactome.org
www.ebi.ac.uk/intact
www.ebi.ac.uk/biomodels-main
www.ebi.ac.uk/Tools/biojs/registry





EMBL-EBI's molecular interaction database, IntAct, is a member of the International Molecular Exchange Consortium (IMEx). Here, we show IntAct data that depicts a map of molecular interactions in Parkinson's disease. These data can help researchers understand the molecular pathways involved in this neurodegenerative disease. This figure shows how proteins from different species are used to look for evidence of interactions of proteins associated with Parkinson's disease. The coloured dots represent molecules: blue represents human and purple represents mouse. Each interaction 'evidence' is depicted as an edge; interaction types are colour-coded. At this density you can get a sense for how often researchers use proteins from different species in the same interaction experiment. Studying molecular interactions in closely related species in this way can shed light on important molecular interactions in humans.

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: all of the resources we develop are freely available.

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

A major success in 2013 was the strong community acceptance of the ProteomeXchange data dissemination platform. In this EU-funded consortium, PRIDE works with a number of international partners (e.g., PeptideAtlas, UniProt, University of Ghent, University of Liverpool, ETH Zurich, University of Michigan, Wiley-VCH) to co-ordinate data deposition and dissemination strategies for mass spectrometry data, providing a single entry point for data deposition, a shared accession number space and a deposition metadata format. ProteomeXchange started full production in spring 2012 with 102 submissions, growing to 527 in 2013. In addition, we released the PRIDE Cluster algorithm (Gris et al., 2013), used to score peptide identifications using the data previously stored in PRIDE.

The IMEx collaboration for the globally co-ordinated curation of molecular interaction data continued to grow. IMEx partners share formats, identifier spaces and curation strategies, and many directly share the web-based IntAct curation infrastructure. This avoids redundant development while retaining the value of each individual resource. IMEx partners include UniProt (Switzerland and the United Kingdom), I2D (Canada), InnateDB (Ireland and Canada), Molecular Connections (India) and MechanoBio (Singapore). In 2013, the well-known MINT database, a long-standing IMEx partner, joined forces with IntAct and is now sharing the IntAct curation platform, while IntAct contributes to the Mentha resource (Orchard et al., 2014).

Reactome provides review-style, curated and peer-reviewed human pathways in a computationally accessible form. In 2013 we released a complete redevelopment of the Reactome web interface, providing tight integration with other resources. The new diagram viewer provides a more user-friendly visual interface to Reactome pathway data. Within the Reactome browser it is now possible to visualise molecular interactions from IntAct, gene expression data from the EBI Expression Atlas (see figure 1), 3D structures from PDBe, and chemical structures from ChEBI and Rhea for the selected pathway objects (Croft et al., 2014).

Our BioModels team contributed to a large-scale collaborative effort for the semi-automated generation of systems biology models based on KEGG pathways, Biocarta, MetaCyc and SABIO-RK data. More than 142 000 models, covering 1852 species, were generated, opening up new opportunities to explore and refine models. We rose to the challenge of releasing these models through BioModels Database in 2013, increasing the number of available models by two orders of magnitude (Büchel et al., 2013).

The Proteomics Services Team is also a major contributor to BioJS, a JavaScript-based library of reusable components for presenting biological data. In its early development phase, BioJS raised strong interest in the bioinformatics community, including an ISMB highlights track presentation of the reference publication (Gómez et al., 2013).

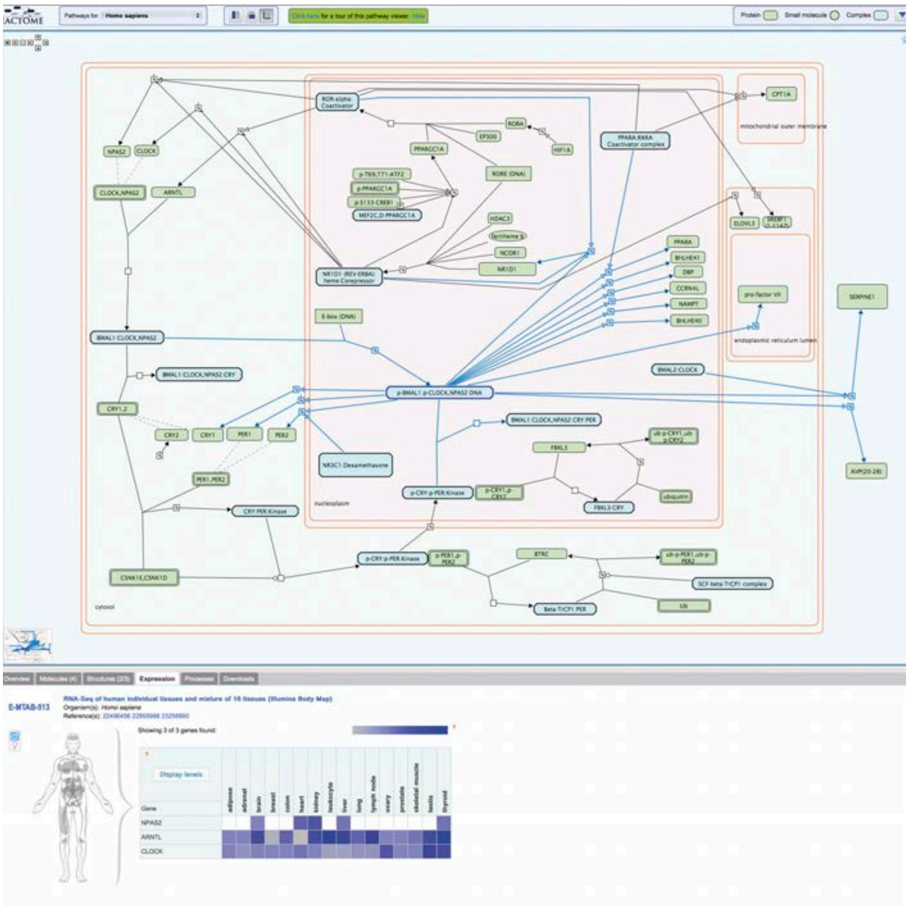
Future plans

The first stage of the redeveloped PRIDE database system, PRIDE Archive, will be released in early 2014, providing a stable base for further rapid growth of proteomics data. We will begin providing quality-controlled subsets and derived datasets from PRIDE, evolving this resource from a primary database into a systems biology source of protein expression data.

Henning Hermjakob

MSc Bioinformatics University of Bielefeld, Germany, 1995. Research Assistant at the German National Centre for Biotechnology (GBF), 1996.

At EMBL-EBI since 1997



The Reactome pathway browser visualisation of gene expression data for the current pathway, using a widget from the Expression Atlas.

We will also release the first version of the new EMBL-EBI Complex Portal, which provides a reference database of stable molecular complexes, integrated with and serving many biomolecular resources.

The Reactome web interface was updated in 2013 and in 2014 we plan to release a completely redeveloped set of Reactome pathway analysis tools, as well as much faster and more stable release building software.

The new JUMMP platform will be the underlying infrastructure for two modelling resources: the DDMoRe model repository for pharmacodynamic models and a 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.

We will continue to work with journals, editors and data producers to make more data publicly available by utilising community-supported standards.

Selected publications

Büchel, F., Rodriguez, N., Swainston, N., et al. (2013) Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Systems Biol. 7, 116.

Croft, D., Mundo, A.F., Haw, R., et al. (2014) The Reactome pathway knowledgebase. Nucleic Acids Res. 42, D472-477.

Gómez, J., García, L.J., Salazar, G.A., et al. (2013) BioJS: an open source JavaScript framework for biological data visualization. Bioinformatics (Oxford) 29, 1103-1104.

Griss, J., Foster, J.M., Hermjakob, H., et al. (2013) PRIDE Cluster: building a consensus of proteomics data. Nature Methods 10, 95-96.

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, D358-D363.

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. The scientific literature provides the essential context for research, and is increasingly linked to the underlying data. Our cross-domain tools and resources, including ontologies, are pivotal in harmonising these diverse data types and ensuring they manage their data in a consistent manner.

In 2013 the newly formed Samples, Phenotypes and Ontologies team played an important role in creating the new Resource Description Framework (RDF) platform. The new platform, built in response to input from industry, provides access to bioinformatics resources that support Semantic Web technologies.

Europe PubMed became EMBL-EBI's central literature resource in 2013, and is linked to all EMBL-EBI databases. Over 13 million articles in Europe PMC have been cited at least once, representing the largest public-domain citation network in the world.

Europe PubMed Central

Europe PMC contains about 26 million abstracts (including PubMed, patents, and Agricola records) and 2.6 million full text, life science research articles. Of these full text articles, over 500,000 are Open Access and can be downloaded from the Europe PMC FTP site. As well as a sophisticated search and retrieval across all this content, it provides information on how many times the articles have been cited and by whom, links to related data resources, and text-mined terms. Europe PMC labs showcases integrated text-mining tools. Principal Investigators on grants awarded by the 19 Europe PMC Funders can use the 'Europe PMC plus' site to self-deposit full text articles and link those articles to the grant that supported the work.

BioSamples Database

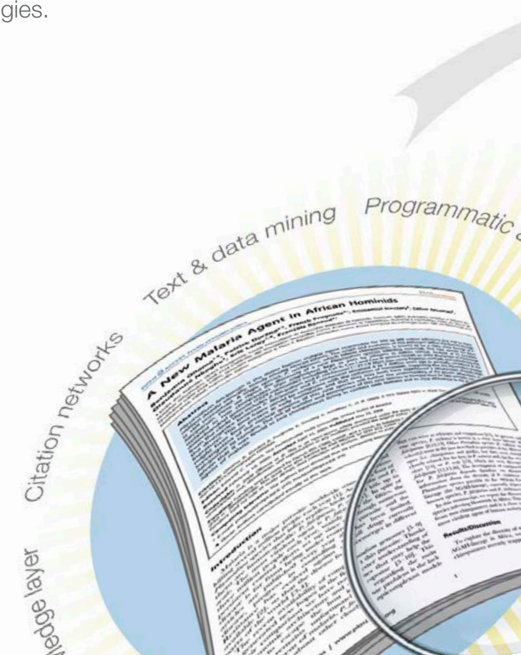
The EBI BioSamples Database holds information about biological samples, particularly samples referenced from other EMBL-EBI databases. A well curated set of reference samples is available and will be exchanged with NCBI. Reference layer biological samples are often reused in experiments, for example cell lines. Samples in this database can be referenced by accession numbers from data submissions to other EMBL-EBI resources.

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 is being mapped to Ensembl variation.

Gene Ontology

GO is a major bioinformatics initiative 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 EBI, the editors play a key role in managing the distributed task of developing and improving the GO ontologies.



Johanna McEntyre

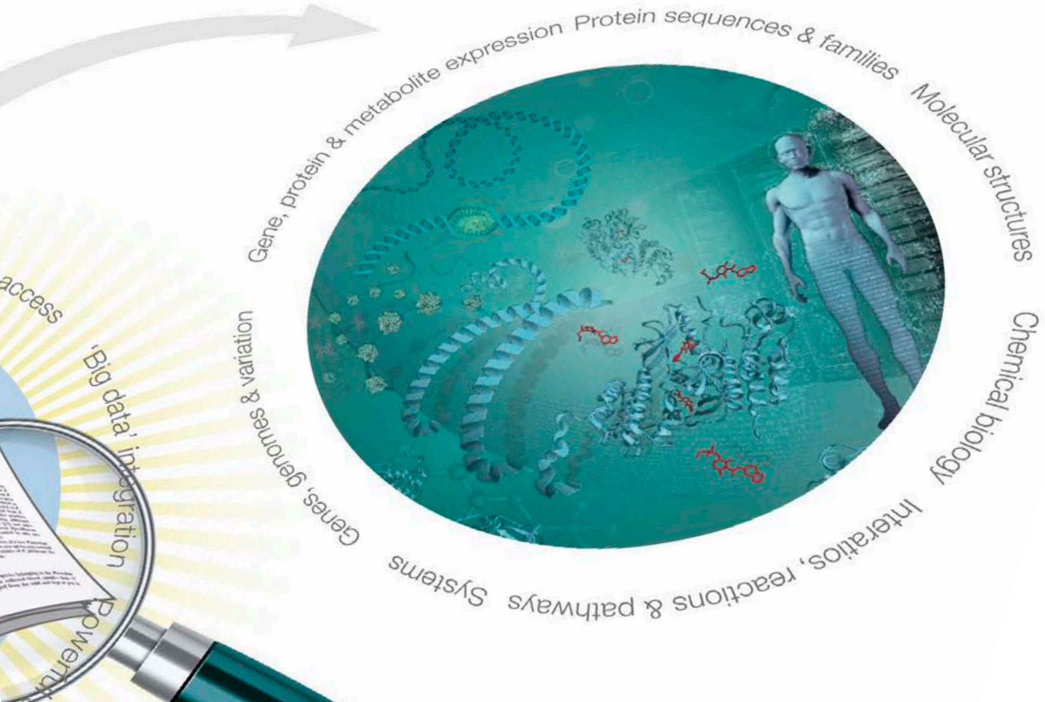
- *Officially retired CiteXplore, making Europe PubMed Central EMBL-EBI's central literature resource with links from all EMBL-EBI databases;*
- *Showed that over 13 million articles in Europe PMC have been cited at least once, representing the largest public-domain citation network in the world;*
- *Developed a new External Links Service that enables databases, institutional repositories (holding full text versions of articles), and text-miners to publish links from Europe PubMed Central to related resources;*
- *Extended our text-mining programme to include mining of citations to 11 key life-science databases and DOIs representing data in resources such as Data Dryad, Pangaea and FigShare;*
- *Launched a new advanced search, and introduced filter for information found in methods and figure legends;*
- *Pioneered the integration of ORCIDs, allowing people to claim their articles unambiguously, and then show these associations publicly.*

www.europepmc.org

Helen Parkinson

- *Launched a new Resource Description Framework (RDF) platform to provide access to resources that support Semantic Web technologies;*
- *Delivered the Gene Expression Atlas RDF and BioSamples RDF, along with a related BioConductor package/AtlasRDF that enables enrichment analysis of RDF data;*
- *Developed and implemented a new web portal for Infrafrontier, the infrastructure for mouse model data;*
- *Integrated the Gene Ontology with ChEBI;*
- *Included a genetic disease hierarchy in the Experimental Factor Ontology;*
- *Standardised the Uberon multispecies vertebrate anatomy ontology;*
- *Issued the first release of Zooma, an application for mapping annotations to ontologies that supports curatorial activities for EMBL-EBI service teams and external users;*
- *In the context of BioMedBridges and in collaboration with the Danish ELIXIR node, delivered a new resource providing access to tools, applications and services in the European service registry.*

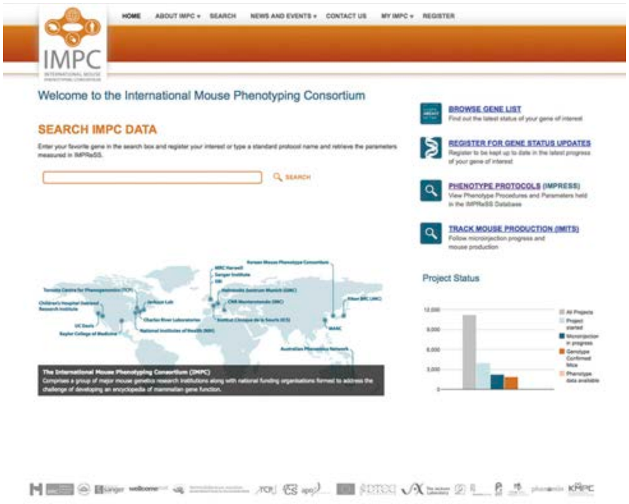
www.biomedbridges.eu
www.ebi.ac.uk/rdf
www.ebi.ac.uk/biosamples
www.ebi.ac.uk/efo
www.infrafrontier.eu
www.mousephenotype.org


Protein sequences & families
Molecular structures
Chemical biology
Interactios, reactions & pathways
Systems
Genes, genomes & variation
Gene, protein & metabolite expression
access
'Big data' integration
Powerf

Samples, phenotypes and ontologies

The team grew considerably in 2013 as the Gene Ontology Editorial Office, Mouse Informatics and Functional Genomics Production teams merged to form the Samples, Phenotype and Ontologies team.

The team grew considerably in 2013 as the Gene Ontology Editorial Office, Mouse Informatics and Functional Genomics Production teams merged to form the Samples, Phenotype and Ontologies team. Like Literature Services, the team operates in the cross-domain space between EMBL-EBI clusters. We provide ontology resources such as the Gene Ontology and Experimental Factor Ontology, 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.


HOME
ABOUT IMPC ▾
SEARCH
NEWS AND EVENTS ▾
CONTACT US
MY IMPC ▾
REGISTER
IMPC
SEARCH
Toronto Centre for Phenogenomics (TCP)
Children's Hospital Oakland Research Institute
UC Davis
Baylor College of Medicine
Jackson Lab
Charles River Laboratories
National Institutes of Health (NIH)
MRC Harwell
Sanger Institute
Helmholtz Zentrum Munich (GMC)
CNR Monterotondo (EMC)
Institut Clinique de la Souris (ICS)
Korean Mouse Phenotype Consortium
Riken BRC (JMC)
MARC
Australian Phenomics Network
12,000
9,000
6,000
3,000
0
All Projects
Project started
Microinjection in progress
Genotype Confirmed Mice
Phenotype data available
wellcome
TCP
ICS
KMPC

Welcome to the International Mouse Phenotyping Consortium

SEARCH IMPC DATA

Enter your favorite gene in the search box and register your interest or type a standard protocol name and retrieve the parameters measured in IMPReSS.

BROWSE GENE LIST
Find out the latest status of your gene of interest

REGISTER FOR GENE STATUS UPDATES
Register to be kept up to date in the latest progress of your gene of interest

PHENOTYPE PROTOCOLS (IMPRESS)
View Phenotype Procedures and Parameters held in the IMPReSS Database

TRACK MOUSE PRODUCTION (IMITS)
Follow microinjection progress and mouse production

The International Mouse Phenotyping Consortium (IMPC)
Comprises a group of major mouse genetics research institutions along with national funding organisations formed to address the challenge of developing an encyclopedia of mammalian gene function.

Project Status

The International Mouse Phenotyping Consortium produces knockout mice and carries out high-throughput phenotyping of each line in order to determine the function of every gene in the mouse genome. These mice are preserved in repositories and made available to the scientific community representing a valuable resource for basic scientific research and reducing the need for labs to produce their own knockout mice. Our team plays a central role in the project, with emphasis on data standards.

The team started three newly funded projects in 2013. PhenoImageShare (funded by the BBSRC in the United Kingdom) addresses the indexing of image annotation in the context of genomic data, so that images are accessible and queryable with biomolecular datasets. Embryonic Phenotyping (funded by the NIH in the United States) captures and integrates mouse embryonic images with genomic and phenotypic data. DIACHRON (funded by the European Commission under FP7) addresses both the use of RDF technology to represent scientific data and strategies for exploring data life cycles in the interests of enabling data preservation. The inception of these new projects expands our efforts in RDF generation and application building, ontology delivery and provision of community access to data.

Major achievements

In 2013 EMBL-EBI launched a new Resource Description Framework (RDF) platform in order to provide access to resources that support Semantic Web technologies (Jenkinson et al., 2013). As part of this effort our team delivered the Gene Expression Atlas RDF and BioSamples RDF, along with a related BioConductor package/AtlasRDF that enables enrichment analysis of RDF data.

We developed new infrastructure for the BioSamples database in 2013, including a new user interface and programmatic access. The infrastructure now provides integrated access to information about 2.8 million samples from life science experiments.

We developed and implemented anew web portal for Infrafrontier, the infrastructure for mouse model data. The portal improves access to mutant mice for the research community.

Our team's ontology achievements in 2013 included integration of the Gene Ontology with ChEBI (Hill et al., 2013), the inclusion of a genetic disease hierarchy in the experimental Factor Ontology and standardisation of the Uberon multispecies vertebrate anatomy ontology. We also issued the first release of Zooma, an application for mapping annotations to ontologies that supports curatorial activities for EMBL-EBI service teams and external users.

Our work on the BioMedBridges project led to the first version of a new resource providing access to tools, applications and services in Europe service registry in collaboration with the Danish ELIXIR node. This registry allows users to find and link to tools by type, area of application or interface type and is developed as part of the BioMedBridges and ELIXIR projects in collaboration with the Danish ELIXIR node.. The project increases visibility of tools and improves access for users and uses the EDAM ontology to describe tools and their functions.

Reflecting EMBL-EBI's activities in the cross-domain space, we published 10 papers in 2013 describing the team's data resources and collaborative projects.

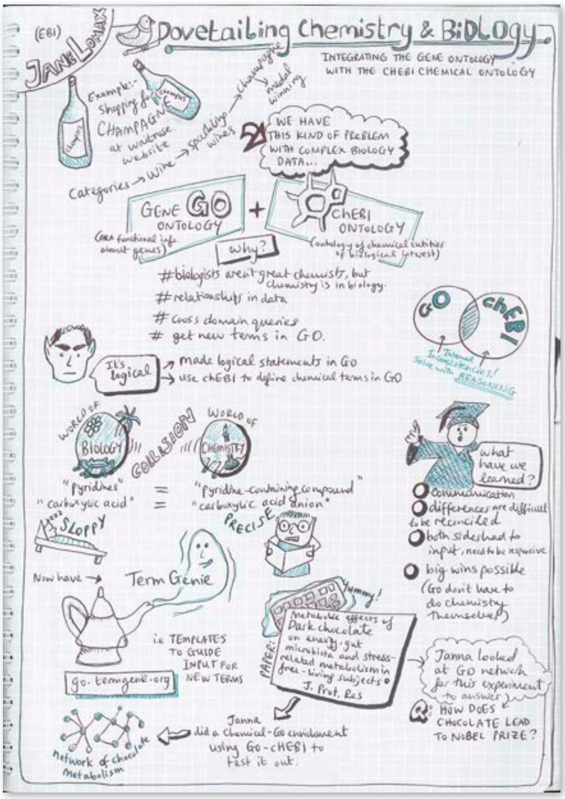
Helen Parkinson

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

At EMBL-EBI since 2000.

Future plans

In 2014 we will continue to work with the disease communities to improve the utility of the open disease ontologies for our data. We will apply these in the BioSamples database, delivering improved access to sample information for translational users. We will deliver a new web design for our major mouse data portal and integrate human disease and phenotype data with the mouse data resources. We will continue to improve our RDF and ontology offerings and start work on a new ontology browser for EMBL-EBI.


(EBI)
JANE LOMAX
Dovetailing Chemistry & BiOLOgy
INTEGRATING THE GENE ONTOLOGY WITH THE CHEBI CHEMICAL ONTOLOGY
Example:- Shopping for CHAMPAGNE at waitrose website
Champagne
Categories → wine → sparkling wines → champagne → medal winning
WE HAVE THIS KIND OF PROBLEM WITH COMPLEX BIOLOGY DATA...
GENE GO ONTOLOGY
(GO = functional info about genes)
+
ChEBI ONTOLOGY
(ontology of chemical entities of biological interest)
why?
#biologists aren't great chemists, but chemistry is in biology.
#relationships in data
# cross domain queries
# get new terms in GO.
It's logical
→ made logical statements in GO
→ use chEBI to define chemical terms in GO
GO
ChEBI
Internal Inconsistencies! Solve with REASONING
WORLD OF BIOLOGY
COLLISION
WORLD OF CHEMISTRY
"pyridines" = "pyridine-containing compound"
"carboxylic acid" = "carboxylic acid anion"
zzz SLOPPY
PRECISE
what have we learned?
communication
differences are difficult to be reconciled
both sides had to input, need to be responsive
big wins possible (GO don't have to do chemistry themselves)
Now have → Term Genie
i.e TEMPLATES TO GUIDE INPUT FOR NEW TERMS
go.termgenie.org
Yummy!
PAPER: Metabolic effects of Dark chocolate on energy, gut microbiota and stress-related metabolism in free-living subjects • J. Prot. Res
Janna looked at GO network for this experiment to answer
Q: HOW DOES CHOCOLATE LEAD TO NOBEL PRIZE?
Janna did a Chemical-GO enrichment using GO-CHEBI to test it out.
network of chocolate metabolism

Integrating the ChEBI ontology for cheminformatics and the Gene Ontology.

Selected publications

Ison, J., Kalaš , M., Jonassen, I., et al. (2013) EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics, and formats. Bioinformatics 29, 1325-1332.

Koscielny, G., Yaikhom, G., Iyer, V., et al. (2013) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res. DOI: 10.1093/nar/gkt977

Rustici, G., Kolesnikov, N., Brandizi, M., et al. (2013) ArrayExpress update–trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987-D990.

Lappalainen, T., Sammeth, M., Friedländer, M., et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511.

‘tHoen, P.A.C., Friedländer, M.R., Almlöf, J., et al. (2013) Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nature Biotech. 31, 1015-1022.

Roncaglia, P., Martone, M.E., Hill, D.P., et al. (2013) The Gene Ontology (GO) Cellular Component Ontology: integration with SAO (Subcellular Anatomy Ontology) and other recent developments. J. Biomed. Semantics 4, 20.

Hill, D.P., Adams, N., Bada, M., et al. (2013) Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics 14, 513.

Petryszak R., Burdett, T., Fiorelli, B., et al. (2013) Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 42, D926-D932.

The Gene Ontology Consortium (2013) Gene Ontology Annotations and Resources. Nucleic Acids Res. 41, D530-D535.

Literature services

Direct access to the scientific literature and the data that underlie it have become increasingly important as data-driven science continues to trend upwards. The Literature Services team addresses this in a number of ways, supporting the wider scientific research community and our data-provider colleagues at EMBL-EBI by providing valuable, multi-layer functionality in Europe PubMed Central.

Europe PubMed Central, now the sole literature database offered by EMBL-EBI, contains over 26 million abstracts and 2.7 million full text articles. The abstracts component includes all of PubMed, agricultural abstracts from Agricola and patents from the European Patent Office. About 700 000 of the full-text articles are open access, so they are free to read and to reuse in ways such as text mining. Europe PubMed Central is funded by 24 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. One of the key approaches we employ to meet this goal is to engage with individual scientists, text miners and database managers to understand how layers of value can be built upon the basic article content. We provide the infrastructure that enables the enrichment of the literature by individuals and computational methods developed throughout the community, and publish the results to maximise the usefulness of the core content and allow their widest possible reuse.

Summary of achievements

In 2013 we focused on the core strengths of our fast and powerful literature search system, and made several significant steps towards engaging the wider scientific community:

- *CiteXplore was officially retired in February 2013, and Europe PubMed Central adopted as EMBL-EBI's sole literature resource. All EMBL-EBI databases now direct literature links to Europe PubMed Central;*
- *We calculated citation counts of records in Europe PMC and showed that over 13 million articles have been cited at least once, representing the largest public-domain citation network in the world;*
- *We link the literature to EMBL-EBI data resources by using the references appended to data records by curators and submitters. Over 1.2 million articles are linked by this method;*
- *We developed a new External Links Service that enables databases, institutional repositories (holding full text versions of articles), and text-miners to publish links from Europe PubMed Central to related resources. These are displayed on the website and distributed in web services;*
- *Our core text-mining programme identifies gene symbols, organisms, diseases, Gene Ontology terms and chemicals in full text articles. In 2013 we extended the programme to include mining of citations to 11 key life-science databases and DOIs representing data in resources such as Data Dryad, Pangaea and FigShare;*
- *We launched an easy-to-use Advanced Search that employs autosuggest features, allowing users to fully exploit Europe PMC's powerful search;*
- *We introduced the function of limiting searches to full-text sections such as methods or figure legends, and use this feature ourselves in the core full-text search to avoid returning hits to articles that only mention the term in the reference list;*
- *We pioneered the integration of ORCIDs, allowing people to claim their articles unambiguously, and then show these associations publically in Europe PMC. We also made searching by author name significantly more accurate in support of authors identifying of articles through the new Europe PubMed Central linking tool;*
- *We initiated a Text Mining for Service Provision interest group, which meets monthly to explore ways in which text-mining can assist database curators and enrich article content for general use.*

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.

Major achievements

We have continued to build our article and data citation networks, incorporating 13 million cited articles and 1.2 million articles linked to database records.

In 2013 we improved author search behaviour and integrated ORCIDs into literature search mechanisms. This work was presented as a new opportunity early in the year and proved of high impact across the user base, with over 600 000 articles in Europe PMC having an associated ORCID and almost 10,000 ORCIDs having been linked to articles using the Europe PMC claiming tool.

The External Links Service launched in 2013 allows text miners to publish results in the context of Europe PMC. A very welcome side effect of this development is that it doubles as a method to link with repositories such as the French HAL open articles repository, extending the number of freely available articles for our users.

We have introduced an Advanced Search, which makes it easier for users to use the full power of Europe PMC search, as well as introducing the section-level search which means that searches can be restricted to sub-components of full-text articles such as figure legends, methods or discussion sections.

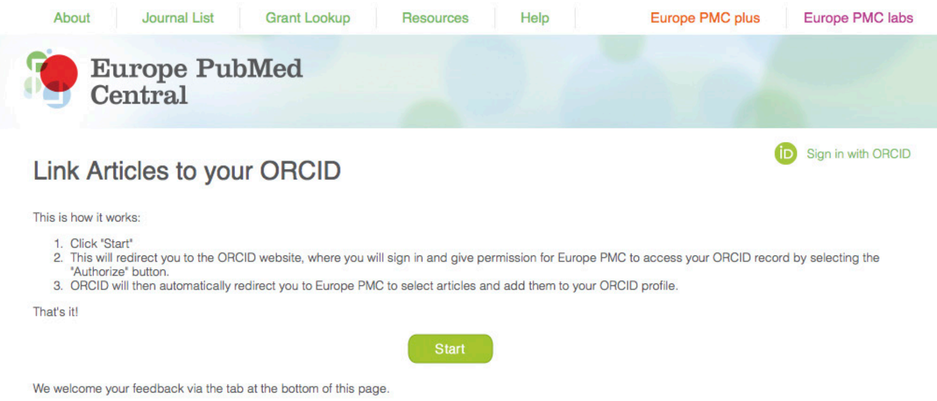
Text mining data citations from full text articles and providing these as various services has enriched EMBL-EBI databases with crosslinks, and has had impact in the wider publishing community in the area of Altmetrics.

In December we organised a Wellcome Trust Retreat on Literature-Data Integration, which was a very well-attended and fruitful event.

Future plans

In 2014 we plan to develop functionality for individual users to set up an account in which they can save searches, set up content alerts and specify preferences for which External Links they wish to see. This will form the basis of an emerging social layer built on top of Europe PubMed Central content, and will incorporate the use of ORCIDs as part of the development.

We will continue to improve the search behaviour of Europe PMC, enriching the full text search in particular and extending this to books and reports. We will also build on our efforts to integrate data from life-science databases throughout Europe with associated literature in Europe PMC, and share this information widely to assist database development and curation efforts at EMBL-EBI.



Link Articles to your ORCID

This is how it works:

1. Click 'Start'
2. This will redirect you to the ORCID website, where you will sign in and give permission for Europe PMC to access your ORCID record by selecting the "Authorize" button.
3. ORCID will then automatically redirect you to Europe PMC to select articles and add them to your ORCID profile.

That's it!

We welcome your feedback via the tab at the bottom of this page.

An abstract in Europe PubMed Central, illustrating ORCIDs associated with the article.

Research highlights

DNA storage

A new, scalable method to store data in the form of DNA makes it possible to store at least 100 million hours of high-definition video in about a teacup full of synthetic DNA.

Comprehensive map of functional genetic variation in humans

The largest-ever dataset linking human genomes to gene activity at the level of RNA was published by the GEUVADIS consortium

Where do turtles 'fit' in the evolutionary tree?

Researchers in the International Turtle Genome Consortium looked at how the Chinese soft-shell turtle’s genome behaves during different stages of development, and unearthed several intriguing facts about these unusual creatures.

The Single Cell Genomics Centre

The Wellcome Trust Sanger Institute and EMBL-EBI launched the Single Cell Genomics Centre, which seeks to answer key biological questions by exploring cellular genetics at the highest resolution possible.

Systematic assessment of RNA analysis software

A systematic assessment of RNA-seq programs, published in *Nature Methods*, may inspire new computing approaches to handle technologies for gene expression analysis.

Perfect proteins preferred

New research, published in *Cell*, shows that the specific order in which proteins assemble into complexes is extremely important to how they function.

Mapping metabolism

A comprehensive map of human metabolism, published in *Nature Biotechnology* and freely available through the BioModels Database, contains verified information on thousands of metabolites and reactions. The resource is a goldmine for systems biologists.

Exploring survival

SurvCurv, published in *Ageing Cell*, is the first comprehensive, open resource to enable the large-scale analysis of survival data in model organisms.

Wired for change

A new study of gene expression, published in *Cell*, reveals the first steps of evolution in gene regulation in mice.

The mutation police

How does our genome keep mutations in check? New research published in *Cell* shows how a protein called hnRNP C suppresses pseudoexons, providing insights into the development of diseases such as cancer and neurodegeneration.

How is a fish like a person?

An international team of researchers showed exactly how our genes line up with those of the zebrafish.

Research summaries

Beltrao group

- Received a HFSP Career Development Award to study the evolutionary dynamics, specificity and functional relevance of post-translational regulatory networks;
- In collaboration with the Villen lab (Washington University), studied the phosphorylation and ubiquitylation of S. cerevisiae proteins in order to analyse the cross-talk between the two modifications.

Bertone group

- Determined a novel component of the pluripotency regulatory network downstream of the LIF signalling pathway;
- Characterised the function of the non-coding RNA 7SK in the regulation of pluripotency genes in mouse embryonic stem cells;
- Explored the potential efficacy of epigenetic therapy in reducing tumour malignancy of glioblastoma multiforme using a cancer stem cell model;
- Identified a series of small-molecule inhibitors to polo-like kinase 1 that selectively disrupt mitosis in glioblastoma stem cells.

Birney group

- Along with Nick Goldman, published a high-profile paper on a method to store digital information in DNA;
- Staff scientist Ian Dunham released the FORGE analysis tool for functional element overlap analysis of GWAS experimental results, which identifies tissue-specific signals within the set of GWAS SNPs. The initial functional elements considered are DNAse1 hotspots from ENCODE or Roadmap Epigenomics projects generated by the Hotspot method;
- 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

- Completed the analysis of human gene-expression datasets in order to establish the transcript composition and showed that, in a given condition, most protein-coding genes have one major transcript expressed at significantly higher level than others;
- Demonstrated that in human tissues, the major transcripts contribute almost 85% to the total mRNA from protein-coding loci, and that often the same major transcript is expressed in many tissues;
- Completed our analysis of differential transcript expression and fusion genes in the RNAseq data generated by the EU-funded GEUVADIS consortium;
- As part of the EurcanPlatform consortium, participated in the identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing;
- In the CAGEKID consortium, completed differential gene and transcript expression and fusion gene analysis in 50 pairs of cancer and normal tissues in kidney cancer.

Enright group

- Developed Kraken, a new set of computational tools for small RNA sequencing data analysis;
- Unravelled new epigenetic mechanisms for piwiRNA control of gene silencing in the mouse germline, in collaboration with the O'Carroll lab at EMBL Monterotondo;
- Explored the role of the lin-28 gene and let-7 microRNA in malignant germ-cell tumours with collaborators at Addenbrookes hospital;
- Developed new tools for statistical analysis of small RNA 5' modification.

Flicek team

- In our exploration of the rapid evolution of tissue-specific transcriptional regulation in closely related mammals, showed how collections of transcription factors evolve rapidly (David Thybert);
- Using similar cell lines from multiple primate species, explored how the only known vertebrate insulator, CTCF, is stabilised evolutionarily by its interaction with the co-binding transcription factor YY1 (Petra Schwalie);
- Explored systematic functional annotation of noncoding sequence variants (Graham Ritchie, ESPOD).

Goldman group

- Described a method for storing information in DNA, with robust error correction, suitable for large-scale and long-term information storage;
- Released the 'rlsim' RNA-seq library construction simulation framework;
- Developed novel computational methods to perform evolutionary analysis of ovarian cancer progression within individuals;
- Studied robustness of orthology inference and evaluated quality of functional annotations;

- *Performed a comprehensive study of multiple sequence alignment benchmarking strategies.*

Marioni group

- *Helped co-ordinate development of the Sanger-EMBL-EBI Single Cell Genomics Centre with the Teichmann lab and others;*
- *Developed some of the first methodological tools for handling single-cell RNA-seq data (Brennecke et al., 2013; Kim and Marioni, 2013; Pettit & Marioni, 2013);*
- *Developed new statistical tools for handling high-dimensional data with multiple sources of correlation;*
- *Built new collaborations with the Logan group at the Sanger to study olfaction.*

Overington team

- *Completed ADME SARfari, a comparative genomics system of drug-metabolising systems that enables identification of the absorption, distribution, metabolism, and excretion protein targets with which a compound is likely to interact.*

Saez-Rodriguez group

- *Developed methods within our platform CellNOpt to efficiently model computationally very large scale signalling networks;*
- *Co-organised the eighth edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods);*
- *Developed methods to predict drug response in cancer using genomic and chemical data.*

Stegle Group

- *Developed and applied new statistical tools to integrate molecular phenotype data and global phenotypes in systems-level models;*
- *Joined the Human Induced pluripotent Stem Cell Initiative;*
- *Successfully acquired EMBO funding to run a new EMBO training course on genotype-phenotype mapping, starting in summer 2013.*

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

- *Smoothly transitioned the Teichmann group from the MRC Laboratory of Molecular Biology to the Genome Campus;*
- *Co-founded the Sanger-EBI Single Cell Genomics Centre and published the first paper on our single cell RNA-sequencing work in collaboration with the Heisler and Marioni groups (Brennecke et al., 2013);*
- *In collaboration with the Robinson group, showed that heteromeric protein complexes have well-defined assembly pathways that are constrained in evolution (Marsh et al., 2013).*

Thornton group

- *Published our analysis of human nsSNP data from the 1000 Genomes Project, developing a pipeline to map the observed mutations onto protein sequences and structures.*
  Demonstrated a radical difference between the distribution of variants observed in the 1000 Genomes data and those in OMIM. Further analysis of cancer mutations in COSMIC is underway for comparison;
- *Published our database of survival curves, SurvCurv, and used it to perform meta-analyses of the influence of strain, sex and bacterial infection over many independent experiments;*
- *Published LigSearch, a knowledge-based web server to identify likely ligands for a protein target. This was developed to facilitate crystallisation of proteins by adding an appropriate ligand to the crystallisation medium;*
- *Completed 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. An equivalent analysis of the isomerases is underway;*
- *Completed the EC-BLAST software tool, which compares enzyme reactions and produces a hit list of the most similar reactions. The manuscript was accepted for publication.*

Research achievements 2013

Approximately one quarter of the European Bioinformatics institute is devoted to investigator-led, blue-skies research. Sixteen principal investigators, each with a small group of pre- and post-docs researchers, perform basic research using computational approaches to understand the processes of life. As the role of bioinformatics continues to expand, our researchers are making major contributions to life science research in many different ways.

A unique environment for research

EMBL-EBI research is well supported by EMBL's member states and our external funders (see page 36), and benefits from a critical mass of expertise in molecular biology on the Wellcome Trust Genome campus. Our institute is well known for its careful stewarding of global data resources, and the research we do is conducted in an environment of great scientific and technical excellence. The topics we explore touch on all aspects of the life sciences, and our working methods are highly collaborative and interdisciplinary.

Predocs and postdocs

In 2013 we hosted 37 PhD students, all enrolled at the University of Cambridge, and welcomed seven newcomers. Four of our students successfully defended their doctoral theses and a further seven submitted theirs for examination in 2014. Our institute is fortunate to have over 60 post-doctoral researchers, who contribute substantially to our vibrant, multidisciplinary environment and benefit from excellent computing facilities and systems support.

New experimental data

Some of our computational research provides insights into living systems through analyses of new experimental data, and these projects are conducted in collaboration with researchers in EMBL and other leading institutes throughout the world.

A new study of gene expression by Schwalie and colleagues in the Flieck team reveals the first steps of evolution in gene regulation in mice. This has implications for the study of differences in gene regulation between people. Similarly, a collaboration between the Teichmann group and researchers at Oxford and the MRC Laboratory of Molecular Biology showed that the specific order in which proteins assemble into complexes is driven by evolution, and is extremely important to how they function (Marsh et al., 2013).

Paul Bertone and colleagues, working to understand how stem cells 'mature', determined a novel component of the pluripotency regulatory network, downstream of the leukemia inhibitory factor (LIF) signalling pathway, and characterised the function of the non-coding RNA 7SK in the regulation of pluripotency genes in mouse embryonic stem cells. Their work has shed important light on a critical factor in the maintenance of embryonic stem-cell self-renewal.

New tools for research

Some of our research results in new tools and resources that we share with scientists worldwide. For example, Ziehm and colleagues created a new resource that can help scientists study ageing. Published in the journal *Ageing Cell*, SurvCurv is the first comprehensive, open resource to enable the large-scale analysis of survival data in model organisms.

In 2013 Enright and his colleagues developed Kraken, a new set of computational tools for small RNA sequencing data analysis, and for the statistical analysis of small RNA 5' modification. These tools, already in use by a number of laboratories, can be applied to mRNA sequencing datasets.

A key technology development over the past few years has been the emergence of machines to measure single RNA-seq data. Marioni and colleagues developed some of the first methodological tools for handling such data (Brennecke et al., 2013; Kim and Marioni, 2013; Pettit and Marioni, 2013).

In a completely different area of biological research, Thornton and colleagues developed EC-BLAST, which compares enzyme reactions and produces a 'hit list' of the most similar reactions (Rahman et al., 2013). The program makes it possible to quickly compare the functions of thousands of catalysts, facilitating research into anything from drug interactions to the efficient production of biofuels.

Similarly, the Overington teams's new ADME SARfari, a comparative genomics system of drug-metabolising systems, enables identification of the absorption, distribution, metabolism and excretion protein targets with which a compound is likely to interact.

One of the challenges of systems biology is to model very large scale signalling networks. To address this, Saez-Rodriguez and colleagues developed new, highly efficient methods within their CellNOpt platform. CellNOpt software uses information on signaling pathways encoded as a Prior Knowledge Network, and trains it against high-throughput biochemical data to create cell-specific models

Novel methodologies

Some research is about testing the power of new algorithms and their ability to make predictions or to analyse data effectively. Such work is technically demanding and critical for the advancement of science. A systematic assessment of RNA-seq programs, published by Bertone and colleagues in *Nature Methods*, was performed as part of the RNA-seq Genome Annotation Assessment Project (RGASP), an ENCODE-affiliated initiative. This work determined which approaches work well for specific tasks, and which areas can be improved. This may inspire new computing approaches to handle current and future technologies for gene expression analysis.

In a similar vein, Goldman and colleagues studied the robustness of orthology inference and the quality of functional annotations as well as a comprehensive study of multiple sequence alignment benchmarking strategies.

New perspectives

Some research uses the power of computational approaches to combine and overview our current knowledge and so derive new insights. As part of the EU-funded GEUVADIS consortium, Brazma and colleagues presented the largest-ever dataset linking human genomes to gene activity at the level of RNA. GEUVADIS, comprising over 50 scientists from nine European institutes, measured gene expression by sequencing RNA in human cells from 462 individuals whose full genome sequences had already been published as part of the 1000 Genomes Project. Their 2013 study adds a functional interpretation to the most important catalogue of human genomes. The data are freely available in ArrayExpress.

Human health

Increasingly, the attention of our researchers has moved into the medical domain, with many new collaborations developing and producing strong publications of medical relevance.

In 2013 Bertone and his colleagues published work exploring the potential efficacy of epigenetic therapy in reducing tumour malignancy of glioblastoma multiforme using a cancer stem cell model. Goldman and colleagues developed novel computational methods to perform evolutionary analysis of ovarian cancer progression within individuals, and Saez-Rodriguez, in collaboration with colleagues at the Wellcome Trust Sanger Institute, developed methods to predict drug response in cancer using genomic and chemical data.

Re-purposing DNA

EMBL research groups have absolute acadmic freedom and their research projects are driven by curiosity. This environment extends our work well beyond data analysis. Using their knowledge of DNA and information processing, Goldman and Birney created a novel, scalable method for storing digital information in the form of DNA, a material that lasts for tens of thousands of years. Published in the journal *Nature*, the new data archiving method makes it possible to store at least 100 million hours of high-definition video in about a cup of DNA.

Single-cell genomics

In 2013, together with the Wellcome Trust Sanger Institute, we launched the Single Cell Genomics Centre (SCGC), which seeks to answer key biological questions by exploring cellular genetics at the highest resolution possible. The centre will focus on the exploration of cell function in normal development and immune function as well as cancer. Several of our researchers are deeply involved in this initiative and are already generating experimental data and new methods.

Looking forward

EMBL-EBI researchers were highly productive in many areas of biology during 2013, from development to ageing and disease, contributing novel findings, novel algorithms and tools and making creative contributions in synthetic biology.

Much of this research is supported through competitive grants, which largely fund collaborative research with external colleagues. We look forward to embracing new challenges in research arising from recent developments in sequencing, proteomics, metabolomics and imaging technologies. We will endeavour to provide ever deeper insights into how life at the molecular level shapes the life of an organism. We look forward with excitement to what new discoveries 2014 will bring.

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 the function and evolution of post-translational regulatory networks and the evolution of genetic and chemical–genetic interactions. Looking beyond evolutionary processes, 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 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.

Summary of progress

- *The group was awarded a HFSP Career Development Award to study the evolutionary dynamics, specificity and functional relevance of post-translational regulatory networks;*
- *In collaboration with the Villen lab (Washington University) we studied the phosphorylation and ubiquitylation of S. cerevisiae proteins in order to analyse the cross-talk between the two modifications.*

Major achievements

During 2013 the group was awarded an HFSP Career Development Award to study the evolutionary dynamics, specificity and functional relevance of post-translational regulatory networks. This funding will be directed at further developing a cross-species repository of PTMs, computational methods to predict PTM regulatory networks and methods to study the function, dynamics and specificity of PTM signalling.

The group was also awarded an EIPOD postdoctoral fellowship (for Brandon Invergo) to study the PTM regulatory networks in the malaria parasite Plasmodium. The project will be developed in collaboration with the Wellcome Trust Sanger Institute groups of Jyoti Choudhary (proteomics) and Oliver Billker (malaria research).

PTMs are known to act in concert to modulate protein function. Given that the function of most PTMs is currently unknown, developing methods to study such cross-talks between different PTMs is crucial. In 2013, in collaboration with Judit Villen's group from the University of Washington, we studied the cross-talk between phosphorylation and ubiquitylation in S. cerevisiae. We used a new mass-spectrometry approach to simultaneously identify both types of modifications and to quantify changes in these PTMs after proteasome inhibition. Phosphorylation sites found in ubiquitylated proteins were more likely to match known phospho-degron motifs and to be closer in sequence and structural distance. Analysis of these data allowed for the prediction of novel cross-regulatory events. We also identified a specific region within protein kinases that is often found to be ubiquitylated. This suggests that ubiquitylation may cross-regulate phosphosites by affecting the activity of kinases.

Pedro Beltrao

PhD in Biology, University of Aveiro (research conducted at EMBL-Heidelberg), 09/2007. Post-doctoral research at the University of California San Francisco.

Group leader at EMBL-EBI since 2013..

Future plans

In 2014 group will continue to study the evolution of cellular interactions networks with a specific focus on post-translational regulatory networks, genetic and chemical-genetic networks. We will continue to explore the evolution of function, specificity and conditional regulation of PTM signalling as well as the evolution and conditional regulation of genetic interactions networks. In addition, we are developing methods to study how the genetic variability observed in different individuals of the same species might impact on the response to environmental conditions or drugs.

Selected publications

Beltrao, P., Bork, P., Krogan, N.J. and van Noort, V. (2013) Evolution and functional cross-talk of protein post-translational modifications. Mol Sys Biol 9, 714.

Swaney, D.L., Beltrao, P., Starita, L,, et al. (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation Nat Methods 10, 676-682.

Beltrao, P., Albanèse, V., Kenner, L.R., et al. (2012) Systematic functional prioritization of protein post-translational modifications. Cell 150, 413-425.

Ryan, C.J., Roguev, A., Patrick, K., et al. (2012) Hierarchical modularity and the evolution of genetic interactomes across species. Mol Cell 46, 691-704.

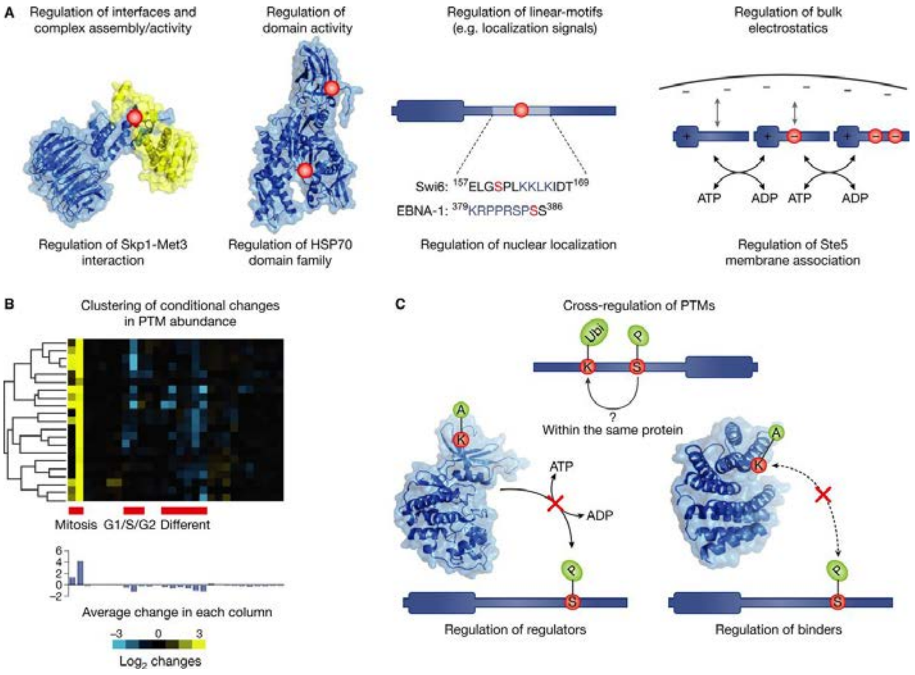

A
Regulation of interfaces and complex assembly/activity
Regulation of domain activity
Regulation of linear-motifs (e.g. localization signals)
Regulation of bulk electrostatics
Swi6: 157ELGSPLKKLKIDT169
EBNA-1: 379KRPPRSPSS386
ATP
ADP
ATP
ADP
Regulation of Skp1-Met3 interaction
Regulation of HSP70 domain family
Regulation of nuclear localization
Regulation of Ste5 membrane association
B
Clustering of conditional changes in PTM abundance
Mitosis
G1/S/G2
Different
6
4
2
0
−2
Average change in each column
−3
0
3
Log2 changes
C
Cross-regulation of PTMs
Ub
P
K
S
?
Within the same protein
A
K
ATP
ADP
P
S
Regulation of regulators
Regulation of binders

Figure. 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]

Bertone group

Pluripotency, reprogramming and differentiation

We investigate the cellular and molecular attributes of embryonic and tissue-specific stem cells using a combination of experimental and computational methods. We develop and apply genomic technologies to the analysis of stem cell function to address fundamental aspects of development and disease.

Our group focuses on two areas: the misregulation in cancer of fundamental processes that regulate cell differentiation, and tumourigenesis of neural cancer stem cells.

Embryonic stem (ES) cells are similar to the transient population of self-renewing cells within the inner cell mass of the pre-implantation blastocyst (epiblast), which are capable of pluripotential differentiation to all specialised cell types comprising the adult organism. These cells undergo continuous self-renewal to produce identical daughter cells, or can develop into specialised progenitors and terminally differentiated cells. Each regenerative or differentiative cell division involves a decision whereby an individual stem cell remains in self-renewal or commits to a particular lineage. The properties of proliferation, differentiation and lineage specialisation are fundamental to cellular diversification and growth patterning during organismal development, and in the initiation of cellular repair processes. The fundamental processes that regulate cell differentiation are likely to be misregulated in cancer, and are not yet well understood.

We study neural cancer stem cells derived from human glioblastoma multiforme tumours. Using neural stem (NS) cell derivation protocols, it is possible to expand tumour-initiating, glioblastoma-derived neural stem (GNS) cells continuously in vitro. Although the normal and disease-related counterparts are highly similar in morphology and lineage marker expression, GNS cells harbour genetic mutations typical of gliomas and give rise to authentic tumours following orthotopic xenotransplantation. We apply genomic technologies to determine transcriptional changes and chromosomal architecture of patient-derived GNS cell lines and their individual genetic variants. These data provide a unified framework for the genomic analysis of stem-cell populations that drive cancer progression, and contribute to the molecular understanding of tumourigenesis.

Major achievements

ES cells are propagated in a pluripotent, self-renewing state through the leukemia inhibitory factor (LIF) signalling pathway. LIF is the primary cytokine responsible for promoting self-renewal of ES cells in the absence of mitotically inactivated fibroblasts, and activates the Janus kinase (Jak) and Signal Transducer and Activator of Transcription (Stat) pathway through transactivation of Stat3. ES cells are dependent on this pathway and exhibit enhanced self-renewal in response to LIF signalling. Culturing ES cells in chemically defined media in the presence of two inhibitors of the Gsk3 and MEK/Erk pathways allows us to culture pluripotent ES cells in the absence of LIF, taking advantage of this to modulate LIF exposure in conditions designed to reveal Stat3-mediated transcriptional regulation. We performed comprehensive RNA-seq analysis in ES cells to identify Stat3 targets upon transactivation by LIF signalling. Transcription factors of interest induced upon LIF exposure include Tfcp2l1, Klf4, and Gbx2. ChIP-seq analysis to identify genome-wide binding patterns of Klf4, Tfcp2l1, and Stat3 reveal many shared regulatory targets, and functional assays demonstrate that undifferentiated ES cells can be maintained in the absence of LIF through overexpression of either Klf4 or Tfcp2l1. We found that Tcfcp2l1 phenocopies LIF stimulation, thus identifying a critical factor in the maintenance of ES cell self-renewal.

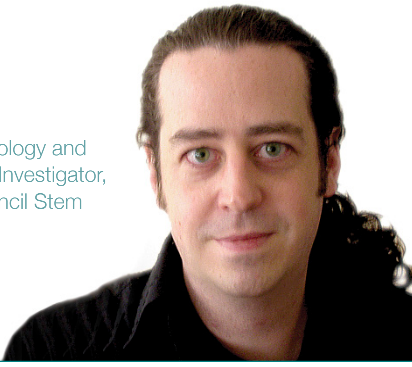
The application of genomic approaches allows the identification of novel regulators addition to transcription factors and chromatin modifiers. In collaboration with the Kouzarides lab at the Gurdon Institute we identified a non-coding RNA, 7SK, to be associated with repression of lineage-specification genes in ES cells. In pluripotent cells, lineage-specification genes are poised for transcription via a mechanism involving the bivalent methylation of H3K4 and H3K27. Antisense knockdowns coupled with strand-specific RNA-seq profiling revealed that 7SK prevents failed transcriptional termination and represses a specific cohort of bivalent genes in ES cells. Moreover, we found that 7SK controls transcriptional directionality by suppressing divergent upstream antisense transcription, again at specific loci. 7SK may therefore act in concert with other transcriptional regulators to modulate and refine expression of pluripotency and lineage-specification genes that establish the balance between self-renewal and differentiation.

The dominant model of cancer progression is the multi-step accumulation of genetic changes that activate proto-oncogenes and silence tumour suppressors. Cancer cells are driven by both genetic and epigenetic changes; however, the relative contribution of these influences in promoting the malignant state has been difficult to determine due to lack of experimental tools that enable resetting of epigenetic abnormalities. Together with Steve Pollard's lab at University College London, we used the induced pluripotent stem (iPS) cell methodology to invoke global epigenetic erasure in GNS cells, demonstrating

Paul Bertone

PhD Yale University, 2005.
Joint appointments in EMBL Genome Biology and Developmental Biology Units. Associate Investigator, Wellcome Trust—Medical Research Council Stem Cell Institute, University of Cambridge.

At EMBL–EBI since 2005.



that highly malignant and aneuploid GNS cells can be epigenetically reprogrammed. These glioblastoma-iPS cells (GiPSCs) were subsequently re-differentiated to assess the impact on tumourigenicity upon restoring normal epigenetic marks on a stable cancer genome. Despite widespread epigenetic alterations, GiPSCs-derived neural progenitors remained highly malignant. Only when cells were directed to non-neural types did we observe sustained expression of reactivated tumour suppressors and reduced infiltration. These results suggest that imposing an epigenome associated with an alternative developmental lineage can suppress some aspects of malignant behaviour. However, in the context of native cell lineages, resetting GBM-associated epigenetic abnormalities is not sufficient to override the inherent regulatory defects imparted by the cancer genome.

Characterising the cancer stem cell compartment of GBM tumours to understand how they differ from normal-tissue stem-cell counterparts can identify new therapeutic opportunities. With the Pollard lab we combined live-cell microscopy with software image processing to set up a system for high-throughput drug screening in GNS cells. In this approach, cells are cultured in multi-well plates in the presence of various small-molecule inhibitors, and imaged in a self-contained microscopy incubator over several days to produce time-lapse image stacks. Feature detection is then applied to identify events such as apoptosis and different phases of cell division, using a pre-defined training set. Upon exposure to a panel of 160 commercially available kinase inhibitors, we identified several compounds that induce mitotic arrest at prometaphase. This phenotype is apparent in GNS cells but not normal NS cells, and triggered by suppression of polo-like kinase 1 (Plk1). Sensitivity to Plk inhibitors may be explained by a lack of a p53-mediated compensatory pathway, as we observed a similar phenotype in p53 knockout NS cells. GBM stem cells may therefore be acutely susceptible to Plk1 inhibition, which might further be exploited as therapeutic agents to disrupt invasive cell proliferation.

Future plans

We have in place the most robust and stable systems for stem-cell derivation and propagation, where controlled experiments can be performed in well-defined conditions. This is particularly valuable for studying cell populations that would normally be inaccessible in the developing embryo. However, to realise the potential of ES cells in species other than mouse, precise knowledge is needed of the biological state of these cells, particularly the molecular processes that maintain pluripotency and direct differentiation. We are working to translate the knowledge and methods that have been successful in mouse ES cell biology to other mammalian species. This involves characterising germline-competent ES cells from the laboratory rat, along with the production of pluripotent human iPS cells using alternative reprogramming strategies. Through deep transcriptome sequencing we have already shown a broad equivalence in self-renewal capacity and cellular state, albeit with intriguing species-specific differences. Thus, while ground-state pluripotency can be captured and maintained in several species, the mechanisms used to repress lineage differentiation may be fundamentally different.

Tumour-based cancer studies are limited by a number of factors, including cellular diversity of tissue biopsies, lack of corresponding reference samples and inherent restriction to static profiling. Cancer stem cells constitute a renewable resource of homogeneous cells that can be studied in a wide range of experimental contexts and provide key insights toward new therapeutic opportunities. We are carrying out in-depth analyses of our GNS cell bank using comprehensive genetic and transcriptomic profiling, and using the data to develop methods to stratify glioblastoma classes based on the molecular attributes and differentiation capacities of tumour-initiating stem cells. Existing tumour subtypes are associated with diverse clinical outcomes and therefore important for prognostic value, but as with any analysis of complex tissues, previous results have suffered from sample heterogeneity. With access to the underlying stem-cell populations derived from parental tumours, we are refining existing subtype classification to improve the diagnostic utility of this approach. We are also performing functional experiments to identify alterations in GNS cells that impart tumourigenic potential.

Selected publications

Martello, G., et al. (2013). Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. EMBO J. 32, 2561-2574.

Castelo-Branco, G., et al. (2013). The non-coding snRNA 7SK controls transcriptional termination, poising, and bidirectionality in embryonic stem cells. Genome Biol. 14, R98.

Stricker, S.H., et al. (2013) Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. Genes Dev. 27, 654-669.

Danovi, D., et al. (2013) A high-content small molecule screen identifies sensitivity of glioblastoma stem cells to inhibition of polo-like kinase 1. PLoS One 8, e77053.

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 modest-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.

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. 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).

Summary of progress

- *Along with Nick Goldman, published a high-profile paper on a method to store digital information in DNA;*
- *Staff scientist Ian Dunham released the FORGE analysis tool, which performs functional element overlap analysis of the results of Genome-Wide Association Study (GWAS) Experiments, to identify tissue specific signals within the set of GWAS SNPs. In the initial implementation, the functional elements considered are DNAse1 hotspots from ENCODE or Roadmap Epigenomics projects generated by the Hotspot method;*
- *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.*

Major achievements

Our group carries out two major types of projects, one of which is a series of molecular and other phenotype association studies. Initial results from our work with Eileen Furlong's group at EMBL on the analysis of Drosophila expression in the Drosophila Genetic Reference Panel (DRGP) have been promising, and we hope to publish them in 2014.

We are working in a variety of human systems to explore association of molecular events in both normal and disease samples. The latter is a collaboration with Francis Collins at the National Institutes of Health in the United States 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, to sequence-level data. We collaborate as part of the UK10K project in this area. We use both genome-wide association data (the FORGE analysis server is one output of this) and Cancer data (as part of the BASIS project).

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.

Ewan Birney

PhD 2000, Wellcome Trust Sanger Institute.

At EMBL since 2000. Joint Associate Director since 2012.

Future projects and goals

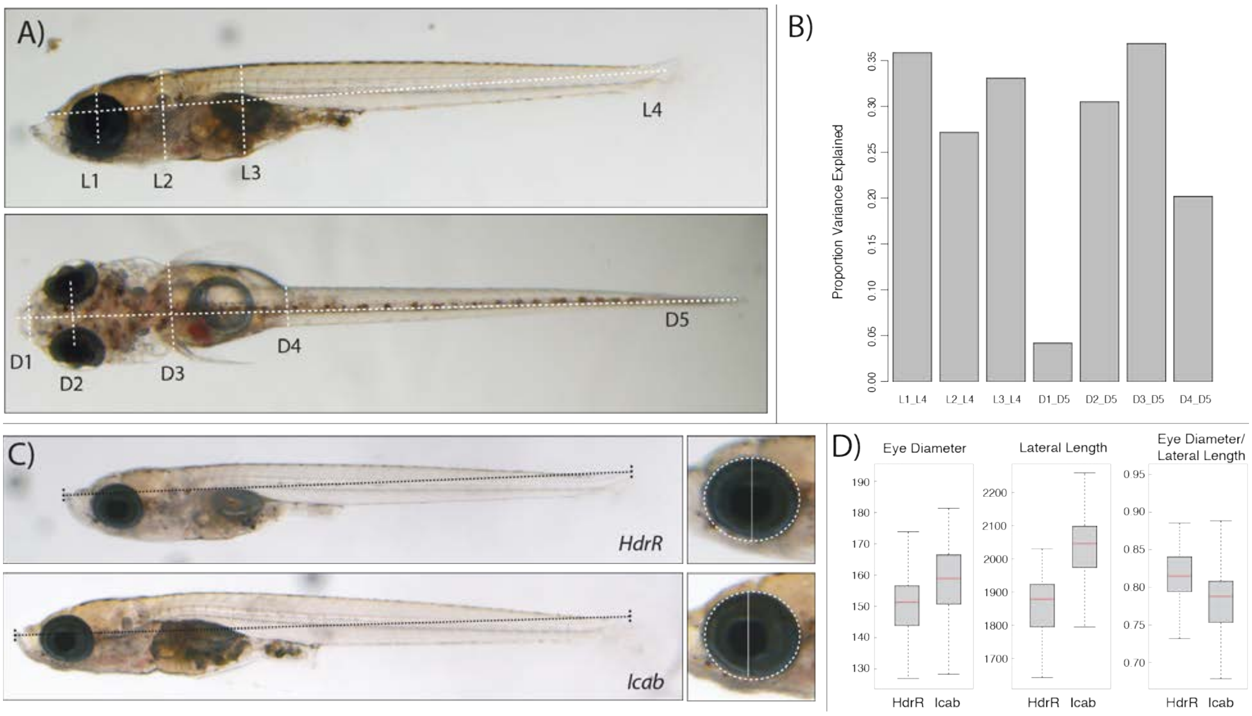
In 2014 the Birney group will continue to work on sequence algorithms and intra-species variation. Our work with human data will focus on molecular phenotypes in an induced pluripotent stem cell (iPSC) panel generated as part of the HipSci consortium, and on a project based on normal human cardiac data. 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.

Selected publications

Goldman, N., Bertone, P., Chen, S., et al. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77-80.

Spivakov, M., Auer, T.O., Peravali, R., (2014) et al. Genomic and phenotypic characterization of a wild medaka population: Towards the establishment of an isogenic population genetic resource in fish. Genes Genomes Gen. 9 January, g3.113.008722

Wright CF, Middleton A, Burton H, et al. (2013) Policy challenges of clinical genome sequencing. BMJ (Clinical Research Ed.) 347, f6845.



Example of differential phenotypes of inbred Medaka fish. (A) Four lateral (L1-L4) and five dorsal (D1-D5) morphometric distances were extracted and analysed for each inbred strain. (B) The proportion of variance explained by the difference between strains as a fraction of the total variance. The variables are the measurements corrected by the appropriate body length measurement (L4 and D5, respectively). (C) Substantial differences in lateral length and eye diameter between fish from the two Southern inbred strains HdrR and Icab. (D) The distribution of eye diameter, lateral length and the ratio between eye diameter and lateral length for HdrR and Icab. Eye diameter and lateral length: y-axis = value in pixel. HdrR (n=70), Icab (n=78).

Enright group

Functional genomics and analysis of small RNA function

Complete genome sequencing projects are generating enormous amounts of data. Although 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.

Summary of progress

- *Developed Kraken, a new set of computational tools for small RNA sequencing data analysis;*
- *Unravelled new epigenetic mechanisms for piwiRNA control of gene silencing in the mouse germline, in collaboration with the O'Carroll lab at EMBL Monterotondo;*
- *Explored the role of the lin-28 gene and let-7 microRNA in malignant germ-cell tumours with collaborators at Addenbrookes hospital;*
- *Developed new tools for statistical analysis of small RNA 5' modification.*

Major achievements

We have developed a number of tools for next-generation sequence data analysis and quality control over the last few years. Mat Davis and Stijn van Dongen have built a new pipeline system based on these tools, which was published in 2013 and is already being used by a number of laboratories. These tools were developed around the needs of small RNA sequencing data analysis and can be applied to mRNA sequencing datasets. In 2013 Matloob Quereshi developed a graphical interface for this system in collaboration with Prof. Tom Freeman's group at the University of Edinburgh.

Our on-going collaboration with Dónal O'Carroll's group at EMBL, Monterotondo focused on two major goals. Firstly, we expanded our initial analyses of piwi-associated RNAs in the mouse germline to look at epigenetic modifications and their effect on piwi-RNA mediated silencing of transposons. We found that the euchromatic repressive histone H3 dimethylated lysine 9 modification cosuppresses L1 expression. We also showed that multiple epigenetic mechanisms, in conjunction with the piRNA pathway, sequentially enforce L1 silencing and genomic stability during mitotic and meiotic stages of adult spermatogenesis.

In 2013 we began a large-scale census of long non-coding RNAs in the mouse germline. This massive task, undertaken by Mat Davis and Harpreet Saini, involves de novo assembly of short-read sequencing data for over 2 billion sequences. We developed computational pipelines and systems for the processing, assembly and classification of a large-number of lncRNAs with significant expression profiles over the course of mouse germline development. Giovanni Bussotti is involved in a large-scale study of lncRNAs in D. melanogaster with the Furlong lab at EMBL Heidelberg.

We are developing novel statistical tools for the prediction and analysis of 3' modification (e.g. Urydilation by TUT enzymes) of microRNAs and additional tools for the prediction of microRNA targets from CLIP assays. This work is being carried out by Tomasso Leonardi, Dimitris Vitsios and Leonor Quintais.

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

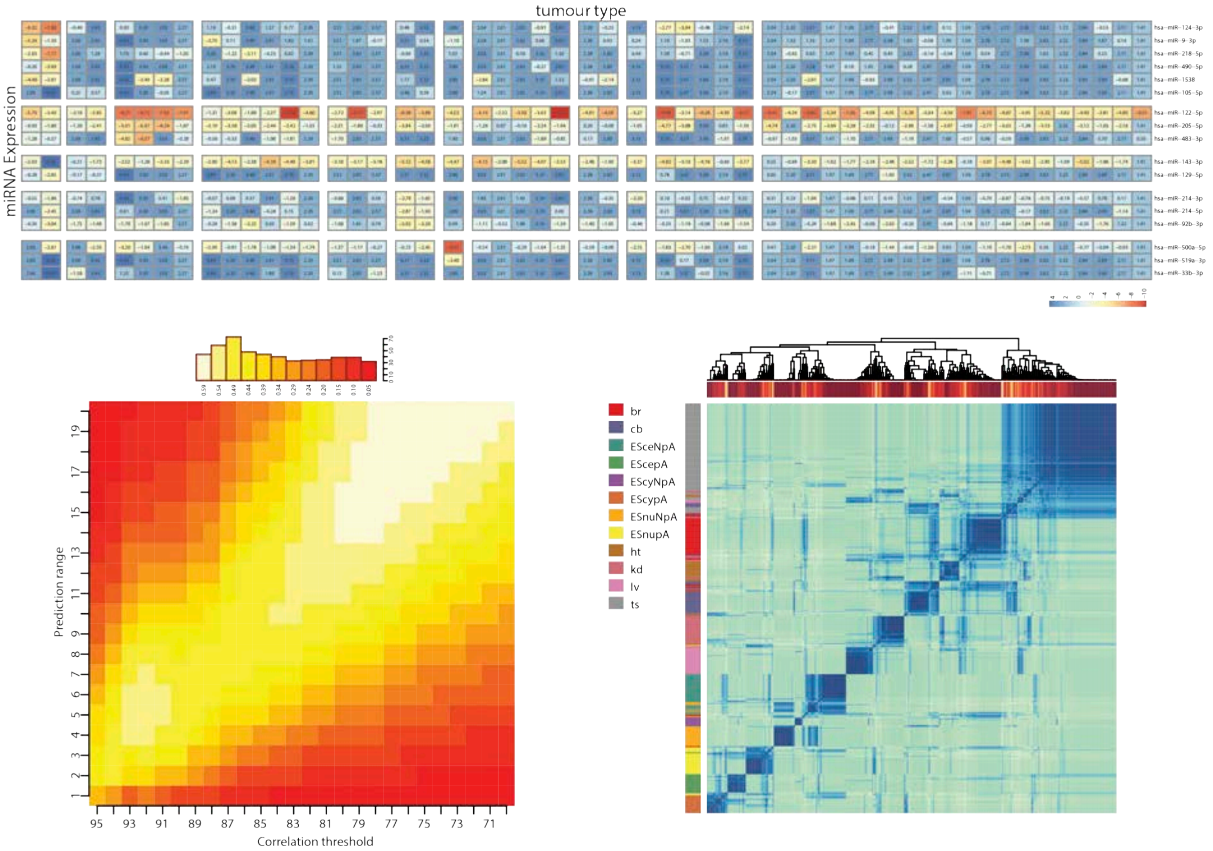
Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. In 2014 we will continue to build an accurate database of piRNA loci in animals and explore the importance and evolution of these molecules. We are extremely interested in the evolution of regulatory RNAs and developing phylogenetic techniques appropriate for short non-coding RNA. We are equally interested in the analysis of lncRNAs and how they fit into the non-coding RNA landscape. 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

Davis, M.P., van Dongen, S., Abreu-Goodger, C., et al. (2013) Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Nat Methods 63, 41-49.

Di Giacomo M, Comazzetto S, Saini H, et al. (2013) Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. Mol Cell 50, 601-608.

Guerra-Assunção, J.A. and Enright A.J. (2012) Large-scale analysis of microRNA evolution. BMC Genomics 13, 218.



Partitioned heat map of differentially expressed microRNAs in select cohorts of solid tumours. Red, high expression in tumour vs. control; blue, low expression in tumour vs. control. Bottom left: Threshold analysis for the creation of gene-expression networks. This plot illustrates for different correlation thresholds the predictiveness of the network topology for two-step relationships. Red, low predictiveness; yellow, high predictiveness (F-measure). This can be used to evaluate a correlation threshold for the creation of gene-expression networks. Bottom right: heatmap showing pair-wise Pearson correlation coefficients of the expression of lncRNAs across tissues. The yellow-to-red bar at the top indicates the tissue specificity score of each lncRNAs, with yellow being 0 and red being 1. The color bar on the left hand side indicates for each lncRNA the tissue of maximal expression.

Goldman group

Evolutionary tools for genomic analysis

Evolution is the historical cause of the diversity of all life. The 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 that data 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, including the analysis of next-generation sequencing (NGS) data. This vast source of new data promises great gains in understanding genomes and brings with it many new challenges.

Major achievements

January 2013 saw the publication of the Group's work using state-of-the-art genomics techniques to encode arbitrary digital information in DNA, in collaboration with Ewan Birney and Agilent Technologies in the United States. As an inherently digital medium, DNA has the capacity to hold vast amounts of information, readily stored for long periods of time in a compact form. Arguably it is ideal for digital information archives, but previously only trivial amounts of information have been stored this way, using techniques that were not amenable to scaling up. We devised and implemented an encoding-decoding ('codec') procedure that can reliably store more information than has been handled before. We successfully encoded five computer files — totalling 739kB and including text, PDF, JPEG and MP3 formats — into a DNA code, synthesised this DNA, sequenced it and reconstructed the original data with 100% accuracy. This represents three orders of magnitude more information than any previous approach. We demonstrated the use of an error-correcting code, and analysed the costs and scalability of our DNA-storage scheme to global-scale information volumes under realistic scenarios for technological advances. Our results indicate that DNA-storage may become a realistic technology for large-scale digital archiving. A patent application has been filed covering some of the encoding techniques devised. Our paper was published in Nature and received extensive international media coverage (TV, radio, film, print and internet media including social networks).

The bioinformatics community has produced numerous data analysis tools incorporating methods to correct sample-specific biases in RNA-seq. However, few advanced simulation tools exist to enable benchmarking of competing correction methods. We released the 'rlsim' RNA-seq library construction simulation software, which incorporates integrated parameter estimation methods, and applied the framework to several publicly available datasets in order to quantify the realism of our simulations and to survey the sample-specificity of biases.

Building upon our expertise in molecular evolution simulation, we undertook a systematic investigation of the robustness of orthology inference algorithms to gene duplication, insertion, deletion, lateral gene transfer and sequencing artifacts. In a related study, we demonstrated that the simple but commonly used 'bidirectional best hits' method misses many orthologs in duplication-rich clades such as plants and animals.

Understanding ongoing evolutionary change is also important on a scale within individual humans. Genomic change within cancer tumours results in intra-tumour heterogeneity (ITH). The expansion of genetically distinct sub-clonal populations may explain the emergence of drug resistance and if so its detection would have prognostic and predictive utility. We developed a novel algorithm — named 'MEDICC' — for phylogenetic reconstruction and ITH quantification based on copy-number profiles. Further work on cancer has seen us collaborate with Peter Campbell's group at the Wellcome Trust Sanger Institute, and we have contributed modules to their established 'ASCAT' algorithm to improve detection of homozygous vs. heterozygous SNPs, improving estimation of germline genotypes using multiple samples from the same patient and automatic inference of trees from sub-clonal allele frequencies.

A long-term interest of the group has been the visualisation of biological data. In 2013 we collaborated with Science Practice, a London-based design studio, to develop a new visualisation technique for multiple sequence alignments (MSAs). The 'Sequence Bundles' visualisation overcomes a major drawback of the de facto standard 'Sequence

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.

Logos' tools, as it is capable of retaining the horizontal dependencies between sites. Sequence Bundles are additionally capable of depicting differing evolutionary signals across different subgroups, e.g. monophyletic clades from a phylogenetic tree, and are especially suited for large alignments with several thousand sequences. The Sequence Bundles visualisation won an Honourable Mention at the BioVis 2013 Awards in Atlanta, United States, October 2013 (http://biovis.net/year/2013/about).

The group continues to study mathematical models of molecular evolution, and is especially concerned with tests of positive selection. Existing tests for positive selection using the ratio of non-synonymous to synonymous substitutions (dN/dS ratio, also denoted ω) tend to be very conservative and lack statistical power. We have developed a new test for positive selection, built upon the group's existing 'SLR' (Sitewise Likelihood Ratio) software, which has greatly improved power whilst retaining control of the false positive rate. We also investigated approaches for integrating structural information into sequence-based methods for detecting both positive and purifying selection and identified a number of viable avenues for further research, including clustering sites under similar selective constraint and introducing site interdependence to increase the power of tests for positive selection

Future plans

The group remains 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 shall 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, where we are 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 remain 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 continue to look to medical applications of NGS and phylogenetics as a source of inspiring collaborations, and hope to start to bring molecular evolution into a clinical setting where it may soon be applicable in 'real time' to help inform doctors' decisions and treatments

Selected publications

de Beer, T.A.P., et al. (2013) Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes project dataset. PLoS Comp Biol 9, e1003382.

Cvejic, A., et al. (2013) SMIM1 underlies the Vel blood group and influences red blood cell traits. Nature Genetics 45, 542–545.

Dalquen, D.A., et al. (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PloS One 8, e56925.

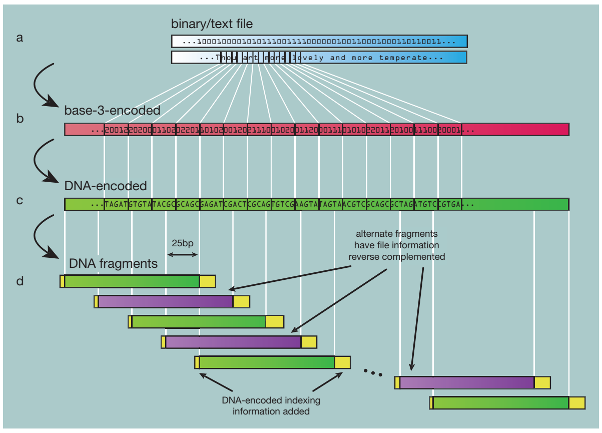
Dalquen, D.A. and Dessimoz, C. (2013) Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol 5, 1800–1806.

Engström, P.G., et al. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nature Methods 10, 1185–1191.

Goldman, N., et al. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77–80.

Schwarz, R.F., et al. (2013) Phylogenetic quantification of intra-tumour heterogeneity. arXiv preprint arXiv:1306.1685.

Sipos, B., et al. (2013) Realistic simulations reveal extensive sample-specificity of RNA-seq biases. arXiv preprint arXiv:1308.3172.



a
binary/text file
...1000100001010111001111000000100110001000110110011...
...Thou art more lovely and more temperate...
b
base-3-encoded
...20012202000110202201101020012021110010200112000111010102201120100111002000l...
c
DNA-encoded
...TAGATGTGTATACGCGCAGCGAGATCGACTCGCAGTGTCGAAGTATAGTAACGTCGCAGCGCTAGATGTCCGTGA...
d
DNA fragments
25bp
alternate fragments have file information reverse complemented
DNA-encoded indexing information added

Digital information encoding in DNA. Digital information (a, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (b, red) using a Huffman code that replaces each byte with five or six base-3 digits (trits). This in turn was converted in silico to our DNA code (c, green) by replacement of each trit with one of the three nucleotides different from the previous one used, ensuring no homopolymers were generated. This formed the basis for a large number of overlapping segments of length 100 bases with overlap of 75 bases, creating fourfold redundancy (d, green and, with alternate segments reverse complemented for added data security, violet). Indexing DNA codes were added (yellow), also encoded as non-repeating DNA nucleotides. Reproduced with permission from Nature Publishing Group.

Marioni group

Computational and evolutionary genomics

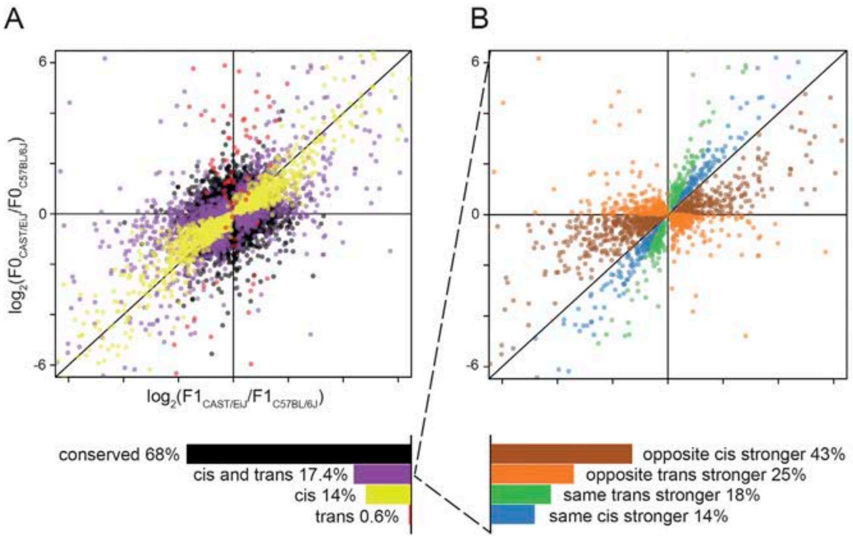
Our research focuses on developing the computational and statistical tools necessary to exploit high-throughput genomics data in order to understand the regulation of gene expression and to model developmental and evolutionary processes.

Within this context, we focus our work on three specific areas. First, we want to understand how the divergence of gene expression levels is regulated. By associating changes in expression with a specific regulatory mechanism, critical insights into speciation and differences in phenotypes between individuals can be obtained. Second, we want to use gene expression as a definition of the molecular fingerprint of individual cells to study the evolution of cell types. By comparing the molecular fingerprint associated with a particular tissue across species, it is possible to decipher whether specific cell types arise de novo during speciation or whether they have a common evolutionary ancestor. Third, we want to model spatial variability in gene-expression levels within a tissue or organism. By modelling such variability, heterogeneous patterns of expression within a cell type can be identified, potentially allowing new cell types, perhaps with novel functions, to be uncovered. Additionally, the extent of heterogeneity present across a tumour can also be studied using such approaches.

These three strands of research are brought together by single-cell sequencing technologies. By studying variability in gene-expression and other genome-wide characteristics at a single-cell level, our ability to assay regulatory variation, molecular fingerprints and spatial patterns of expression will be revolutionised. As founding members of the Sanger-EMBL-EBI Single Cell Genomics Centre, we are closely involved in data generation and in using these data, especially single-cell RNA-sequencing, to answer exciting biological questions. However, to exploit these data to the fullest extent, it is critical to develop the appropriate statistical and computational tools. These are the key challenges we will embrace in the coming years.

Summary of progress

- *Helped co-ordinate development of the Sanger-EMBL-EBI Single Cell Genomics Centre with the Teichmann lab and others;*
- *Developed some of the first methodological tools for handling single-cell RNA-seq data (Brennecke et al., 2013; Kim and Marioni, 2013; Pettit & Marioni, 2013);*
- *Developed new statistical tools for handling high-dimensional data with multiple sources of correlation;*
- *Built new collaborations with the Logan group at the Sanger to study olfaction.*



Classifying genes by their regulatory function. We used RNA-seq data generated from F0 mice and their F1 hybrids to classify genes into sets depending upon their regulatory mechanism. [Goncalves et al., Genome Research 2012.]

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.

Major achievements

In 2013 the Marioni group made significant progress in the area of single-cell transcriptomics, publishing three significant papers. The first exploited single-cell RNA-sequencing to obtain insights into the kinetics of transcription (Kim and Marioni, 2013). The standard kinetic model of transcription assumes that a gene fluctuates randomly between ‘on’ and ‘off’ promoter states, with mRNA being transcribed only in the ‘on’ state. A two-state Markov process governs transitions between the promoter states, with k_{on} and k_{off} describing the rates (per unit time, p.u.t.) at which the gene becomes active and inactive, respectively. When the gene is in the active promoter state, it is assumed to be transcribed at a rate, s (p.u.t), and the number of mRNA molecules of the gene are assumed to decay at a rate, d, (p.u.t.). Transcriptional bursting, which defines the expression profile of the gene under study, is characterised by two parameters: the burst size, s/koff, which describes the average number of synthesized mRNA molecules when a gene is in the active state and the burst frequency, kon, the number of bursts p.u.t. Under this model, a set of differential equations can be derived describing how the number of mRNA molecules of a given gene within a cell changes over time.

We showed that the steady state solution to this system, which describes the probability of observing a given number of molecules at any time point, can be written as a convolution of Poisson and Beta distributions (Kim and Marioni, 2013). Subsequently, using single-cell RNA-sequencing to measure gene expression across a population of cells at the same time point data, we demonstrated that we can efficiently infer the kinetic parameters for each gene by using a hierarchical Bayesian model.

The group made a major contribution to modeling technical noise and identifying highly variable genes from single-cell RNA-sequencing experiments (Brennecke et al., 2013). In collaboration with the group of Marcus Heisler (EMBL Developmental Biology Unit) we showed how extrinsic spike-in molecules could be used to quantify the extent of technical variability in single-cell RNA-sequencing data and how this varied with gene-expression strength. Building upon this, we developed a statistical test to identify genes that showed more variation in expression across cells than would be expected by chance, thus highlighting those genes that might play a physiologically important role in explaining heterogeneity between the population of cells under study.

We developed a visualisation tool to study expression patterns at the single-cell resolution when information about the spatial location of each cell within the tissue under study (Pettit & Marioni, 2013). We are currently building upon this method to cluster cells into groups that can help us characterise cell-type identity.

In parallel to approaches for studying variation in expression at the single-cell level, we continue to work on approaches for studying gene expression levels using bulk RNA-sequencing data. With the Odom lab at Cancer Research UK, we are working on approaches for studying the interface between transcription and translation by assaying mRNA and tRNA gene expression levels during mouse development (Schmitt, Rudolph et al., in preparation). We are also involved in external collaborations with the labs of Francois Spitz at EMBL Heidelberg and George Vassiliou at the Wellcome Trust Sanger Institute that exploit RNA-seq to study how expression changes affect cranio-facial abnormalities and clonality, respectively.

Finally group continues to develop computational approaches based upon matrix variate statistics that can handle transposable data, where the correlation structure both within and between the row and the column variables is of interest (Touloumis et al., in revision).

Future plans

In 2014 and beyond, the Marioni lab will continue to focus on developing computational tools for understanding the regulation of gene expression levels. A strong focus will be on the development of methods for analysing single-cell RNA-sequencing data (in conjunction with other members of the Sanger-EMBL-EBI Single-Cell Genomics Centre), which has the potential to reveal novel insights into cell type identity and tumourigenesis.

Selected publications

Kim, J.K. and Marioni, J.C. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. DOI: 10.1186/gb-2013-14-1-r7.

Brennecke, P., Anders, S., Kim, J.K., et al. (2013) Accounting for technical noise in single-cell RNA-sequencing experiments. Nat. Methods 10, 1093–1095.

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 p¬¬rocess 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 benefit 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.

Our research is hypothesis-driven and tailored towards producing mathematical models that integrate diverse data sources. To this end, we collaborate closely with experimental groups. Our models integrate a range of data, from genomic to biochemical, with various sources of prior knowledge, 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. For 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.

Summary of progress

- *Developed methods within our platform CellNOpt to efficiently model computationally very large scale signalling networks;*
- *Co-organised the eighth edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods);*
- *Developed methods to predict drug response in cancer using genomic and chemical data.*

Major achievements

In 2013 we developed various methods to analyse large drug screenings in cancer cell lines—particularly for the prediction of drug response from genetic and chemical features—in collaboration with the Genomics of Drug Sensitivity in Cancer Group at the Sanger Institute and Massachusetts General Hospital. Specifically, we developed methods that integrate various types of genomic data with our prior knowledge on pathways and gene regulation. These methods point at novel biomarkers of resistance or sensitivity to drug treatments.

We further developed our software tool CellNOpt to model signalling networks. We devoted substantial efforts to developing methods to handle two types of data: single-cell phosphorylation data and mass spectrometry phosphoproteomic data. We applied these methods in various contexts, primarily to dissect the mode of action of drugs on cancer cells.

Our group co-organised the eighth edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods, coordinated by Sage Bionetworks and G. Stolovitzky at IBM), a community effort organised around challenges to advance the inference of mathematical models of cellular networks. In 2013 we were involved in a challenge to predict the toxic effect of chemical compounds: the NIEHS-NCATS-NC DREAM Tox Challenge. We also took part in the Heritage Provider Network (HPN)-DREAM Breast Cancer Network Inference Challenge to identify signal transduction networks in breast cancer from phosphorylation data upon perturbation with drugs and ligands.

Future plans

In 2014 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 phospho-proteomics data with our software tool CellNOpt, and finding ways to employ different proteomics technologies and sources of information about pathways. We will also develop methods to infer drug mode of action and drug repurposing by integrating genomic and transcriptomic data with drug screening data. Using these 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

Julio Saez-Rodriguez

PhD University of Magdeburg, 2007. Postdoctoral work at Harvard Medical School and M.I.T.

At EMBL-EBI since 2010.

Joint appointment at Genome Biology Unit (EMBL-HD).

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?

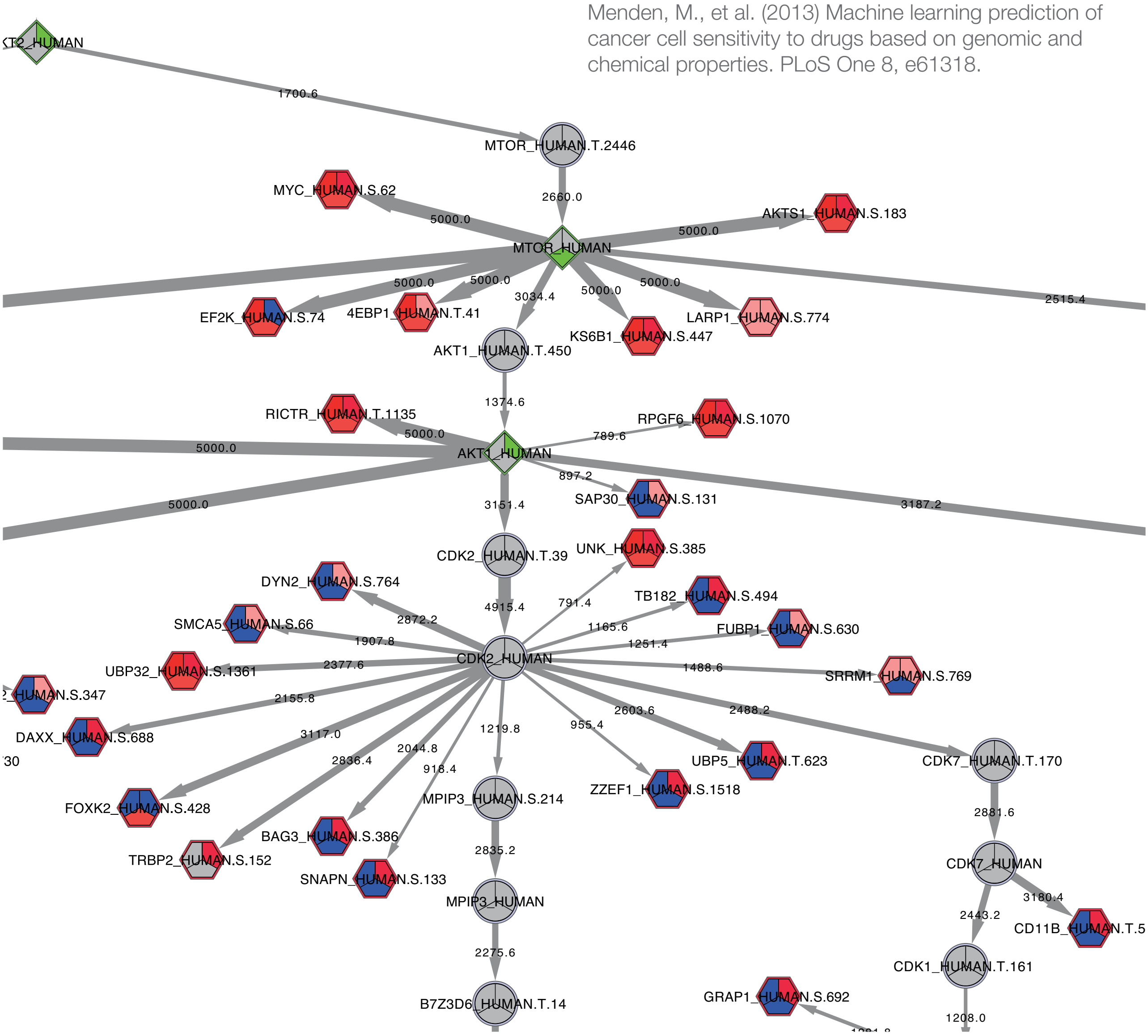
Selected publications

Meyer, P., et al. (2014) Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. BMC Sys Bio, 8, 13.

Chaouiya, C., et al. (2013) SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC Sys Biol. 7, 135.

Guziolowski, C., et al. (2013) Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. Bioinformatics 29, 2320.

Menden, M., et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One 8, e61318.



Subset of a model of a signalling network in a breast cancer cell obtained from phosphoproteomic data.

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 such as transcription and translation?

To answer these questions, we build on statistics as our main tool. 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. With this in mind, we develop statistical analysis methods in the areas of gene regulation, genome wide association studies (GWAS) and causal discovery in molecular systems. Our methodological work is tied in with experimental collaborations, in which we study the variability of molecular phenotypes in different systems including yeast models, plants genomics and human genetics.

Summary of progress

- *Developed and applied new statistical tools to integrate molecular phenotype data and global phenotypes in systems-level models;*
- *Joined the Human Induces pluripotent Stem Cell Initiative (HipSci);*
- *Successfully acquired EMBO funding to run a new EMBO training course on genotype-phenotype mapping starting in summer 2014.*

Major achievements

In 2013 we continued to develop and apply methods for linking genetic variation data and molecular phenotypes. We contributed to major studies mapping the genetic component of gene expression variation in humans, including contributions to an initial publication of the GEUVADIS consortium (Lappalainen et al., 2013).

To fully explain differences in gene expression between individuals, it is necessary to consider genetic effects in combination with other sources of variation. Together with colleagues from Sheffield (Fusi et al., 2013) we devised new statistical approaches to study genetic effects on gene expression in the context of different environmental exposures. The genotype–environment interactions uncovered by our new approach shed light on the interplay of genetic signals and external stimuli.

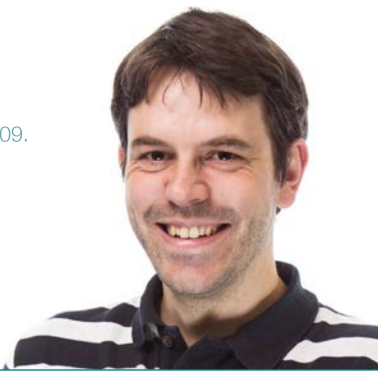
In addition to studying genetic effects on molecular phenotypes in isolation, we developed and applied statistical tools for tying together multiple layers of molecular readouts with global phenotypes. In a collaborative effort with Lars Steinmetz at EMBL Heidelberg and Julien Gagneur at Ludwig Maximilian University Munich, we demonstrated how these methods allow for teasing apart cause–consequence relationships when studying gene expression levels in the context of genotype-phenotype relationships of fitness traits (Gagneur et al., 2013). This work laid the groundwork for other projects in human genetics in which we exploit rich molecular phenotype data to dissect the GWAS loci.

Complementing our contributions in the field of statistical and molecular genetics, the group conducted research in the field of machine learning. In collaboration with colleagues at Max Planck Tübingen, we developed new models that allow for jointly analysing statistical associations of multiple response variables in a single model (Rakitsch et al., 2013). These theoretical developments form the basis of integrative genetic approaches to study high-dimensional phenotypes.

Oliver Stegel

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 2014 we will continue to devise statistical methods to model and analyse data from high-throughput genetic and molecular profiling experiments. Technically, the development of new approaches for tying together quantitative readouts across multiple molecular layers will increase in importance. To this end, we develop causal inference methods to deduce functional relationships from the wealth of correlative omics datasets being generated. We are particularly interested in applying these methods to data from the Human Induced Pluripotent Stem Cell Initiative, in which we are a partner.

Selected publications

Gagneur, J., Stegle, O., Zhu, C., et al. (2013) Genotype-environment interactions reveal pathways that mediate genetic effects on phenotype. PLoS Genet. 9, e1003803.

Gan, X., Stegle, O., Behr, J., et al. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477, 419-423.

Parts, L., Stegle, O., Winn, J., et al. (2011) Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet. 7, e1001276.

Rakitsch B., et al. (2013) It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In Advances in Neural Information processing Systems, 19.

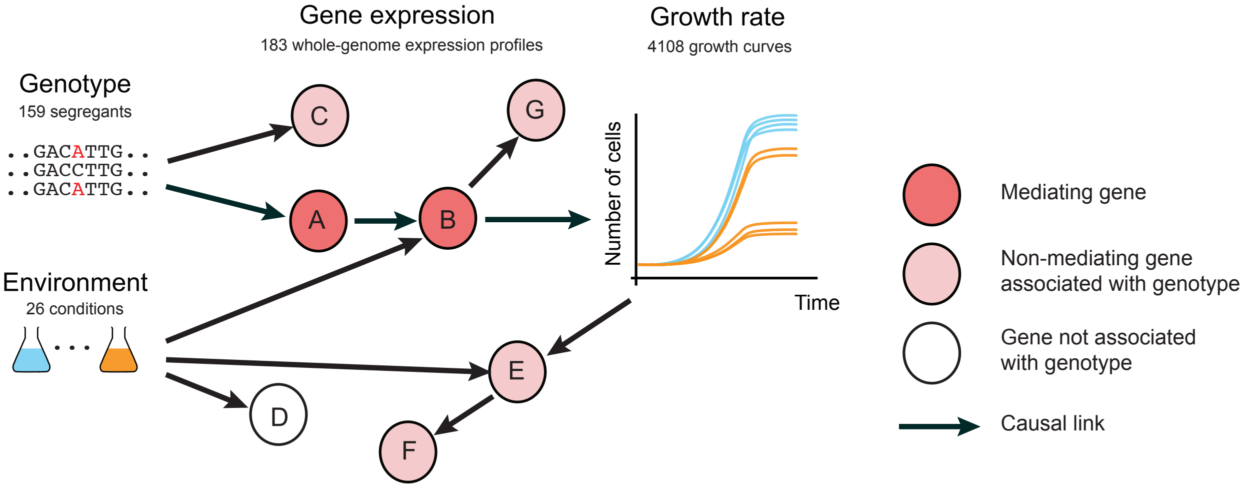


Illustration of causal molecular pathways that link genetic and environmental factors to global phenotypes such as fitness or disease (figure adapted from Gagneur et al., 2013). Statistical approaches allow for distinguishing functional molecular intermediates (mediating genes) from merely correlated markers (non-mediating genes).

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 now 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 graded or switch-like.

Our group also investigates the 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.

Summary of progress

- *Smoothly transitioned the Teichmann group from the MRC Laboratory of Molecular Biology to the Genome Campus;*
- *Co-founded the Sanger-EBI Single Cell Genomics Centre and published the first paper on our single cell RNA-sequencing work in collaboration with the Heisler and Marioni groups (Brennecke et al., 2013);*
- *In collaboration with the Robinson group, showed that heteromeric protein complexes have well-defined assembly pathways that are constrained in evolution (Marsh et al., 2013).*

Major achievements

In 2013 the Teichmann group transitioned from the MRC Laboratory of Molecular Biology eight miles south to the Genome Campus. The move went smoothly and both the wet lab at Sanger and dry lab at EMBL-EBI are now in full swing. We are founding members of the Sanger-EBI Single Cell Genomics Centre, which launched in 2013.

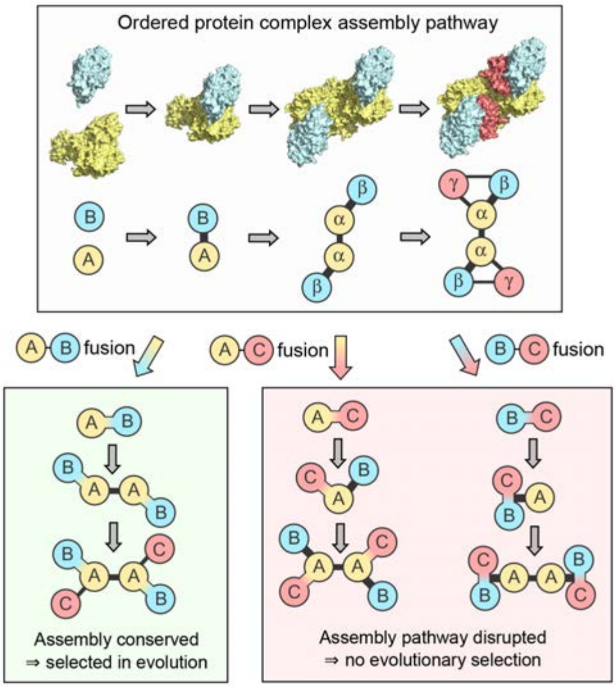
In our work on structural bioinformatics of protein complexes, our major breakthrough work in 2013 (Marsh et al., 2013) showed that proteins encode a single major assembly pathway for each protein complex, and this pathway is under evolutionary selection pressure. These insights open up new possibilities for predicting assembly pathways and designing protein complexes for protein engineering, as well as guiding design of small-molecule inhibitors. This work was carried out in collaboration with world-leading mass spectrometry experts in the Robinson group, who use electrospray mass spectrometry to analyse proteins in the gas phase. Our study centred on protein characteristics that affect disassembly routes in the gas phase (Hall et al., 2013).

We also reviewed work on the conformational dynamics and flexibility of protein complexes, and articulated how this relates to their evolution (Marsh & Teichmann, 2013). In another review, we discussed how graph theoretical methods can be used to describe residue interactions within and between polypeptide chains (Zhang, Perica & Teichmann, 2013).

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 Dept Physics/Cavendish Laboratory, University of Cambridge, 2013-2016.

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


Ordered protein complex assembly pathway
B
A
α
β
γ
A B fusion
A C fusion
B C fusion
C
Assembly conserved
⇒ selected in evolution
Assembly pathway disrupted
⇒ no evolutionary selection

Graphical abstract from Marsh et al. (2013): Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153, 461-470.

Several years ago we observed that mRNAs in metazoan cell types can be divided into two major abundance classes (Hebenstreit et al., 2011), and presented evidence that the lower abundance class corresponds to stochastic, noisy gene expression. This therefore implies a method for thresholding bulk RNA-sequencing data based on a bimodal distribution. In 2013 we applied this principle to study a specific type of fibroblasts involved in immune evasion in mouse models of cancer, in collaboration with the Fearon group. Our analysis showed that this type of fibroblast, purified from different tissues, shares a common lineage (Roberts et al., 2013), and helped identify the mechanism of immune evasion (Feig et al., 2013).

Moving beyond traditional bulk RNA-sequencing, high-throughput single cell RNA-sequencing is now routine at the Sanger-EBI Single Cell Genomics Centre. In order to calibrate the technical noise in highly parallel single cell RNA-sequencing on the Fluidigm C1 robot, we used synthetic RNAs spiked in to our mouse T helper cells. We evaluated the cell-to-cell variation of these synthetic RNAs to dissect the technical from biological variation across individual cells (Brennecke et al., 2013). This method has become our bread-and-butter for single-cell RNA-sequencing as part of many other projects.

Future plans

We will continue our projects in structural bioinformatics of protein complex assembly and expand our programme in genomics of gene expression. A major thrust of the group will be single cell transcriptomics of the dynamics of immune responses to pathogens. This will reveal the full spectrum of CD4+ T cell types, and the evolution of the T cell response during an infection time course. 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 work towards integrating high-throughput high-content imaging with single-cell RNA sequencing.

Methods development in single cell bioinformatics approaches is a further strand of research that we will pursue energetically over the next few years. 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

Brennecke, P., Anders, S., Kim, J.K., et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10, 1093-1095.

Feig, C., Jones, J.O., Kraman, M., et al. (2013) Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. Proc. Nat. Acad. Sci. U.S.A. 110, 20212-20217.

Hall, Z., Hernández, H., Marsh, J.A., et al. (2013) The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure 21, 1325-1337.

Marsh, J.A., Hernández, H., Hall, Z., et al. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153, 461-470.

Roberts, E.W., Deonarine, A., Jones, J.O., et al. (2013) Depletion of stromal cells expressing fibroblast activation protein-   from skeletal muscle and bone marrow results in cachexia and anemia. J. Exp. Med. 210, 1137-1151.

Zhang, X., Perica, T. and Teichmann, S.A. (2013) Evolution of protein structures and interactions from the perspective of residue contact networks. Curr Opin Struct Biol. 23, 954-963.

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 three-dimensional structure and evolution.

We explore how enzymes perform catalysis by gathering relevant data from the literature and developing novel software tools, which 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 in different contexts and for different species. To understand more about the molecular basis of ageing in different organisms, we participate in a strong 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.

Summary of progress

- *Published our analysis of human nsSNP data from the 1000 Genomes Project, developing a pipeline to map the observed mutations onto protein sequences and structures. Demonstrated a radical difference between the distribution of variants observed in the 1000 Genomes data and those in OMIM. Further analysis of cancer mutations in COSMIC is underway for comparison;*
- *Published our database of survival curves, SurvCurv, and used it to perform meta-analyses of the influence of strain, sex and bacterial infection over many independent experiments;*
- *Published LigSearch, a knowledge-based web server to identify likely ligands for a protein target. This was developed to facilitate crystallisation of proteins by adding an appropriate ligand to the crystallisation medium;*
- *Completed 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. An equivalent analysis of the isomerases is underway;*
- *Completed the EC-BLAST software tool, which compares enzyme reactions and produces a hit list of the most similar reactions. The manuscript was accepted for publication.*

Major achievements

We compared the differences between 'natural' and disease-associated amino acid variants using sequence and structural information from the 1000 Genomes Project, the OMIM database and UniProtKB Humsavar. The results highlight the complex interplay of features from the level of the DNA to protein sequence and structure. We found that the codon CpG dinucleotide content plays a large role in determining which amino acids mutate; this in turn affects the mutability of amino acids. We observed a clear difference between non-disease and disease variants where amino acids that are naturally very mutable show the opposite trend in disease-associated data. Our results show evidence for some selection, mainly in that the variants occur slightly more often on the surface of the protein and are much less likely to be annotated as functional residues by chance than expected. However, even the best definition of 'functional', taken from structural data, is limited. Even with these caveats, it is clear that the 1000 Genomes variants eschew functional residues as defined here; a trend that is, surprisingly, even stronger in the OMIM and Humsavar data.

Survival records of longevity experiments are a key component in research on ageing. However, there have been very few cross-study analyses that go beyond comparisons of median lifespans or similar summary information. We used a large set of full survival data from various studies to address questions in ageing that are beyond the scope of individual studies. We characterised survival differences between female and male flies of different genetic Drosophila strains, and showed significant differences between strains. We further analysed the variation in survival of control cohorts recorded under highly similar conditions within different Drosophila strains, and found that overall transgenic constructs of the UAS/GAL4 expression system, which should have no effect (e.g. a

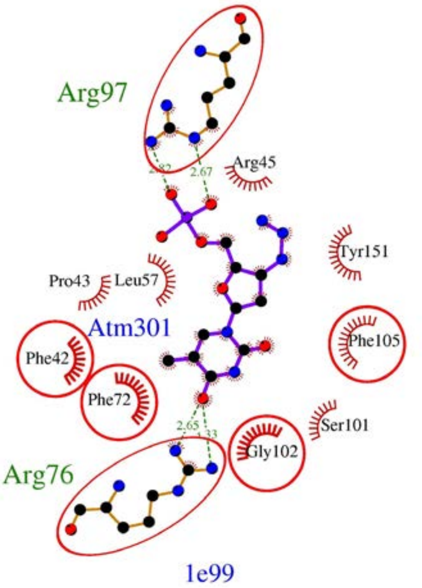
Janet Thornton

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.

GAL4 construct alone), extend lifespan significantly in the w1118 strain. Using a large dataset comprised of various studies, we did not find any evidence for an association between larger lifespan extensions and shorter lifespans of the control in Drosophila. This demonstrates that lifespan-extending treatments are not purely rescuing weak backgrounds.

Identifying which ligands might bind to a protein before crystallisation trials could provide significant savings in time and resources. In 2013 we developed LigPlot, a web server for predicting ligands that might bind to and stabilise a given protein. Using a protein sequence and/or structure, the system searches against a variety of databases, combining available knowledge, and provides a clustered and ranked output of possible ligands (see Figure).



A schematic of the protein-ligand interactions in two distantly related proteins. (a) thymidylate synthase from S. aureus (PDB entry 4eaq), and (b) human thymidylate synthase, (1e99). The ligands (blue bonds) in both are identical: 3'-azido-3'-deoxythymidine-5'-monophosphate. Equivalent protein residues in the two plots are circled in red and occupy the same positions in each plot e.g. Glu37 is equivalent to Phe42, Phe66 is equivalent to Phe72 and Tyr100 is equivalent to Phe 105. Hydrogen bonds are depicted by green dotted lines and labelled with their length in Ångströms, while hydrophobic interactions are represented by the red arcs whose spokes radiate towards the ligand. The diagram was generated using the LigPlot+ program [Laskowski & Swindells, 2011].

Future plans

Our work on understanding enzymes and their mechanisms using structural and chemical information will include a study of how enzymes, their families and pathways have evolved. We will study sequence variation in different individuals, including humans, flies and bacteria, and explore how genetic variations impact on the structure and function of a protein and sometimes cause disease. We will seek a better understanding of reaction space and its impact on pathways, and to use this new knowledge to improve chemistry queries across our databases. Using evolutionary approaches, we hope to improve our prediction of protein function from sequence and structure. We will also improve our analyses of survival curves and combine data with network analysis for flies, worms and mice in order to compare the different pathways and ultimately explore effects related to human variation and age.

Selected publications

Ziehm, M. and Thornton, J.M. (2013) Unlocking the potential of survival data for model organisms through a new database and online analysis platform: SurvCurv. Aging Cell 12, 910-916.

Ziehm, M., Piper, M.D. and Thornton, J.M. (2013) Analysing variation in Drosophila aging across independent experimental studies: a meta-analysis of survival data. Aging Cell 12, 917-922.

de Beer, T.A.P., Laskowski, R.A., Duban, M-E., et al. (2013) LigSearch: a knowledge-based web server to identify likely ligands for a protein target. Acta Crystallographica Section D 69, 2395-2402.

de Beer, T.A.P., Laskowski, R.A., Parks, S.L., et al. (2013) Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes Project dataset. PLoS Comp Biol. 9, e1003382.

Links

LigSearch: www.ebi.ac.uk/thornton-srv/databases/LigSearch

SurvCurv: www.ebi.ac.uk/thornton-srv/databases/SurvCurv

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 comprising of EMBL-EBI faculty, local academics and, where appropriate, industry partners. Our PhD students are awarded their degrees by the University of Cambridge. EMBL-EBI benefited from the presence of 37 PhD students in 2013, welcoming seven newcomers. During 2013, four students successfully defended their theses and were awarded PhDs from the University of Cambridge. These were Filipe Cadete, Andre Faure, Ben Stauch and Matthias Ziehm. Seven students submitted their theses and will defend in 2014.

Andre Faure

Studies of key modulators of higher-order chromatin structure

Mammalian genomes are organised in a way that ensures the coordinated activity of their 20 000 or so genes.

Andre's work in the Flicek group focussed on DNA-associated factors with ties to their 3D genome structure. Using functional genomics approaches, he investigated the roles of cohesin and CTCF in cell-type-specific gene regulation. By integrating genome-wide binding, expression and chromatin conformation datasets he showed that cohesin stabilises large protein–DNA complexes at cis-regulatory regions, as well as intra-chromosomal looping interactions that enable discrete gene expression states. These findings strengthen the link between these proteins and the dynamic control of genome architecture.

Ben Stauch

Methods for the investigation of protein-ligand Complexes

Ben Stauch's work in the Overington group explored two approaches for characterising the binding of small organic molecules to proteins. These interactions play a central role in many biological processes and offer a means to influence these processes in a therapeutic manner. In collaboration with an experimental group at EMBL Hamburg, Ben validated an approach that utilises the noble gas xenon as a probe for detecting ‘druggable’ ligand-binding sites in the early phase of structure-determination projects. Xenon is already used to obtain information required for the structure-determination process. In some instances, xenon-derived druggability information comes ‘for free’ and can be used to streamline the drug-discovery process for novel protein targets during early phases of the structure determination.

In collaboration with colleagues at EMBL Heidelberg, Ben used data obtained by NMR spectroscopy to determine the exact orientation of two ligands binding to a protein kinase. He demonstrated that by treating the protein as a plastic and dynamic system rather than in a widely used, simplified, rigid representation, the methodology could be improved, allowing the determination of true ligand orientations with greater confidence. This shows the potential of NMR spectroscopy-based methods for the study of ligand-binding modes, and provides an alternative to the predominant use of X-ray crystallography in this field.

EMBL International PhD Programme events at EMBL-EBI in 2013

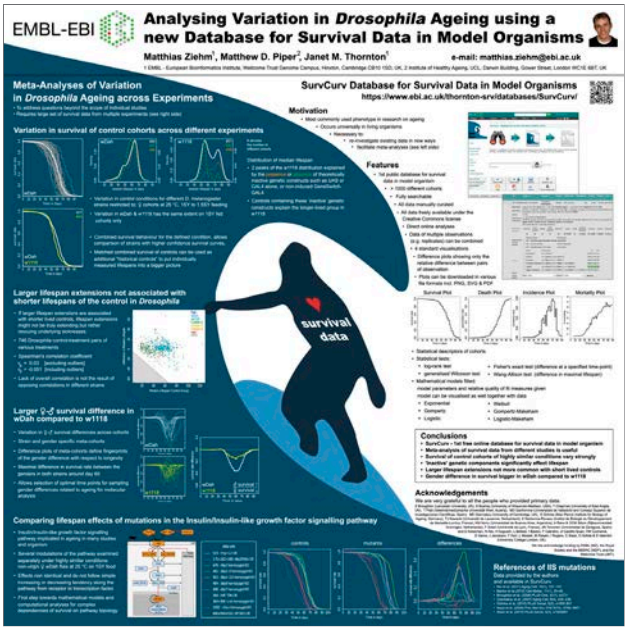
- *Statistics training course jointly held with WT Sanger Institute at EMBL-EBI;*
- *'Primers for Predocs' course at EMBL-EBI;*
- *PhD Student Seminar Day at EMBL-EBI;*
- *Bioinformatics course for all second-year EMBL PhD Programme students: organised and run by EMBL-EBI students;*
- *eSCAMPS: An event featuring talks by leaders in the field as well as by PhD students.*

Matthias Ziehm

Computational biology of longevity in model organisms

Matthias worked in the Thornton group on longevity in model organisms, a key phenotype to establish environmental and genetic factors influencing ageing.

He created SurvCurv, the first database and online resource for survival data from model organisms, and populated it with manually curated data. He used the large dataset he collected to address questions on ageing which are beyond the scope of individual studies, such as hidden factors of variation in survival of control cohorts and correlation of 'effect size' with baseline survival. Matthias also worked towards repurposing compounds into experimental tools for research on ageing in invertebrate model organisms. To do this, he combined a large amount of diverse data types – including implication in ageing, evolutionary relationship, protein structures, compound binding affinities and chemical properties – into a ranking procedure.



EMBL-EBI

Analysing Variation in *Drosophila* Ageing using a new Database for Survival Data in Model Organisms

Matthias Ziehm[1], Matthew D. Piper[2], Janet M. Thornton[1]

e-mail: matthias.ziehm@ebi.ac.uk

1 EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, 2 Institute of Healthy Ageing, UCL, Darwin Building, Gower Street, London WC1E 6BT, UK

Meta-Analyses of Variation in *Drosophila* Ageing across Experiments

- To address questions beyond the scope of individual studies
- Requires large set of survival data from multiple experiments (see right side)

Variation in survival of control cohorts across different experiments

Distribution of median lifespan

- 2 peaks of the w1118 distribution explained by the presence or absence of theoretically inactive genetic constructs such as UAS or GAL4 alone, or non-induced GeneSwitch-GAL4
- Controls containing these 'inactive' genetic constructs explain the longer-lived group in w1118
- Variation in control conditions for different D. melanogaster strains restricted to: ♀ cohorts at 25 °C, 1SY to 1.5SY feeding
- Variation in wDah & w1118 has the same extent on 1SY fed cohorts only
- Combined survival behaviour for the defined condition, allows comparison of strains with higher confidence survival curves.
- Matched combined survival of controls can be used as additional "historical controls" to put individually measured lifespans into a bigger picture

Larger lifespan extensions not associated with shorter lifespans of the control in *Drosophila*

- If larger lifespan extensions are associated with shorter lived controls, lifespan extensions might not be truly extending but rather rescuing underlying sicknesses
- 746 Drosophila control-treatment pairs of various treatments
- Spearman's correlation coefficient
  r_s = 0.03 [excluding outliers]
  r_s = -0.001 [including outliers]
- Lack of overall correlation is not the result of opposing correlations in different strains

Larger ♀-♂ survival difference in wDah compared to w1118

- Variation in ♀-♂ survival differences across cohorts
- Strain and gender specific meta-cohorts
- Difference plots of meta-cohorts define fingerprints of the gender difference with respect to longevity
- Maximal difference in survival rate between the genders in both strains around day 60
- Allows selection of optimal time points for sampling gender differences related to ageing for molecular analysis

Comparing lifespan effects of mutations in the Insulin/Insulin-like growth factor signalling pathway

- Insulin/Insulin-like growth factor signalling pathway implicated in ageing in many studies and organism
- Several modulations of the pathway examined separately under highly similar conditions: non-virgin ♀ wDah flies at 25 °C on 1SY food
- Effects non identical and do not follow simple increasing or decreasing tendency along the pathway from receptor to transcription factor
- First step towards mathematical models and computational analyses for complex dependencies of survival on pathway topology

SurvCurv Database for Survival Data in Model Organisms

https://www.ebi.ac.uk/thornton-srv/databases/SurvCurv/

Motivation

- Most commonly used phenotype in research on ageing
  - Occurs universally in living organisms
    - Necessary to:
      - re-investigate existing data in new ways
      - facilitate meta-analyses (see left side)

Features

- 1st public database for survival data in model organism
  - > 1000 different cohorts
  - Fully searchable
  - All data manually curated
  - All data freely available under the Creative Commons license
  - Direct online analyses
  - Data of multiple observations (e.g. replicates) can be combined
  - 4 standard visualisations
  - Difference plots showing only the relative difference between pairs of observation
  - Plots can be downloaded in various file formats incl. PNG, SVG & PDF
- Statistical descriptors of cohorts
- Statistical tests:
  - log-rank test
  - generalised Wilcoxon test
  - Fisher's exact test (difference at a specified time point)
  - Wang Allison test (difference in maximal lifespan)
- Mathematical models fitted:
  model parameters and relative quality of fit measures given
  model can be visualised as well together with data
  - Exponential
  - Gompertz
  - Logistic
  - Weibull
  - Gompertz-Makeham
  - Logistic-Makeham

Conclusions

- SurvCurv - 1st free online database for survival data in model organism
- Meta-analysis of survival data from different studies is useful
- Survival of control cohorts of highly similar conditions vary strongly
- 'Inactive' genetic components significantly effect lifespan
- Larger lifespan extensions not more common with short lived controls
- Gender difference in survival bigger in wDah compared to w1118

Acknowledgements

We are very grateful to all the people who provided primary data: [illegible]

References of IIS mutations

Data provided by the authors and available in SurvCurv

[illegible]

Matthias's prize-wimming poster shows the newly established online analysis resource and database 'SurvCurv', containing a large set of manually curated survival data in model organisms. These data are key to research in ageing, but had not been amenable to large scale analyses before. Using SurvCurv, he examined a series of questions that are beyond the scope of individual studies, including the variation of lifespans among wild type control cohorts.

Support highlights

Training our users

More than 140 members of EMBL-EBI staff took part in 250 training events, reaching over 8800 people in 28 countries.

Our Train online platform was used by 82 600 unique IP addresses – doubling in uptake over 2012 levels.

The EMBL-EBI Training Programme supports bioinformatics traners throughout the world, and in 2013 worked with the Australian Bioinformatics Network and the University of Leicester to train 205 researchers over six sites.

Expanding infrastructure

Our Systems and networking team grew EMBL-EBI's infrastructure to handle virtualisation of our database servers, and migrated major components from physical to virtualised infrastructure for improved efficiency in our data centres.

Our administration team organised and delivered the administrative aspect of the United Kingdom's Large Facilities Capital Fund Programme, providing both a building on campus and equipment/ and off-site data facilities until the end of 2019.

Improving interfaces

The new EMBL-EBI website, launched in March, offers our users vastly improved navigation between resources, browsing privacy and a more powerful search.

Many teams at EMBL-EBI worked together to create and deploy a new portal for disseminating Semantic-Web supporting information and documentation.

Our web development team vastly improved the workflow for displaying peer-reviewed publications on the website, and helped devise a sustainable solution for reporting research outcomes throughout EMBL.

Our user experience researchers engaged with a wide range of bioinformatics service users, running a large-scale survey to learn more about how scientists interact with our resources.

Cloud solutions

Our Systems and networking team coordinated Embassy Cloud trials throughout the year, and took part in other cloud-based projects such as the International Cancer Genome Consortium (ICGC) Pan-Cancer initiative and the Tara Oceans consortium.

Lightpath to Finland

Our ELIXIR pilot project with CSC Finland progressed towards providing a lightpath over European Academic networks between CSC and EMBL-EBI. This dedicated connection will enable uninterrupted transfer of large datasets between the two institutes.

Raising our profile

Our External Relations team made EMBL-EBI's visual brand stronger and more accessible to our many ambassadors, creating and sharing a range of new publications, slides, graphics and templates.

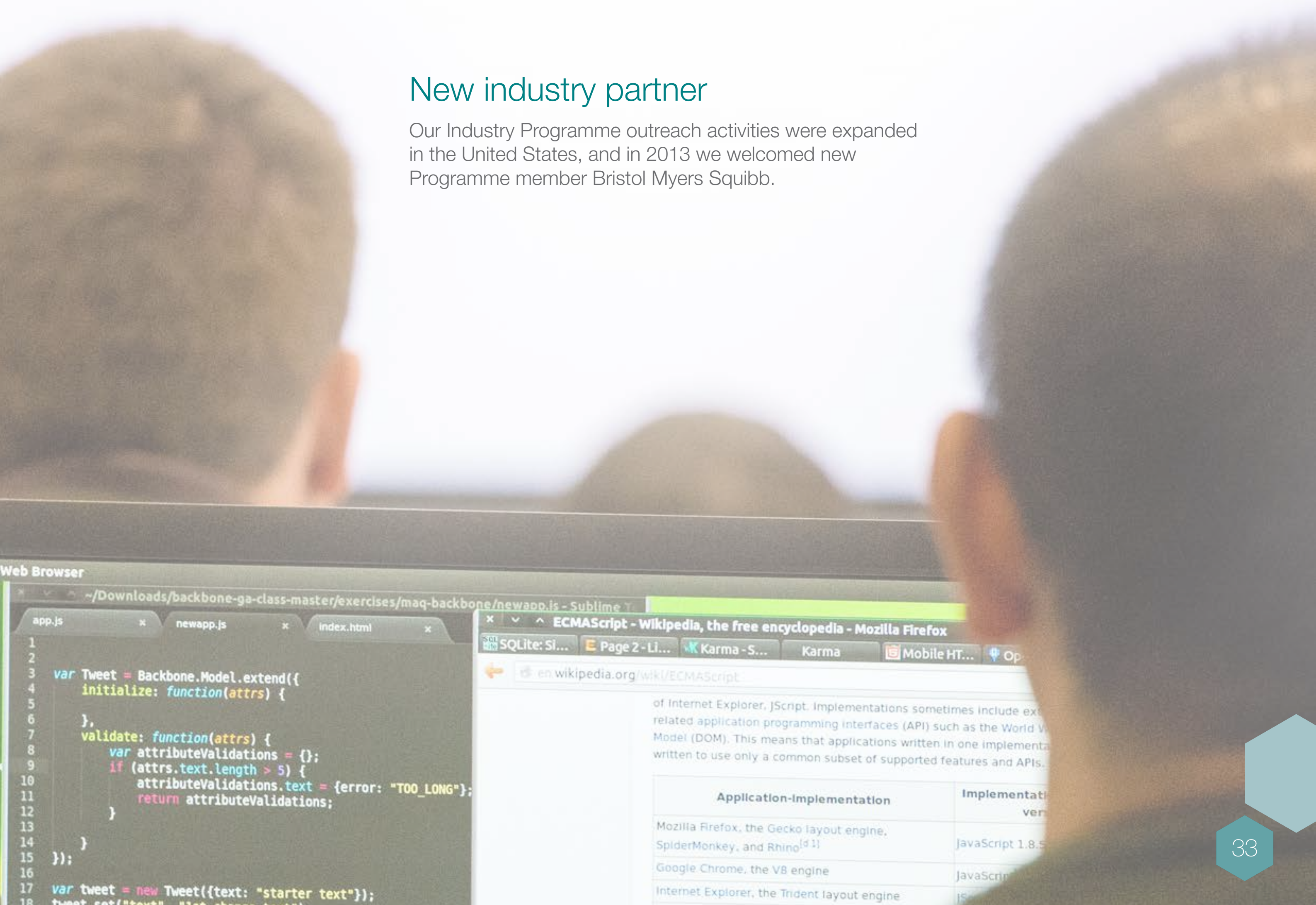
Many teams worked to redesign, organise and implement a brand-new Intranet, which is now easier to update with new logistical and promotional materials.

The Administration, Training and External Relations teams worked to create a practical and beautiful working environment in the new South building, which welcomes a steady stream of visitors throughout the year.

An opening ceremony for the new building, with guests of honour Rt Hon David Willetts MP, Professor Patrick Valance of GSK, BBSRC CEO Jackie Hunter and many long-time supporters from industry and partner organisations, featured an exhibition about the relevance of bioinformatics to society.

New industry partner

Our Industry Programme outreach activities were expanded in the United States, and in 2013 we welcomed new Programme member Bristol Myers Squibb.


Web Browser
~/Downloads/backbone-ga-class-master/exercises/maq-backbone/newapp.js - Sublime T
app.js
newapp.js
index.html
ECMAScript - Wikipedia, the free encyclopedia - Mozilla Firefox
SQLite: Si...
Page 2 - Li...
Karma - S...
Karma
Mobile HT...
Op
en.wikipedia.org/wiki/ECMAScript

```
var Tweet = Backbone.Model.extend({
    initialize: function(attrs) {

    },
    validate: function(attrs) {
        var attributeValidations = {};
        if (attrs.text.length > 5) {
            attributeValidations.text = {error: "TOO_LONG"};
            return attributeValidations;
        }

    }
});

var tweet = new Tweet({text: "starter text"});
```

of Internet Explorer, JScript. Implementations sometimes include ex
related application programming interfaces (API) such as the World W
Model (DOM). This means that applications written in one implementa
written to use only a common subset of supported features and APIs.

Application-implementation	Implementati ver
Mozilla Firefox, the Gecko layout engine, SpiderMonkey, and Rhino[d 1]	JavaScript 1.8.5
Google Chrome, the V8 engine	JavaScri
Internet Explorer, the Trident layout engine	JS

Support summaries

Training

- Actively involved more than 140 members of personnel in 250 events, reaching an audience of >8800 people in 28 countries on 6 continents;
- Supported our colleagues to create new courses in Train online, EMBL-EBI's free, web-based training resource;
- Usage of Train online more than doubled compared with 2012: 82 600 unique IP addresses accessed it and we have >3000 registered users;
- Through our train-the-trainer programme, supported the Australian Bioinformatics Network and the University of Leicester to train 205 researchers over six sites in 2013;
- Expanded our train-the-trainer programme to include metagenomics data analysis and began a collaboration to build training capacity in Brazil;
- Continued to work towards the implementation of LifeTrain, a pan-European framework for continuing professional development for the biomedical sciences;
- Contributed to the expansion of on-course®, the EMTRAIN course portal, by adding hundreds of short courses;
- Supported five new EMBL-EBI ambassadors to raise awareness of EMBL-EBI to the scientific community;
- Secured external funds to develop online training to support discovery in the pharma, agri-food and consumer-goods industries;
- Helped formalise best working practice for training throughout EMBL, and to develop a coordinated approach to course planning throughout the organisation that emphasises collaboration between sites;

External Relations

- Prepared and coordinated the production of all design elements for the new South building;
- Organised the opening ceremony and exhibition for the new EMBL-EBI South building;
- Collaborated with the Web Development team on a content strategy for the new EMBL-EBI website and staff trained on the new content management system;
- Seeded all content for the global website, with the exception of Training, for its launch in March 2013;
- Created and organised outreach content, for the new EMBL-EBI Intranet;
- Issued 20 press releases and coordinated substantial media work for the breakthrough study on using DNA to store digital information;
- Produced and distributed printed the Annual Scientific Report and a range of brochures;
- Developed templates for printed and digital promotional materials and shared them with staff and our colleagues in Heidelberg;
- Created ELIXIR brand guidelines and developed a strategy for their rollout to ensure brand consistency and quality can be easily sustained;
- Contributed to the public affairs activities of ELIXIR to secure new signatories to the Memorandum of Understanding, laying the groundwork for the arrival of the new ELIXIR Director in May 2013.

Industry programme

- Welcomed new Programme member Bristol Myers Squibb in April 2013;
- Expanded outreach activities in the United States, with our second workshop hosted by Pfizer;
- Organised four regular quarterly strategy meetings in Hinxton;
- Organised nine workshops on topics of high importance to industrial partners;
- Initiated precompetitive project discussions in areas relating to biomarkers, informatics for rare disease research and repurposing, an integrated cell-line repository and resources for causal interaction modelling.

Web production

- Deployed a new EMBL-EBI website;
- Improved user privacy in browsing EMBL-EBI web services;
- Handled a sustained increase in the usage of EMBL-EBI sequence analysis services;
- Integrated the global EBI search and web services within several new data resources;
- Improved usability and response for the global search engine.

Web development

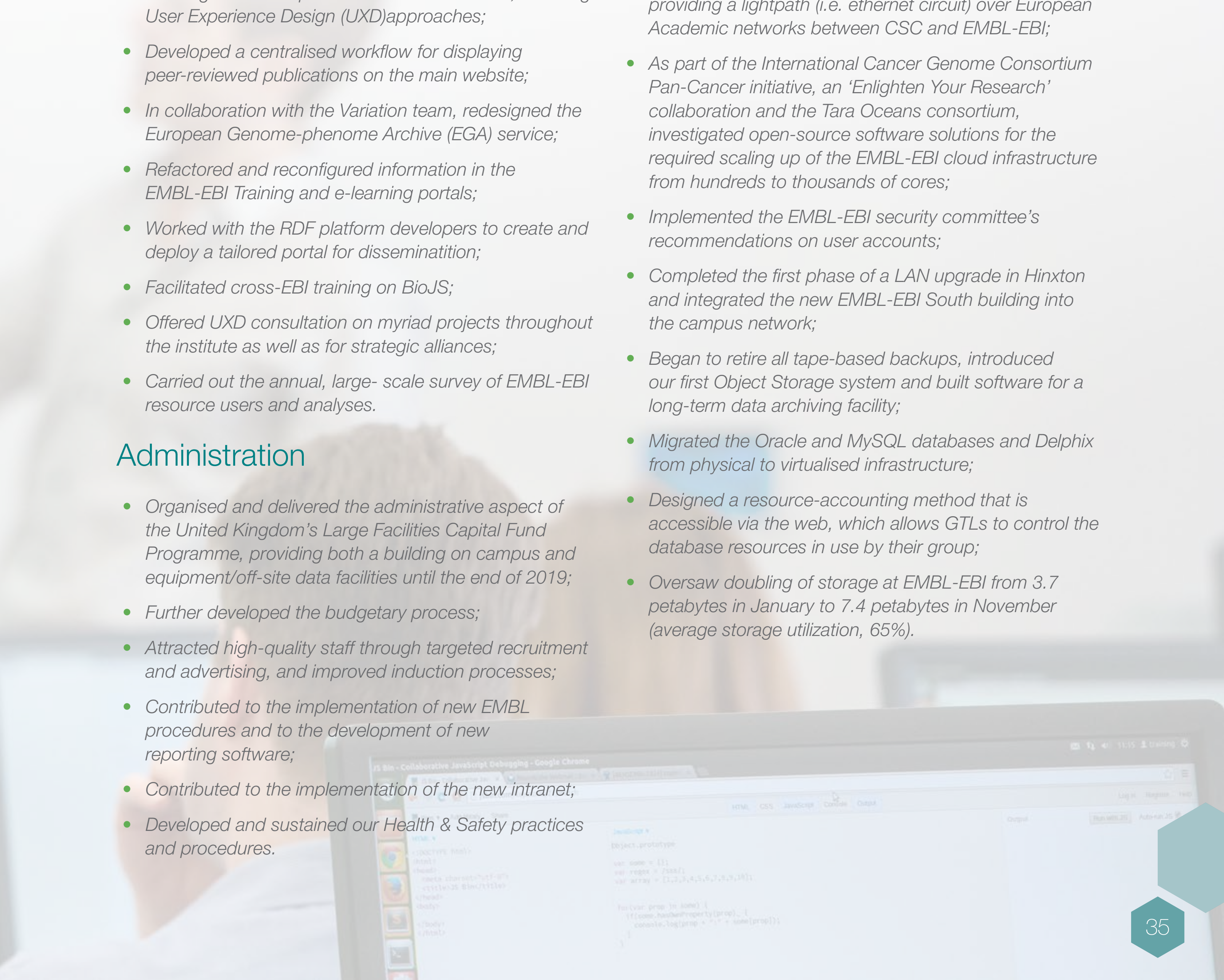
- Launched a new EMBL-EBI website in March;
- Worked with several teams to further develop resources and approaches that encourage users to explore between resources;
- Deployed new templates that web developers can use to generate design-compliant webpages using their own chosen frameworks and technologies;
- Completed the new website's infrastructure and technology shift from largely static HTML to Drupal, deployed on virtual machines for core content;
- Created a new information architecture, then imported and reformatted existing content into that structure;
- Redesigned and implemented the EBI Intranet, following User Experience Design (UXD)approaches;
- Developed a centralised workflow for displaying peer-reviewed publications on the main website;
- In collaboration with the Variation team, redesigned the European Genome-phenome Archive (EGA) service;
- Refactored and reconfigured information in the EMBL-EBI Training and e-learning portals;
- Worked with the RDF platform developers to create and deploy a tailored portal for disseminatition;
- Facilitated cross-EBI training on BioJS;
- Offered UXD consultation on myriad projects throughout the institute as well as for strategic alliances;
- Carried out the annual, large- scale survey of EMBL-EBI resource users and analyses.

Administration

- Organised and delivered the administrative aspect of the United Kingdom's Large Facilities Capital Fund Programme, providing both a building on campus and equipment/off-site data facilities until the end of 2019;
- Further developed the budgetary process;
- Attracted high-quality staff through targeted recruitment and advertising, and improved induction processes;
- Contributed to the implementation of new EMBL procedures and to the development of new reporting software;
- Contributed to the implementation of the new intranet;
- Developed and sustained our Health & Safety practices and procedures.

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;
- 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 group;
- Oversaw doubling of storage at EMBL-EBI from 3.7 petabytes in January to 7.4 petabytes in November (average storage utilization, 65%).



Training

It is essential 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. To that end, EMBL-EBI provides an extensive user training programme, coordinated and funded centrally but with input from technical experts throughout the institute.

Input from technical experts sets EMBL-EBI's courses apart. Our training activities offer a unique interface between service developers and users, making them invaluable in the evolution of existing resources and the creation of new ones. EMBL-EBI's diversifying community of users is reflected in its user-training offering. The programme, courses and materials are created in response to user demand, and cover the full spectrum of EMBL-EBI's activities.

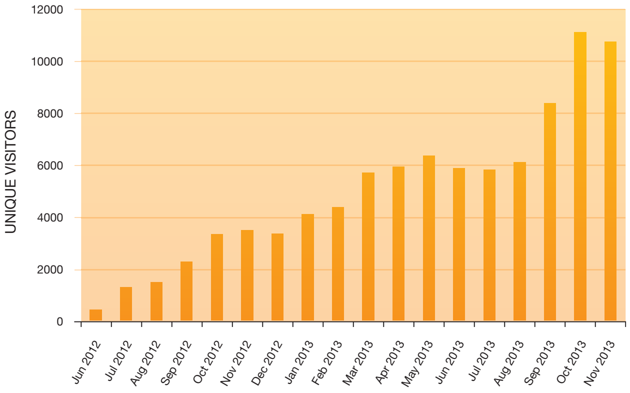
Major achievements

In 2013 we ran a large number of training courses at EMBL-EBI, off-site training events throughout the world, conference exhibitions, careers fairs and workshops. New courses included metabolomics data analysis, metagenomics data analysis and biological interpretation of next-generation sequencing data analysis. We conducted workshops on next-generation sequencing data analysis in India, in collaboration with local trainers at the National Institute of Biomedical Genomics and the National Centre for Biological Sciences in Bangalore. We trained practicing clinicians in the Faroe Islands, agricultural scientists in Slovenia, Spain and Kenya, safety scientists in Italy and the Netherlands, cancer biologists and chemical biologists in Germany, genome biologists in China and graduate students in Zimbabwe. We also delivered our first two knowledge exchange workshops for the BioMedBridges project, bringing together scientists from different research infrastructures to help them build data bridges between their services.

Our programme is made possible by the contributions of trainers 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.

In 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). This has been successfully rolled out in Australia, where we worked with BioPlatforms Australia and CSIRO to facilitate local trainers to train 285 scientists in six states since July 2012. In 2013 we extended our Australian collaboration to include metagenomics data analysis and secured BBSRC funding to build training capacity in Brazil. Closer to home, we supported the University of Leicester to develop its own next-generation-sequencing data analysis courses for graduate students.

The user base of Train online, EMBL-EBI's web-based training resource, has grown significantly: in 2012 we had 33 000 unique users (based on unique IP addresses, which may represent all users at a single institute); in 2013 we topped 82 000 unique users and over 3000 registered users. Guided by our users, we improved the landing pages of our online courses and continue to develop new content.



Growth of Train online's user base since June 2012. Unique users are based on unique IP addresses, which may represent all users at a single institute.

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. LifeTrain, which is working towards a mutually recognised framework for continuing professional development in the biomedical sciences, is gaining momentum, with over 80 individuals and organisations committing to working towards its implementation. Short courses are the fastest growing section of on-course®, EMTRAIN's comprehensive online course catalogue, which now contains more than 2200 short courses. We also made excellent progress on developing the concept for a community portal for course developers.

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.

We extended our collaborations with the broader bioinformatics training community, contributing to the development of GOBLET, a new professional body for bioinformatics trainers that became a legal entity in 2013. We also provided input to a project organised by the Education Committee of the International Society of Computational Biology to define core competencies for bioinformaticians (Welch et al. 2014), and became a partner in the ELIXIR UK node.

We moved into the new training facilities in the new EMBL-EBI South building, which were built with substantial input from the team, especially in terms of audio-visual support. We also worked closely with the Systems and Networking team to develop, test and implement a new training system based on virtual machines.

Summary of progress

- *Actively involved more than 140 members of personnel in 250 events, reaching an audience of >8800 people in 28 countries on 6 continents;*
- *Supported our colleagues to create new courses in Train online, EMBL-EBI's free, web-based training resource;*
- *Usage of Train online more than doubled compared with 2012: 82 600 unique IP addresses accessed it and we have >3000 registered users;*
- *Through our train-the-trainer programme, supported the Australian Bioinformatics Network and the University of Leicester to train 205 researchers over six sites in 2013;*
- *Expanded our train-the-trainer programme to include metagenomics data analysis and began a collaboration to build training capacity in Brazil;*
- *Continued to work towards the implementation of LifeTrain, a pan-European framework for continuing professional development for the biomedical sciences;*
- *Contributed to the expansion of on-course®, the EMTRAIN course portal, by adding hundreds of short courses;*
- *Supported five new EMBL-EBI ambassadors to raise awareness of EMBL-EBI to the scientific community;*
- *Secured external funds to develop online training to support discovery in the pharma, agri-food and consumer-goods industries;*
- *Helped formalise best working practice for training throughout EMBL, and to develop a coordinated approach to course planning throughout the organisation that emphasises collaboration between sites;*
- *Consulted in the development of the new training facilities in the EMBL-EBI South building.*

Future plans

In 2014 we will continue to add new content and functionality to Train online. A new BBSRC-funded modular training partnership in collaboration with Industry will be a major focus of these efforts. 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. We will work towards the implementation of LifeTrain's agreed principles for course providers. This will lay the foundation for a more structured approach to trainer development, within EMBL-EBI and beyond it through our broader interactions with the bioinformatics training community.

Selected publications

Brooksbank, C., Bergman, M.T., Apweiler, R., et al. (2014) The European Bioinformatics Institute's data resources 2014. Nucleic Acids Res. 42, d18-25.

Watson-Haigh, N.S., Shang, C.A., Haimel, M., et al. (2013) Next-generation sequencing: a challenge to meet the increasing demand for training workshops in Australia. Briefings Bioinform. 14, 563-574.

Hardman, M., Brooksbank, C., Johnson, C., et al. (2013) LifeTrain: towards a European framework for continuing professional development in biomedical sciences. Nat. Rev. Drug Discov. 12, 407-408.

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

Industry programme

Since 1996 the Industry Programme has been an integral part of EMBL-EBI, providing on-going and regular contact with key stakeholder groups. The programme is well established as a subscription-funded service for larger companies.

Since 1996 the Industry Programme has been an integral part of EMBL-EBI, providing on-going and regular contact with key stakeholder groups. The programme is well established as a subscription-funded service for larger companies in the pharmaceutical and agri-food industries who wish to leverage more effectively and influence the future direction of EMBL-EBI's genomics, computational biology and chemical biology services and resources.

We support and encourage precompetitive projects by hosting regular strategy meetings and knowledge-exchange workshops covering a broad range of topics. Outputs from pre-competitive projects are made publicly available, benefiting interested parties in EMBL member states and beyond. Our programme also serves as an interface between EMBL-EBI and many external, industry-focussed initiatives and organisations including Innovative Medicines Initiative (IMI), the Pistoia Alliance, the Clinical Data Interchange Standards Consortium and many others. The programme also encourages the involvement of industry in ELIXIR, the pan-European infrastructure for biological information.

Summary of progress

- *Organised four regular quarterly strategy meetings in Hinxton;*
- *Organised nine workshops on topics of high importance to industrial partners;*
- *Expanded outreach activities in the United States, with our second workshop hosted by Pfizer;*
- *Initiated precompetitive project discussions in areas relating to biomarkers, informatics for rare disease research and repurposing, an integrated cell-line repository and resources for causal interaction modelling;*
- *Welcomed new Programme member Bristol Myers Squibb in April 2013.*

Major achievements

During 2013 the programme initiated discussions around many pre-competitive projects that emerged as a result of our knowledge-exchange workshops, which provide members with opportunities to identify and document shared needs they consider to be pre-competitive, and from discussions at the quarterly meetings.

Our quarterly strategy meetings were well attended, and 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). We ran an industry training workshop on RNA-seq data analysis. Reaching out to member companies whose discovery activities are primarily in the United States, we organised our second United States workshop, which covered Oncogenomics and was hosted by Pfizer in Pearl River, NY. Our workshop on data integration marked the formal launch of the EMBL-EBI RDF Platform, which itself was initiated following the Semantic Web for Industry workshop in 2011.

The importance of ELIXIR as a key European research infrastructure cannot be overstated. Its realisation promises to vastly improve the translation of research discoveries into practical applications in medicine and agriculture. Because of the importance of industrial involvement, we have been working closely with companies to secure their participation in the development of ELIXIR.

We reached out to our established network of pharmaceutical industry contacts in Japan, including programme member Astellas Pharma Inc., by organising a set of visits and presentations in and around Tokyo.

Future projects and goals

We see our interactions with industry partners growing even stronger as the need to reduce costs and avoid duplication intensifies. We anticipate that pre-competitive service collaborations, open-source software and standards development will all become basic requirements for life-science companies. During 2014, EMBL-EBI will continue its participation in nine different IMI projects and will seek out more opportunities knowledge-exchange workshops in North America and Asia.

The opening in 2013 of the new South building, which will house the ELIXIR Technical Hub and the Innovation and

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).

Translation suite, presents opportunities for working with larger companies in new ways. We will organise our annual event for small and medium-sized enterprise, the SME Bioinformatics Forum, jointly with the regional SME cluster One Nucleus. We will continue to plan this annual event in collaboration with regional industry organisations throughout Europe, and with the support of the Council of European Bioregions (CEBR).

Industry Programme members 2013

- *Astellas Pharma Inc.*
- *Bayer Pharma AG*
- *Bristol-Myers Squibb*
- *F. Hoffmann-La Roche*
- *Johnson & Johnson Pharma.*
- *Nestlé Research Centre*
- *Novo Nordisk*
- *Sanofi-Aventis R&D*
- *UCB*
- *AstraZeneca*
- *Boehringer Ingelheim*
- *Eli Lilly and Company*
- *GlaxoSmithKline*
- *R&DMerck Serono S.A.*
- *Novartis Pharma AG*
- *Pfizer Inc*
- *Syngenta*
- *Unilever*

Innovative Medicines Initiative projects

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.

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.

OpenPhacts: the Open Pharmacological Concepts Triple Store – Reducing barriers to drug discovery in industry, academia and for small businesses.

STEMBANCC – Providing well-characterised, patient-derived, induced-pluripotent stem cell lines and associated biomaterials in an accessible, sustainable bio-bank.

Industry workshops

- *Oncogenomics*
- *Biomedical ontologies*
- *Biomarkers*
- *Encode and epigenomics*
- *Data integration*
- *Translational informatics*
- *Oncogenomics (held at Pfizer, Pearl River, NY)*
- *Computational tools for chemical biology, phenotypic screening & target de-convolution*
- *RNA-seq data analysis*

Selected publication

Rebholz-Schuhmann, D., Grabmüller, C., Kavaliauskas, S., et al. (2013) A case study: semantic integration of gene-disease associations for type 2 diabetes mellitus from literature and biomedical data resources. Drug Discovery Today. DOI: 10.1016/j.drudis.2013.10.024.

External Relations

As the role of bioinformatics and its application in improving health and economic benefit become more prominent, so the task of engaging EMBL-EBI's multiple stakeholders takes on greater significance. The External Relations team handles public affairs and communications for the institute as a whole, engaging with diverse audiences through a wide range of media.

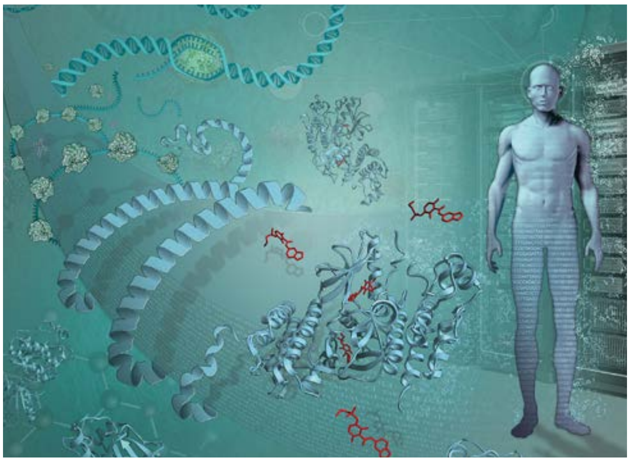
We support the work of EMBL-EBI's many ambassadors, in particular the institute's Director, Associate Directors and team leaders, in fostering good relations with policymakers, funders, potential collaborators and service users throughout the world. We welcome visiting delegations of scientists, politicians 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, 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.

Summary of progress

- *Prepared and coordinated the production of all design elements for the new South building, which opened in October 2013;*
- *Organised an opening ceremony event and poster exhibition for the new building, with guests of honour Rt Hon David Willetts MP and Professor Patrick Vallance, President of Pharmaceuticals R&D at GSK, as well as VIPs from industry and partner organisations;*
- *Collaborated with the Web Development team on a content strategy for the new EMBL-EBI website and trained staff in the use of new content management system;*
- *Seeded all web content including graphics and text for the launch of the new website in March 2013;*
- *Refreshed guidelines, graphics and templates for the new EMBL-EBI intranet, also launched in March 2013;*
- *Hosted the EMBL Alumni Board meeting and two reconnection events for UK residant alumni;*
- *Coordinated inbound international visiting delegations of government and industry representatives;*
- *Produced and distributed printed publications and other promotional materials;*
- *Issued 20 press releases and coordinated substantial media work for the breakthrough study on using DNA to store digital information (528 news stories in top news publications, 5 visiting camera crews);*
- *Developed and shared new templates for printed publications and slide presentations;*
- *Created ELIXIR brand guidelines and developed a strategy for their rollout to ensure brand consistency and quality can be easily sustained;*
- *Contributed to the public affairs activities of ELIXIR to secure new signatories to the Memorandum of Understanding, laying the groundwork for the arrival of the new ELIXIR Director in May 2013.*

Major achievements

The official opening of the new EMBL-EBI South building in October 2013 provided an opportunity to raise the profile of the institute and emphasise our commitment to enabling collaborations with industry. Our team's efforts to strengthen the visual identity of EMBL-EBI in print and digital media extended to creating vibrant graphical elements throughout the South building, and an exhibition showcasing the relevance of the work undertaken at the institute. This included interactive displays designed specifically to engage the Rt Hon David Willetts MP, UK Minister of State for Universities and Science, with the excitement of cutting-edge, life-science research.



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.

EMBL-EBI's new website was launched in March 2013, and in the lead-up we collaborated with the Web Development team and others to establish its look and feel, general guidelines and content strategy. We seeded content and helped train individuals throughout the institute on using the new content management system. We also populated the new Intranet with content, taking the opportunity to refresh our guidelines and templates.

In 2013 we issued press releases regularly (20 press releases, 1426 news stories in news publications (online-only metric), 5 visiting film/TV crews) and coordinated substantial media work for the breakthrough study on using DNA to store digital information. This involved liaising with a large number of journalists, bloggers, film and television crews to field questions on behalf of the authors when possible, and show the institute and campus to best advantage. We continued to expand our social media presence (7500 followers on Twitter, 3000 likes on Facebook) and to plan campaigns for 2014.

Our team produces a range of print-based materials including this Annual Scientific Report, the EMBL-EBI Overview, ELIXIR brochures and others, and contributes to EMBL-wide publications. We also produce high-level slide presentations and other materials for the Director and Associate Directors and others. In 2013 we undertook to establish a process for accessing, reusing and sharing the considerable volume of materials generated in these efforts, and plan to implement an institute-wide system in 2014.

In 2013 we focused on generating templates and sharing them EMBL-wide. These included templates for publications, posters, documents, EMBL-EBI slide presentations and slides/key messages for new initiatives. By distributing high-quality templates with seeded basic content we enabled ambassadors throughout EMBL-EBI to more easily create high-quality materials with which to represent the institute.

We are the first stop for any project seeking advice on branding and visual identity. In 2013 we created clear, simple brand guidelines for ELIXIR and developed a strategy for their rollout to ensure brand consistency and quality can be easily sustained. We began to build an asset library (i.e. graphics, slides) to share with the Nodes. We also worked with several service teams and projects on their logos, elevator pitches and key messages.

As ELIXIR transitioned towards the end of the final year of its preparatory phase, our team helped lay the groundwork for the formalisation of ELIXIR as a legal entity by continuing to raise its profile amongst European scientific and funding stakeholders. With our colleagues in Heidelberg, we supported the process through political engagement with member states and promotion of the brand guidelines at the earliest possible stage. The careful management and stewarding of ELIXIR's brand is crucial as countries with vastly different communications resources begin to promote the new infrastructure at their respective Nodes.

Our alumni are ambassadors for EMBL throughout the world and we help them reconnect with each other and the institute. With the support of the EMBL Alumni Programme, we continued our UK-based alumni activities with an evening event in Cambridge for 28 former staff. We hosted the EMBL Alumni Board semi-annual meeting in the splendid surroundings of the new South building.

Future plans

In 2014 EMBL-EBI will observe the 20th anniversary of its founding. We will lead the organisation of a celebratory event, including an exhibition, booklet and film, for staff and alumni in June. This occasion will provide an opportunity to reflect on the rapid rise of bioinformatics and look to its future, honouring the people who have driven its progress.

We will promote the new The Centre for Therapeutic Target Validation, a collaboration with GSK and the Wellcome Trust Sanger Institute, to a number of audiences. As this first-of-its-kind partnership will be based at EMBL-EBI, we will be initially taking the lead on communications. In 2014 we will change to more sophisticated tools for contacting the media, analysing coverage and sharing reports with our colleagues.

CTTV
Centre for Therapeutic Target Validation

In 2014 we will support the institute's communications activities by establishing: an intranet-based, user-friendly image and slide repository and other tools; a firm strategy for prioritising reusable content; and a regular meeting for website content and social media authors throughout EMBL-EBI. To maximise uptake and reduce waste, we will focus more on our digital, mobile-friendly content.

Our team will continue to support ELIXIR in its outreach strategy, building a network of communications professionals representing the Nodes and helping them integrate the ELIXIR brand into their respective websites.

In 2014 we will devote considerable efforts to forming a society of philanthropic giving that will ultimately provide income streams for EMBL-EBI. We will continue to participate in the campus Sex in Science programme, which generates tackles issues traditionally facing women in science and drives policy and practice changes to redress them.

Publication

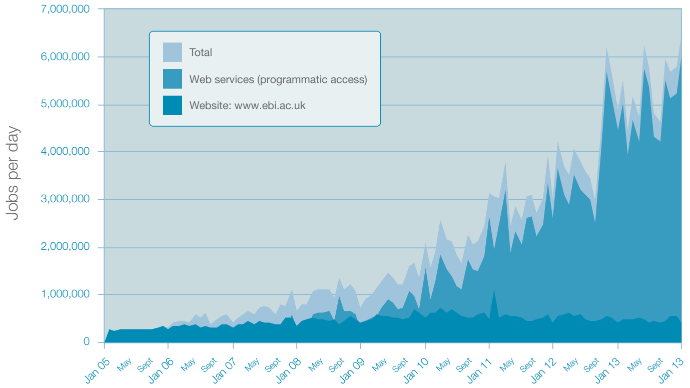
Brooksbank, C., Bergman, M.T., Apweiler, R., et al. (2014) The European Bioinformatics Institute's data resources 2014. Nucleic Acids Res. 42, d18-d25.

Web production

Our team manages the EMBL-EBI web infrastructure, delivers platforms for web service development and provides robust, secure frameworks for deploying public bioinformatics services. We develop and maintain the global EBI search engine, the job-dispatcher framework and corresponding SOAP/REST web services for programmatic access.

We develop and maintain the global EBI search engine, the job-dispatcher framework and corresponding SOAP/REST web services for programmatic access. We also facilitate access to project management core services including document management (Alfresco, 10 sites), project documentation and tracking (Confluence, 1319 users on 100 project spaces; JIRA, 613 users over 65 projects), source code control (CVS, SVN and GIT) and user support.

Jobs per day, 2005 through 2013



Summary of progress

- *Deployed a new EMBL-EBI website;*
- *Improved user privacy in browsing EMBL-EBI web services;*
- *Handled a sustained increase in the usage of EMBL-EBI sequence analysis services;*
- *Integrated the global EBI search and web services within several new data resources;*
- *Improved usability and response for the global search engine.*

Major achievements

During 2013 the team focused on developing and deploying the new global EMBL-EBI website, which is now running in production infrastructure from the London data centres. All web operations from London are now running on infrastructure based on virtual machines (VMware), put together in collaboration with the Systems and Networking Team. As part of this work we developed a new Intranet, maintained core services and delivered user support (4500 tickets in 2013). Our development efforts led to substantial improvements in data management, services monitoring and web-traffic reporting. Web traffic is monitored daily with the production of more than 450 distinct reports that include traffic over http/https, ftp and Aspera.

Secure, assured access

In 2013 the team completed implementation of secure access for all sites under the ebi.ac.uk domain. We were largely responsible for running three service-readiness “fire drills” on EMBL-EBI’s web services, and concluded that services running from the London Data Centres are resilient to network outages between Hinxton and London and have no dependencies on the web infrastructure in Hinxton. We also showed that data-submission services for certain data resources must move to more robust and modern virtual infrastructures as a matter of priority. Based on these findings we are planning improvements in DNS and proxies maintained in Hinxton and London.

Increased usage of services

During 2013 more than 65 million jobs were run using the jdispatcher framework (McWilliams et al., 2013). The most popular services were InterProScan, Clustal Omega and NCBI Blast+. New additions include the HMMer suite for TreeFam and Pfam. Projects that consume these tools within their web sites include uniprot.org, ensemblgenomes.org, PDBe and InterPro. External users of these include Blast2GO, Galaxy, .Net Bio, BlastStation, BioServies, KEIO Bioinformatics Web Services, Yabi, BlastStation, STRAP, T-Coffee, CCP4, Geneious and GMU-metagenomics.

The average number of jobs per month was 5.4 million (up from 4 million in 2013). Figure 1 shows the usage of this system since 2005. The number of datasets available

Rodrigo Lopez

Veterinary Medicine Degree, Oslo Weterinærhøgskole, 1984. MSc in Molecular Toxicology and Informatics, University of Oslo, 1987.

At EMBL-EBI since 1995.

for searching was over 60K in 2013. This is based on the number of class- or taxonomy-specific files being produced to generate Blast and InterProScan services.

Our team informs strategic decisions about which services to continue providing in order to best serve the needs of EMBL-EBI users. In 2013 we were active in an institute-wide assessment of bioinformatics tools (e.g. SOAPLAB, SRS, WSEMBOSS), which will result in decisions about which tools to keep and how best to present them to users.

Search engine integration

We refactored and redesigned many aspects of the global EBI search engine during 2013, for example extending the gene and protein summaries to include bacterial, fungi and plant model organisms. Our search-engine experts have a profound understanding of how the numerous data types used at EMBL-EBI interact with each other, and in 2013 we were able to add new data domains from ENA, BioSamples, the new Expression Atlas, HLA/MHC sequences, MetaboLights and Pombase, to name a few.
To ensure the search engine's high performance we reduced network and resource utilisation by 50%, reduced the web application memory footprint by 40%, reduced the size of indices by 30% and improved cacheing. This removed potential bottlenecks and assured the smooth running of indexing and functioning of the web app in a fully virtualised environment.

Outreach, training and support

In 2013 we handled 2080 help-desk tickets (up from 1350 in 2012). These were primarily requests for help with: programmatic access and data acquisition; technical problems with services; best practices; and training on specific resources. Our team participated in 12 bioinformatics training courses, both in the Training Programme and at external conferences and workshops. We generally provide a comprehensive overview of how best to use EBI resources (including the search), discuss tools and techniques for sequence manipulation and searching (including, for example, analysis of proteomic and immunogenetic sequences); walk people through different aspects of multiple sequence alignment and provide developer training on the use of the Web Service APIs.

Future plans

Changes to the group composition towards the end of 2013 have required a reprioritisation of tasks, and we will be recruiting new staff in 2014. We plan to further improve the interoperability of EBI services and the global search, enhance data logistics to keep up with data growth, maintain operations for EMBL-Australia and deploy a new inventory and web-traffic reporting platform. We also plan to export the jdispatcher framework to partners in Barcelona (i.e. Bioinformatics Barcelona, Centre for Research in Animal Genomics and Universitat Autònoma de Barcelona).

As part of the Technical Services Cluster, led by Steven Newhouse, we will engage in the consolidation of resources at the technical-management level; explore federated authentication for Confluence, Jira and Alfresco; improve communication and outreach for web-services deployment; promote standardisation ; develop and promote the use of well-defined guidelines and teach best practises; run further "fire drills"; and support group and team leaders in aligning their technical approach to meeting the EMBL-EBI service mission. During 2014 we will also further strengthen EMBL-EBI's trusted, reputable services and websites. We will bolster privacy and overall web security on both the development and production fronts, ensuring the implementation of industry-standard best practices, particularly with reference to the web load balancers.

Selected publications

McWilliam H., Li W., Uludag M, et al. (2013) Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41(Suppl), W597-W600.

Robinson, J., Halliwell, J.A., McWilliam, H., et al. (2013) The IMGT/HLA database. Nucleic Acids Res. 41, D1222–D1227.

Robinson, J., Halliwell, J.A., McWilliam, H., et al. (2013) IPD—the Immuno Polymorphism Database. Nucleic Acids Res. 41, D1234–D1240.

Li, W., Kondratowicz B., McWilliam H., et al. (2013) The annotation-enriched non-redundant patent sequence databases. Database (Oxford) 2013, bat005.

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 currently maintain the main website and its content database, the Intranet and Training portals. We also develop and maintain over 30 ancillary web portals and services, including the European Genome-phenome Archive (EGA), ELIXIR, 1000genomes, INSDC, BioMedBridges and HGNC, among many others.

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

Building on the team's success in 2012 of coordinating and developing a global website redesign, we successfully launched a new EMBL-EBI website in March 2013. We collaborated with teams throughout the institute to further develop resources and approaches that both improve service delivery and encourage users to explore between resources. We are working with several services that could not fully comply with the new website guidelines by March 2013, helping them move towards full compliance by March 2015.

We played a key role in developing design guidelines, content style sheets (CSS) and templates that could be used by web developers throughout the institute to generate design-compliant webpages using their own chosen frameworks and technologies. While content providers can now maintain editorial control, the framework ensures a consistent experience for our external users.

In 2013 we completed the new website's infrastructure and technology shift from largely static HTML to Drupal, deployed on virtual machines for core content. We created a new information architecture, then imported and reformatted existing content into that structure. The team created HTML templates as well as a Template Generator Service to make it easy for developers to migrate their services.

Content strategy for the global website relies on a content database that holds key information about individuals, groups and resources, so that any sensitive data can be held securely within the EBI firewall. To fill this need, the team created an internal Drupal portal site. Group and Team Leaders as well as trusted delegates can access this portal so that the information displayed on the main site can more easily be kept up to date. We also completely redesigned and implemented the EBI Intranet in 2012/2013, following similar UXD approaches to the main site.

We developed a centralised workflow for displaying peer-reviewed publications on the main website. The team is involved in discussions and efforts across EMBL to devise a more sustainable solution using technologies such as ORCID and Converis. This project is in conjunction with EMBL Heidelberg and is expected to continue well into 2014.

Projects

In collaboration with the Variation team, we redesigned the European Genome-phenome Archive (EGA) service. This involved upgrading the site from Drupal 6 to Drupal 7 and splitting the current service into two distinct portals, one of which is deployed inside a secure vault. We completely developed both portals and worked to develop a secure communication method to allow both sites to function seamlessly.

As part of the website relaunch, we refactored and reconfigured information in the EMBL-EBI Training portal to be compliant with the new design. We also updated and refitted the Train online e-Learning portal, originally developed by the Web Development and Training teams.

EMBL-EBI's RDF platform aims to bring together the efforts of a number of resources that provide access to their data using Semantic Web technologies. In 2013 our team worked with the platform developers to create and deploy a tailored Drupal portal for disseminating RDF information and documentation.

BioJS, an open-source library of JavaScript components to represent biological data, has great potential to provide standard components for data representation and visualisation. In 2013 we facilitated cross-EBI training on BioJS to evaluate its potential and interest across the institute's developers, and received overwhelmingly positive responses.

Our UXD experts interacted with a most groups throughout 2013, offering consultation on myriad projects and strategic alliances as well as hosting UX-focused workshops, talks, webinars and teleconferences (see list). This part

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.

of the team carries out the annual, large- scale survey of EMBL-EBI resource users and present subsequent analyses to key stakeholders. In 2013 the survey, coordinated by our User Experience experts, included 24 questions and elicited 1232 respondents.

Future plans

Drupal 8 will be released in 2014 and we will implement our strategy for transitioning content in Drupal 6 to Drupal 8, as Drupal 6 will no longer be supported. This process, which particularly impacts Train online, will require considerable time and resource.

We plan to re-architecture some of the main website in order to improve author experience. We will also implement the full set of requirements for the Events platform and deploy a more engaging portal for users to explore EMBL-EBI events.

When the main website launched in 2013, mitigated sites were given a timetable of two years to become fully compliant. Through 2014 we will work with the remaining mitigated site developers to facilitate this upgrade.

BioJS has become more widely adopted within EMBL-EBI services, and greater knowledge of JavaScript is vital in leveraging the technology. Our team will lead the organisation of a JavaScript training symposium at the institute for the benefit of all EBI Web Developers.

In 2014 we will set up a common standardised local development environment, to allow us to work in the same environment as our production servers. This environment should allow flexibility for developer preferences (such as IDE) and also be easily deployable. Our team so far has not had a definitive testing framework. In 2014 we will investigate setting up a testing infrastructure (i.e. unit testing, functional testing, continuous integration, behavioural testing), which will greatly increase the reliability of our sites and integrating new developments.

Analytics are a vital tool to assess the effectiveness of a web resource, or the impact of changes to an interface. In 2014, we will evaluate our current analytics requirements and address any gaps in our understanding. The team will carry out the 2014 EBI user survey and analysis, drawing on best practices learned from the pilot survey in 2013.

We will continue to work with EMBL Heidelberg and EMBL-EBI Literature services to generate a coherent strategy for publication management and a pipeline for extracting publication information for reporting or website display purposes. We will continue to collaborate with the Web Production team on improving the discoverability bioinformatics tools on the website, addressing any development issues that arise.

Selected outreach events

Invited talks

Drupal at the EBI (GMOD, Cambridge)

A Survival guide for Complex UX

User research: The gentle art of not asking users what they want

Designing with the user in mind: how UCD can work for bioinformatics

Investigating usage of EBI services: highlights from the 2013 Service Survey (GTL perspective and EBI day)

Surveys and workshops

EBI Service Survey 2013

BioJS monthly meeting

Industry programme: Biomarkers workshop planning

Industry programme: Rare diseases workshop planning

EMTRAIN/On-course training

Selected publications

de Matos, P., Cham, J.A., Cao, H., et al. (2013) Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. BMC Bioinform. 14, 103.

Alcántara, R., Onwubiko, J., Cao, H., et al. (2012) The EBI Enzyme Portal. Nucleic Acids Res. 41, D773-D780.

Pavelin, K., Pundir, S. and Cham, J.A. (2014) Ten Simple Rules for Running Interactive Workshops. PLoS Comp. Biol. 10, e1003485.

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 (e.g. Oracle, MySQL), 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

Computing, clouds and virtualisation

Throughout 2013 the demand for compute from EMBL-EBI users grew. LCFC funds supported significant growth of the Hinxton data centre cluster, from 12 000 cores in January to 17 000 in May. Total installed RAM was 50 terabytes in January and 74 terabytes by May. Average core utilisation of this cluster increased 23% over 2012 levels. A development 'Hadoop' cluster is in use by several groups and will shortly enter production.

Our database administration staff coordinated efforts to grow our infrastructure to handle virtualisation of EMBL-EBI's database servers. There were 94 Oracle virtual machines (VM) in use across the three data centres as of November 2013. MySQL followed the same pattern. We also built a Delphix VM infrastructure offering 80 terabytes of RamSan™ storage. To support this increase, we installed 20 additional hypervisors. Storage use in the virtual infrastructure is increasing 25% year on year, and new storage appliances have been installed to accommodate this growth.

Our team coordinates Embassy Cloud activities. Trials continued throughout 2013, when we invited 12 organisations to participate and gave them access to the VMware-based cloud. We participate actively in a number of collaborations involving cloud use, for example the International Cancer Genome Consortium (ICGC) Pan-Cancer initiative, an 'Enlighten Your Research' collaboration and the Tara Oceans consortium. All of these collaborations call for a dramatic scaling up of the EMBL-EBI's cloud infrastructure, from hundreds to thousands of cores. This increase poses a number of technical challenges, and we are sought to address cost management issues by investigating Open Source software solutions such as OpenStack and OpenVZ. We also continued to negotiate with VMware over licensing costs for their proprietary software stack.

In 2013 we also undertook significant work to implement the EMBL-EBI security committee's recommendations on user accounts.

Networking

We completed the first phase of a LAN upgrade in Hinxton in 2013. During the first phase of this project we built an entirely new LAN in parallel with the legacy network. A significant proportion of devices, including all newly installed devices, are now using a new infrastructure that offers improved latency, greater bandwidth and resilience to switch failure.

We integrated the new EMBL-EBI South building into the campus network in 2013. This building houses approximately 200 staff, two IT training rooms, a lecture theatre and several industry suites.

We completed an ELIXIR pilot project with CSC Finland, providing a lightpath (i.e. ethernet circuit) over European Academic networks between CSC and EMBL-EBI. The project demonstrated the feasibility of performing replication of large datasets over the lightpath. We will test this in the context of the Enlighten Your Research project with CSC in 2014.

Storage

Our team put into operation all the plans we formulated in 2012. As in previous years, the amount of storage used at EMBL-EBI roughly doubled, from 3.7 petabytes in January to 7.4 petabytes in November. These figures do not include backups, replicas or temporary files. Our average storage utilization is 65%.

A new backup policy was agreed amongst Group and Team Leaders, enabling us to retire all tape-based backups. We introduced our first Object Storage system and built software for a long-term data archiving facility.

Petteri Jokinen

MSc in Computer Science 1990, Helsinki University.

At EMBL-EBI since 1996.

Databases

Longer-term projects during 2013 were characterised by the migration of the Oracle and MySQL databases and of Delphix from physical to virtualised infrastructure. This migration allowed the institute to achieve consolidation of database hardware resources, refresh of technology and reduction of database footprint in the data centres. We also deployed migration from Oracle 11.1 to 11.2. We devoted substantial efforts to introducing OpenSource 'hot' backup Percona Xtrabackup to the MySQL backups at EMBL-EBI. Our databases group acquired MongoDB skills and started to plan a central infrastructure that would provide MongoDB as a service. Our team designed a resource-accounting method that is accessible via the web, allow GTLs to control the database resources that are in use by their group.

Future plans

We plan to develop the Embassy Cloud into an infrastructure capable of supporting organiations requiring access to thousands of cores and petabytes of storage. We already have a list of users including the ICGC's Pan-Cancer initiative, Tara Oceans, GlaxoSmithKline and potential projects from the UK Biobank. Trials are already taking place with some of these users on the existing cloud infrastructure, and these will allow us to closely define user requirements and inform our choice of hardware and software. We intend to build the first facility to house 2000 cores in early 2014.

We plan to implement an Object Tape Archive, reusing existing tape backup infrastructure. We are also working to have a new version of software for our Sequence Archive. This software, which separates data and metadata, will enable Object Storage to be used.

We will investigate the feasibility of using a parallelised high-performance computing file system for the Hinxton data centre cluster and on the Embassy Cloud. We will also ensure that EMBL-EBI will benefit from a direct connection to the Janet6 network in 2014.

Servers at the Wellcome Trust Genome campus, Hinxton.

Administration

The EMBL-EBI Administration team aims to facilitate the work of the institute through 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 Large Facilities Capital Fund (LFCF); and pre- and post-doctoral programmes. We coordinate and integrate administrative activities throughout EMBL-EBI to facilitate interactions with the wider scientific community through, for example, organising meetings and courses and arranging travel for our extremely mobile staff.

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

Major achievements

In the United Kingdom'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. In 2013 our team devoted substantial efforts to implementing this programme, for example supporting project management and liaising with facilities staff.

The EMBL-EBI South building was officially opened by the United Kingdom's Science Minister, the Right Hon. David Willetts, in October 2013. The first suppliers Framework Agreement for the acquisition of equipment to be placed in off-site Data Centres is well established and a number of successful 'mini-competitions' concluded by the Systems and Networking Group. The current license on Data Centre Space expires in Dec 2014 and work is well in-hand on the next OJEU procurement exercise to acquire space for the period 2015-2018.

Work has also commenced the development and gathering of data demonstrating the benefits being delivered under the programme of work, and to engage with EBI staff on the gathering of data to interpret the impact of EMBL-EBI as a whole as a measure of the success of the LFCF programme.

We continued to refine our financial and budgetary processes over 2013. This included a progressive switch to GBP rather than Euro accounting following an EMBL Council decision that the United Kingdom contribution should be made in sterling in order to help minimize exchange rate fluctuations. We are contributing to the development of new EMBL financial reporting tools which will be introduced in 2014.

Our Grants Office is contributing to the development of EMBL wide administrative processes in respect of external funds and is playing a significant role in the review of existing grants.

We have undertaken a restructuring of our Finance and Purchase section to ensure, as resources have grown, that we continue to exercise sound financial control.

The Human Resources (HR) team have also been involved in a number of reviews of processes, including contracts, flexible working, comparability between teams, amongst others. The HR team continue work on improving processes and the development of internal documentation to ensure we provide accurate and consistent advice to staff and supervisors. As an example, the HR Team streamlined the process for hiring visitors and trainees, and issued guidelines to supervisors. We have also now had one full year of running and refining newcomers meetings; meetings which feedback show to be appreciated by attendees. In 2013 we improved the exit processes and the information provided to leavers and the HR team started to collect feedback from leavers which we plan to analyse early in 2014. We have also been involved in making sure that staff based at the EBI have an ORCID number enabling EMBL to keep track of the wealth of publications emanating from the EBI.

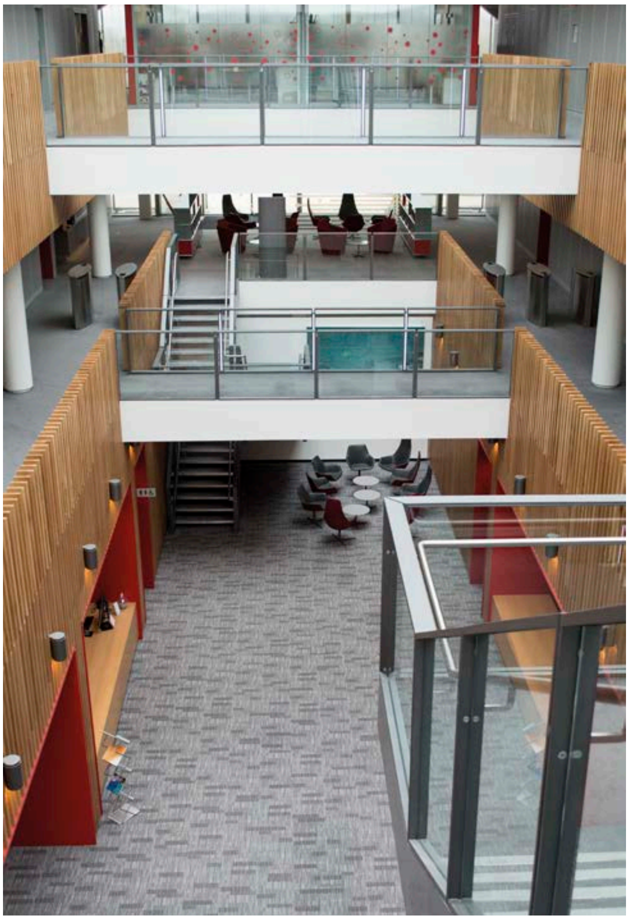
Work has continued on developing the intranet, which now includes a section on recruitment with process guidelines and tips for creating effective vacancy ads and running selection interviews. (revamped in November 2012) as the repository of easily accessible information and help for all members of personnel.

We continue to participate in a wide range of cross-campus activities and initiatives including EMBL-EBI/WTSI meetings, Health and Safety, Campus Library and the 'Sex in Science' programme.

Mark Green

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

At EMBL-EBI since 2003.



Interior of the new EMBL-EBI South building, home of the Training programme and the EXLIR hub.

Summary of progress

- *Organised and delivered the administrative aspect of the United Kingdom's Large Facilities Capital Fund Programme, providing both a building on campus and equipment/ and off-site data facilities until the end of 2019;*
- *Continued the development of the 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 the EMBL procedures and to the development of new reporting software;*
- *Contributed to the development and implementation of the new intranet;*
- *Continued to develop and sustain our Health & Safety practices and procedures.*

Future plans

We will continue developing 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 develop the quality of guidance provided to Group and Team Leaders and introduce changes to performance assessment to benefit all staff.

We will continue to maintain good interactions between a wide diversity of stakeholders such as the BBSRC, Wellcome Trust and NIH.

Virtually everyone in the Institute has, as a result of the migration to the EMBL-EBI South building, moved workstation, even if they have remained in the 'old' buildings. This requires restarting the programme of ergonomic workplace assessment, necessary for staff working in a computer-intensive environment. We had achieved a very credible 80% assessment rate; the counter has effectively been reset to zero and it will take time to rebuild the depth and scope of coverage. We will continue supporting the procurement process for the acquisition of Data Centre space 2015 -2019, alongside support for the Framework Agreement for suppliers of equipment and the capture and reporting of benefits under the Programme as whole.

We will also be looking at the some of the opportunities now presented, following the reorganization of space in the 'old' EMBL-EBI buildings, to optimize meeting room space.

ELIXIR

ELIXIR integrates Europe's best bioinformatics resources so that scientists can make the most of the data deluge. Initiated in 2007 with backing from the European Commission, ELIXIR now has the support of EMBL and seven member nations; nine further countries have signed a Memorandum of Understanding (MoU) to participate.

ELIXIR integrates Europe's best bioinformatics resources so that scientists can make the most of the data deluge. The preparatory phase kicked off in 2007 with backing from the European Commission, ELIXIR now has the support of EMBL and seven member nations; nine further countries have signed a Memorandum of cUnderstanding (MoU) to participate.

In May 2013, ELIXIR welcomed its first Director: Dr Niklas Blomberg, formerly of pharmaceutical company AstraZeneca. Throughout the year Niklas worked closely with Janet Thornton, who led the preparatory phase of this crucial research infrastructure, towards the initiation of implementation phase. On 18 December 2013, ELIXIR celebrated its launch as a permanent legal entity following the ratification of the ELIXIR Consortium Agreement (ECA) by EMBL and the first five countries.

ELIXIR becomes a legal entity

The public launch of ELIXIR in Brussels on 18 December 2013 comprised a signing ceremony as well as a series of talks and posters about ELIXIR. The event was attended by Robert-Jan Smits, Director General of the European Commission's DG Research, as well as national ELIXIR representatives and high-level national and European funders.

EMBL was the first organisation to ratify the ECA in June 2013, followed in order by UK, Sweden, Switzerland, Czech Republic, Estonia, Norway, Netherlands and Denmark throughout the year. The ELIXIR Nodes in these countries have also now been established. Countries that have signed the MoU include Finland, Greece, Israel, Italy, Portugal, Slovenia and Spain, with Belgium and France signing in 2013.

The ELIXIR Hub

During 2013 the ELIXIR Hub appointed a Chief Technical Officer, a Grants and Programme Manager and an Events Officer. The newly strengthened team is housed in the South building on the Genome Campus, which opened in October 2013.

The team worked throughout the year in consultation with ELIXIR member states to develop a scientific programme for 2014-2018. The ELIXIR Board is scheduled to adopt

ELIXIR moved into its implementation phase following the entering into force of the ELIXIR Consortium Agreement and the public launch in Brussels in December 2013. Pictured: Robert-Jan Smits, Janet Thornton and Rolf Apweiler.

Niklas Bloomberg

BSc in Chemistry, Göteborg University, Sweden. PhD from EMBL, Heidelberg. AstraZeneca, 1999 to 2013.

Founding Director, ELIXIR since 2013.

the programme, which forms the basis for ELIXIR's activities, in 2014. The newly formed Technical Coordinators group will develop ELIXIR's technical strategy on issues including cloud computing and interactions with other e-Infrastructures in 2014. The template Collaboration Agreement, approved in early 2014, will form the basis for services to will be run by ELIXIR Nodes over the coming years.

ELIXIR Pilot Actions

ELIXIR's first pilot actions, funded in 2012, are scalable, technical projects that tackle major European challenges in life-science data access, high-performance computing and the interoperability of public biological and biomedical data resources.

Seamless transfer of major datasets across Europe

Downloading massive datasets for analysis behind a firewall is often necessary for research. EMBL-EBI and the Finnish Node of ELIXIR are building tools for the dedicated, secure and private transfer of large datasets between the UK and Finland. Scaling up these tools will enable the timely and dependable transfer of massive datasets between European institutes over allocated lines.

Secure access to personal genomic data

The European Genome-phenome Archive enables the secure sharing of research data that matches genetic information to disease characteristics. Continually improving secure access to these data is a key driver behind this collaboration between EMBL-EBI and the Finnish Node of ELIXIR. Endorsed by Geant3Plus, this action has become a priority model for similar projects.

Interoperability of protein resources for drug discovery

EMBL-EBI is working with the Swedish ELIXIR Node to make the Human Protein Atlas interoperable with established resources PRIDE, InterPro and the Expression Atlas. This will make navigation seamless between protein resources.

Safeguarding resources that link genes with disease

EMBL-EBI has partnered with the Centre for Genomic Research in Barcelona to further develop the EGA to ensure it can grow to meet demand, and in a short time has eased considerable pressure on the project and opened up new avenues for development. This Pilot Action demonstrates how ELIXIR Nodes can support one another in expanding Europe's bioinformatics capacity.

BioMedBridges at the launch of ELIXIR.
Pictured: Niklas Blomberg, Robert-Jan Smits and José Cotta.

BioMedBridges

BioMedBridges is a joint effort between ten biomedical sciences research infrastructures on the ESFRI roadmap. It spans many disciplines, from structural biology and genomics to translational research and clinical trials. Together, the project partners are developing the shared e-infrastructure—the technical bridges—to allow data integration in the biological, medical, translational and clinical domains and thus strengthen biomedical resources in Europe. The project runs from 2012 to 2015.

BioMedBridges completed its first set of major technical deliverables in 2013, including a comprehensive assessment of the legal and ethical landscape surrounding data sharing and integration a first set of data integration tools. These tools now make existing data more discoverable and enable their efficient use, raising all infrastructures to a higher level of data interoperability readiness.

The services are based on the pilot data integration using REST web services and the identification of feasible pilots for semantic web integration between the project partners. These services include: Euro-BioImaging's server for the secure sharing and integration of medical imaging data; EU-OPENSCREEN's new, connectivity-based search function for UniChem; a method for sharing and visualising sequencing data from environmentally derived biological samples, an in-kind contribution from EMBRC and the Micro B3 project; Euro-BioImaging's tools for the visualisation and leveraging of ontologies in queries; and ECRIN's tool for integrating gene and drug information with a clinical trials registry.

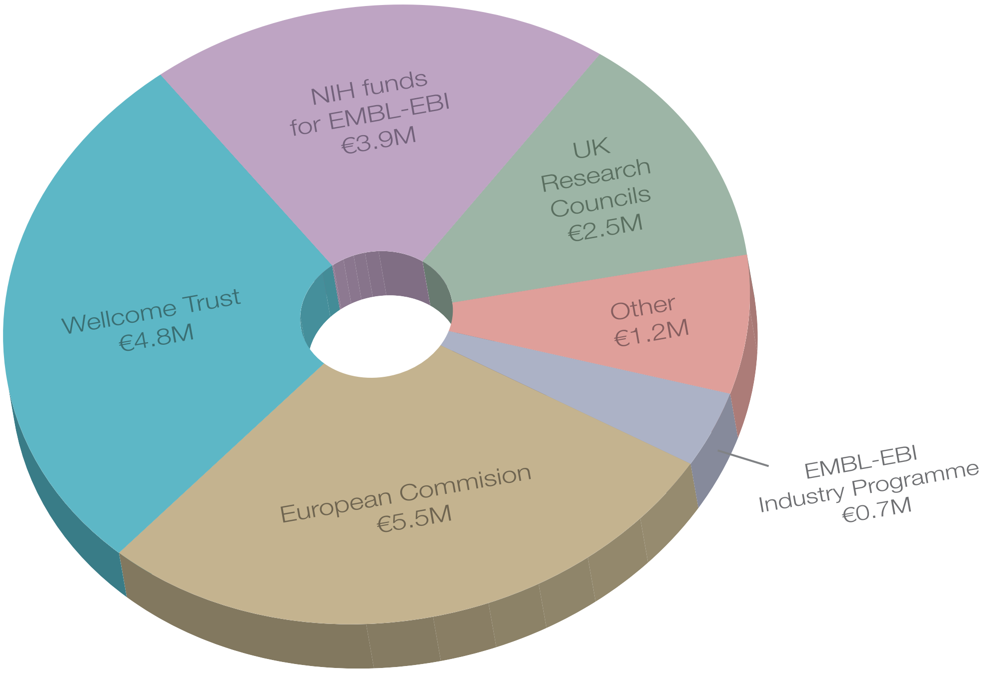
Funding and resource allocation

Despite the on-going challenges to research funding worldwide, including in EMBL member states, EMBL-EBI funding remained stable in 2013. This continued support of our member states and other funding bodies in 2013 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 2013. The 'external funds' shown here represent both funds that were available for our use in 2013 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 2013 was €53.7 million and comes primarily from EMBL member states (€35 million). Our major sources of external funding include the European Commission (€5.5 million), the Wellcome Trust (€4.8 million), the United States National Institutes of Health (€3.9 million to EMBL-EBI for direct use the UK Research Councils (€2.5 million) and the EBI Industry Programme (€0.7 million). We also benefit from a large number of grants from various other sources (total, €1.2 million). These major sources of funding are shown in the figure below.



(*)In addition to these sums EMBL-EBI's funding in 2013 included funds earmarked for project subcontractors of €5.6 million (NIH €4.9m,   Wellcome Trust €0.6m, others € 0.1m)

Capital investment

In 2013 we moved many of our staff into a new EMBL-EBI South building on the Genome Campus, which also houses the ELIXIR Hub and an Innovation and Translation suite for large-scale collaborations with industry. It was also the first full year of running EMBL-EBI services out of the tier-3+ security London Data Centres. Both were made possible by additional funding committed by the UK government's Large Facilities Capital Fund: £10 million (€11.5 million) in 2009 to enable data service provision, including acquisition of space and equipment, and £75 million (€90 million) in 2011 to maintain our high-security data centre space and equipment and build the new EMBL-EBI premises on campus. The new South building, like all structures on the Genome Campus, is owned by the Wellcome Trust and its upkeep is covered by EMBL-EBI.

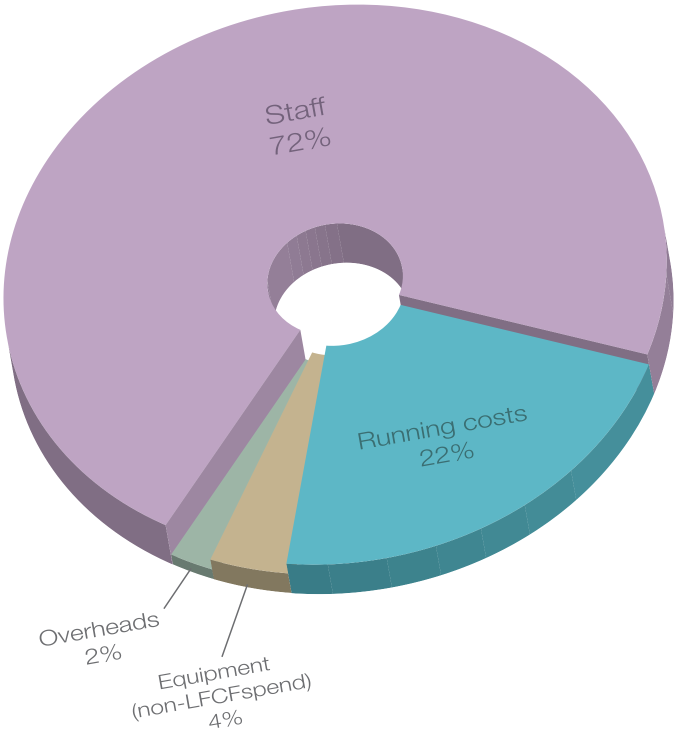
In 2013, LFCF funds were used to finalise the construction of the new South building and purchase equipment for the London data centres. Over the next five years the remaining capital funds received from the UK government will be spent to continue supporting the smooth running of our high-security data centre space and equipment.

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

Year	Funding for London data centres	Funding for Technical Hub (EBI South building)	Total funding received
2012	€ 0.39 M	€ 6.7 M	€ 7.5 M
2013	€ 10.4 M	€ 15.4 M	€ 25.7 M

Spending

This shows the breakdown of EMBL-EBI's total spend for 2013 (€54.8 million excluding sums expended by grant subcontractors).



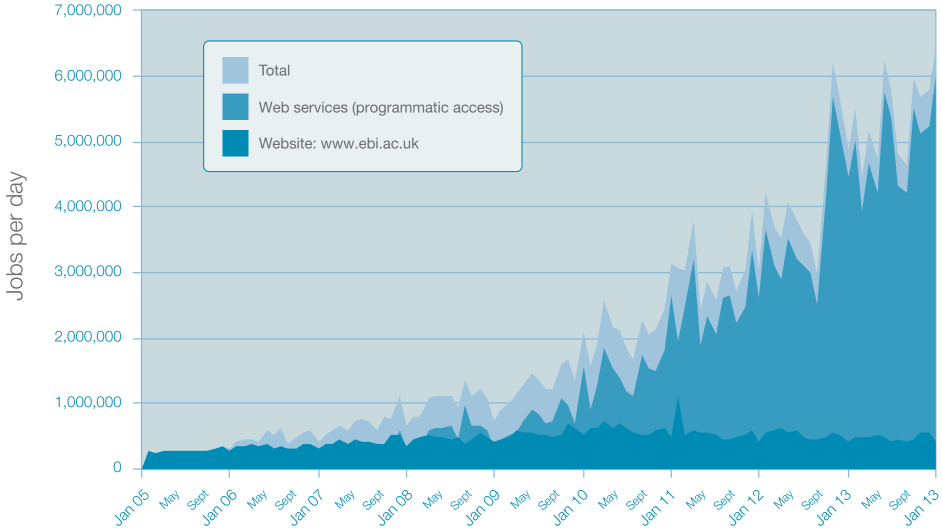
Growth of core resources

In 2013 there were on average 9.8 million requests on our services per day, including Ensembl (compare to 9.2 million in 2012). During 2013 more than 65 million jobs were run using the jdispatcher framework.

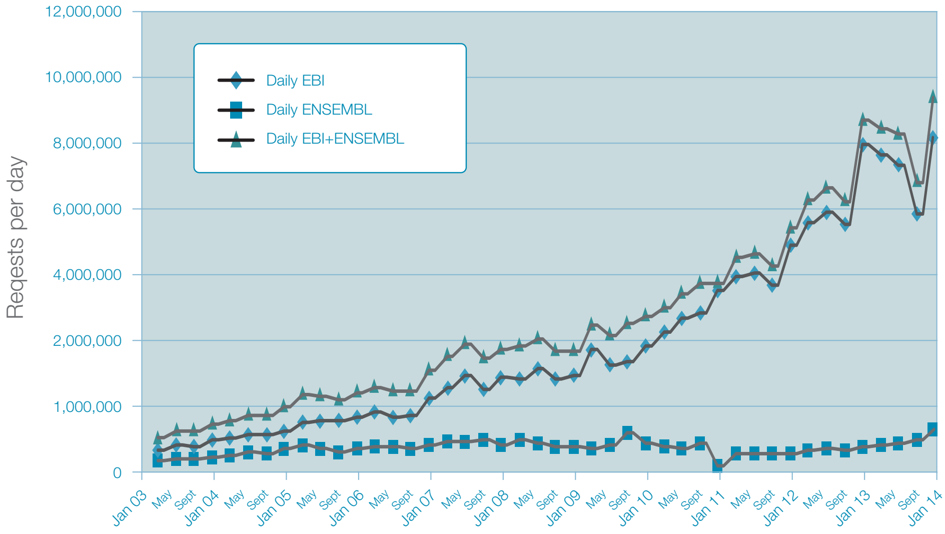
The most popular services were InterProScan, Clustal Omega and NCBI Blast+; new additions include the HMMer suite for TreeFam and Pfam. The average number of jobs per month was 5.4 million (up from 4 million in 2012). The number of datasets available for searching was over 60,000 in 2013. This is based on the number of class- or taxonomy-specific files being produced to generate BLAST and InterProScan services.

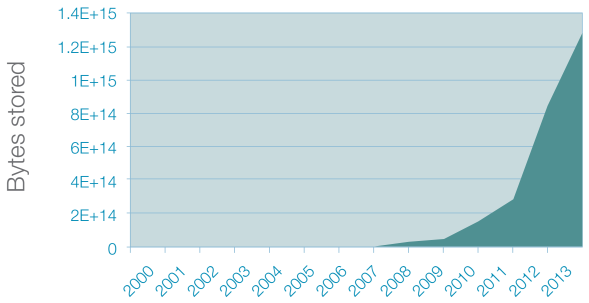
We continue to see steady growth in usage and in the number of computers accessing our services. The number of unique IPs or web addresses accessing our website grew by 22.8% during 2013 (compared to 16.2% growth in 2012). This figure is based on cumulative IP counts for both years. We approach these figures with caution, as an IP address could represent a single person or an entire organisation, so the figures represent a minimum number of users.

Jobs per day, 2005 through 2013

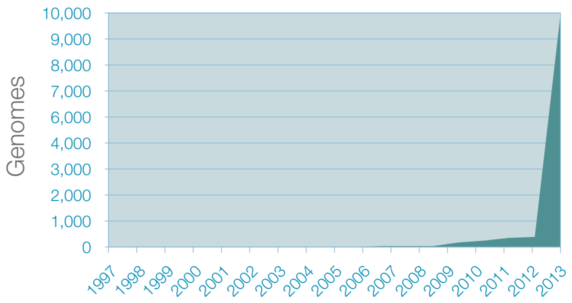


Requests per day, 2003 through 2013 (quarterly)

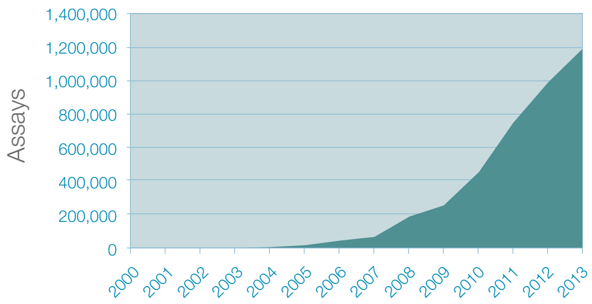




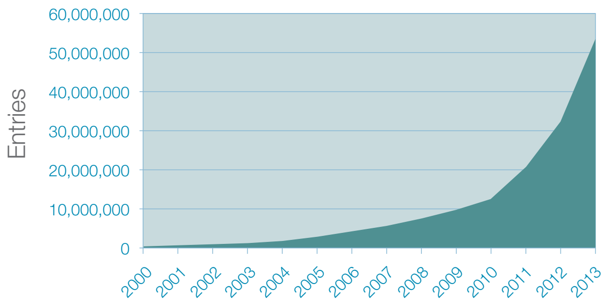
Nucleotide sequence data (compressed)



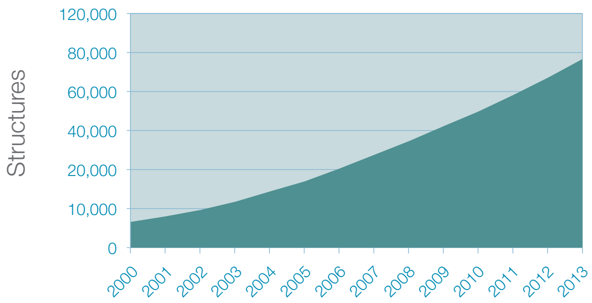
Genomes (all species)



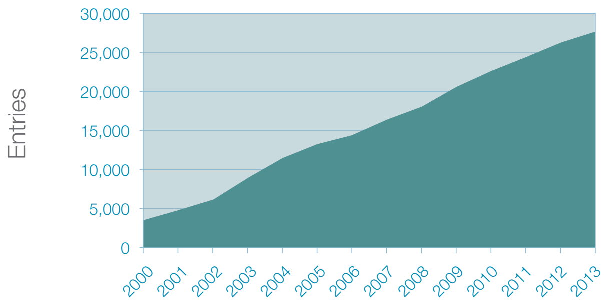
Gene expression data



Protein sequence



Macromolecular structures



Protein families motif and domains

In 2013 all of our core data resources grew substantially. The cost of generating data continues to fall, which has a dramatic impact on EMBL-EBI databases. Innovative ways of storing such data are beginning to have traction to reduce the required storage.

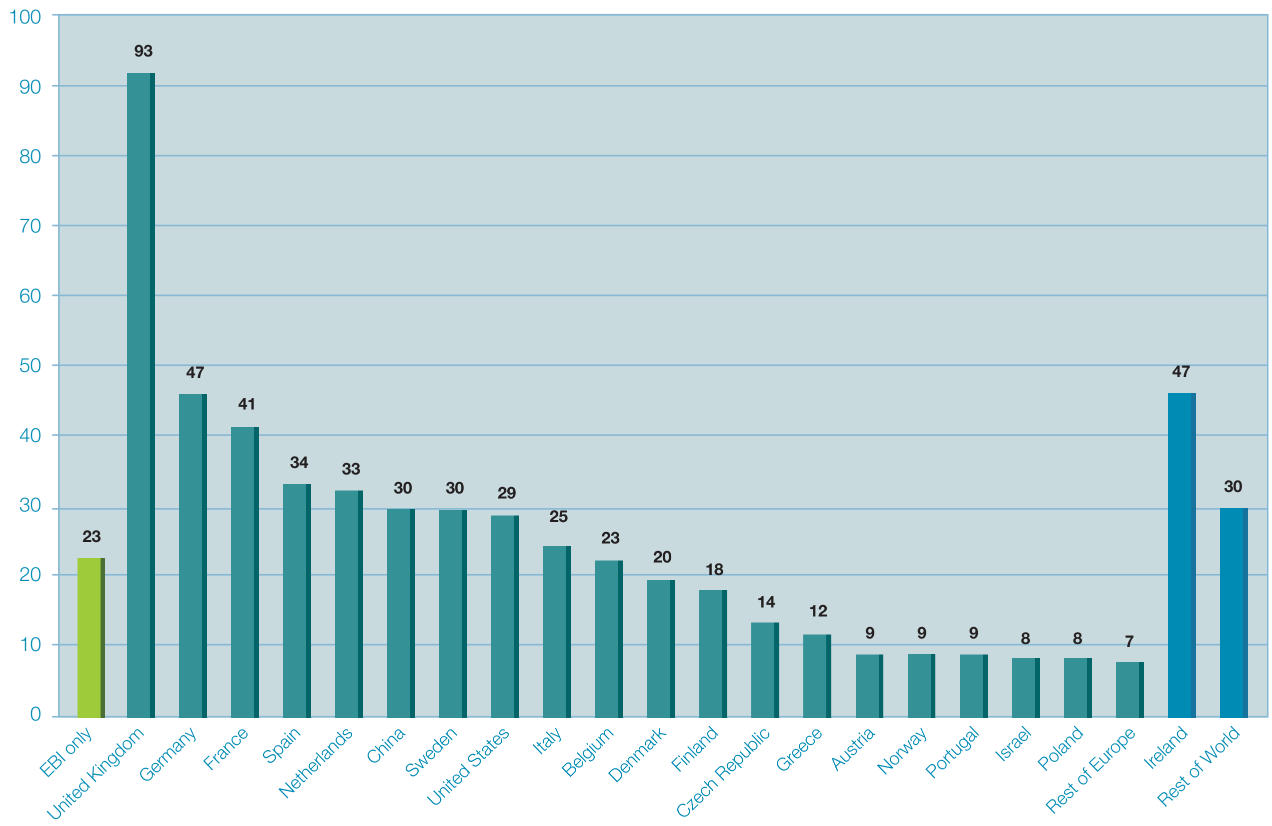
Compressed nucleotide sequence data: 1.14 petabytes stored (compare to 0.75 petabytes in 2012). A petabyte is 1 x 1015 bytes. Disk usage for nucleotide sequence data storage at EMBL-EBI would have grown even more, were it not for the recent implementation of the CRAM data compression algorithm. Genomes, all species and strains: 11 010 (compare to 417 in 2012). Gene expression assays: 1.25 million (compare to 1.04 million in 2012). Protein sequences: 49.2 million (compare to 29.8 million in 2012). Macromolecular structures: 96 574 (compare to 86 954 in 2012). Protein families, motifs and domains - entries in InterPro: 25 326 (compare to 24 117 in 2012).

Scientific collaborations

We are proud to work with our colleagues throughout the world on setting standards, exchanging information, improving methods for analysis and sharing curation of the complex information we make accessible to the public. Our research programme in particular is highly collaborative, and their outstanding work in many different aspects of biology is conducted in partnership with a large network of academic peers.

Funding

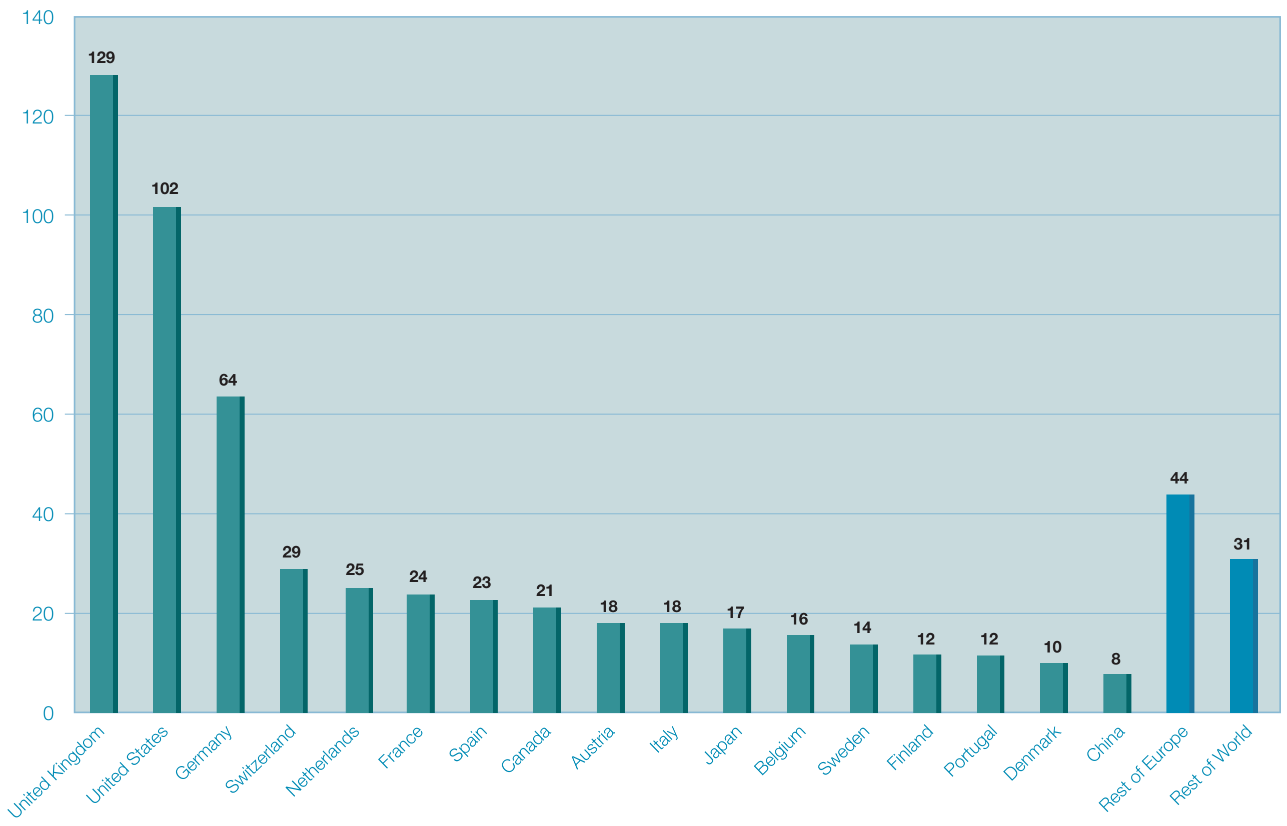
In 2013, EMBL-EBI had joint grant funding with researchers and institutes in 58 countries throughout the world, most notably in the United Kingdom, Germany, France and Spain but also with colleagues in unexpected places like Uganda. Of the 142 grants received, only 23 were exclusively for EMBL-EBI. These figures are potentially underestimated, as all partners are not always listed on grants.



Publications

In 2013 most of our scholarly publications were co-authored with colleagues at other institutes throughout the world. To get an idea of how collaborative we are in terms of publications, we queried the Scopus database to find papers with authors affiliated with EMBL-EBI. This search returned 271 scientific papers, and revealed that in 2013 our most productive partnerships were with people at partner institutes in the United Kingdom, United States and Germany. Figure 3 shows the international scope of our collaborations.

Producing a detailed view of all publications from an institute of our size can be a challenge for a number of reasons. One challenge is capturing papers published after the authors have moved on from EMBL. Another is publication timeframes: some papers published during the year by our current staff are based on work done elsewhere. From a data perspective, public literature resources have only recently begun capturing full author affiliation information. These are just some of the reasons why EMBL has become a member of ORCID, the public, open registry of unique researcher identifiers. In 2013 all members of EMBL-EBI staff were assigned an ORCID and began to claim their articles. In 2014 the laboratory will roll out a new publications reporting system that compares ORCID and employment information, which should improve the accuracy of our research output reporting.

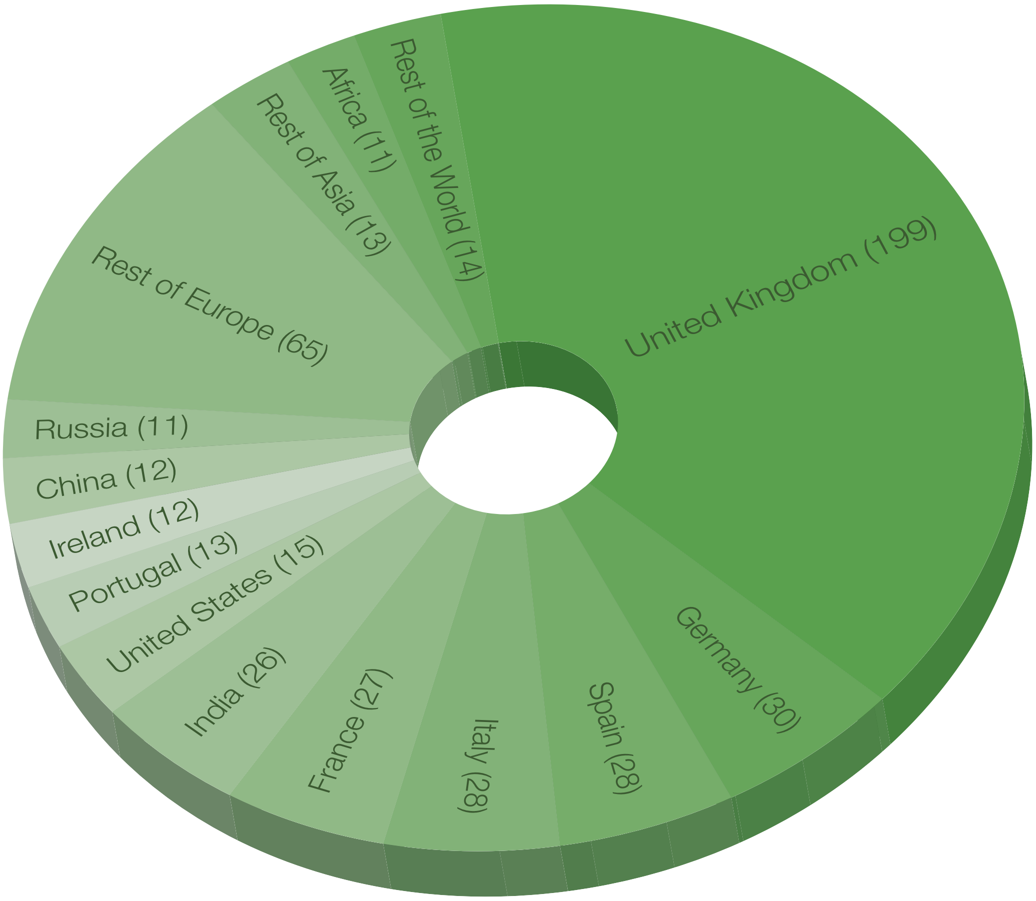


Our staff in 2013

As a European organisation we are proud to report that our personnel in 2013 represented 55 nationalities (53 in 2012).

Our organisational structure reflects our mission: services, research, training and industry support, with overarching internal support. We had 504 members of staff in 2013, and hosted 76 visitors. Of these visitors, eight were transitional visitors from the Wellcome Trust Sanger Institute who rejoined Alex Bateman's team at EMBL-EBI and will included in the full staff figures for 2014. The remaining 68 joined us for longer than one month (compare to 58 in 2012).

We extend a warm welcome to Niklas Blomberg, the new Director of ELIXIR; Steven Newhouse, our new Head of Technical Services; Jason Mundin, seconded from EMBLEM to lead our Innovation and Translation programme; Research Group Leaders Sarah Teichmann and Pedro Beltrao; and Rob Finn, who will run the new Protein Families team in 2014.



EMBL-EBI postdocs organised the 2013 EMBL Postdoc Retreat, a family-friendly event held at Jesus College, University of Cambridge. An Outstanding Postdoc Award was presented by Mark Patterson, Executive Director of eLife, to EMBL Hamburg's Ciaran Carolan.

Scientific Advisory Commitees

EMBL Scientific Advisory Committee

- *Nenad Ban, Zurich, Switzerland (2013-2015)*
- *Denis Duboule, Geneva, Switzerland (2011-2013, 2014-2016)*
- *Roderic Guigo, Spain (2012-2014)*
- *Anthony Hyman, Dresden, Germany (2014-2016)*
- *Reinhard Jahn, Germany (2010-2012, 2013-2015)*
- *Daniel Louvard, France (2012-2014)*
- *Ron Milligan, United States (2012-2014)*
- *Tom Muir, United States (2010-2012, 2013-2015)*
- *Andrea Musacchio, Germany (2011-2013, 2014-2016)*
- *Stefano Piccolo, Padua, Italy (2014-2016)*
- *Venki Ramakrishnan, United Kingdom (2003-2006 and 2013-2015)*
- *Michael Snyder, United States (2011-2013, 2014-2016)*
- *Alexander van Oudenaarden, Utrecht, The Netherlands (2013-2015)*
- *Jean Weissenbach, France (2012-2014)*

Bioinformatics Advisory Committee

- *Philippe Sanseau of GlaxoSmithKline, United Kingdom.*
- *Roderic Guigo of the Centre de Regulacio Genomica, Barcelona, Spain*
- *Tim Hubbard of King's College London, United Kingdom*
- *Olli Kallioniemi, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland*
- *Jonathan Knowles, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland*
- *Matthias Uhlén of the Royal Institute of Technology (KTH), Stockholm, Sweden*
- *BioModels Scientific Advisory Board*
- *Carole Goble, University of Manchester, United Kingdom*
- *Thomas Lemberger, Nature Publishing Group/EMBO*
- *Pedro Mendes, University of Manchester, United Kingdom*
- *Wolfgang Mueller, HITS, Germany*
- *Philippe Sanseau, GSK, United Kingdom*

Cheminformatics: ChEMBL and ChEBI Advisory Board

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

EMDataBank Advisory Board

- *Joachim Frank (Chair), Columbia University, United States*
- *Achilleas Frangakis, Goethe University Frankfurt, Germany*
- *Richard Henderson, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom*
- *Maryanne Martone, University of California San Diego, United States*
- *Michael Rossmann, Purdue University, United States*
- *Andrej Sali, University of California, United States*
- *Paula Flicker (observer), National Institute of General Medical Sciences, United States*

Ensembl Scientific Advisory Board

- *Toby Bloom, Broad Institute of Massachusetts Institute of Technology and Harvard, Boston, MA, United States*
- *Deanna Church, Personalis, Menlo Park, CA, United States*
- *Jim Reecy, Iowa State University, Iowa, United States*
- *Hugues Roest Crolli, United States, Ecole Normale Supérieure, Paris, France*
- *Erich Jarvis, Duke University Medical Center, Durham, NC, United States*
- *Ian Bird, CERN, Switzerland*
- *Mark Diekhans, Center for Biomolecular Science and Engineering, University of California Santa Cruz, United States*

- *Anne Ferguson-Smith, University of Cambridge, United Kingdom*
- *Matt Hurles, Wellcome Trust Sanger Institute, Hinxton, United Kingdom.*

Ensembl Genomes Scientific Advisory Board

- *Martin Donnelly, University of Liverpool, United Kingdom*
- *Klaus Mayer, Helmholtz Institute for Pharmaceutical Research, Saarland, Germany*
- *Claudine Medigue, Genoscope, France*
- *Allison Milller, University of St. Louis, United States*
- *Rolf Mueller, Helmholtz Institute, Bonn, Germany*
- *Chris Rawlings, Rothamsted Research, United Kingdom*
- *Jason Staijich, University of Riverside, United States*
- *Denis Tagu, INRA, 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 Genomica, Barcelona, Spain*
- *Tim Hubbard (Chair), Wellcome Trust Sanger Institute, Hinxton, United Kingdom*
- *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*

Gene Expression Scientific Advisory Board

- *Frank Holstege (Chair), University Medical Center Utrecht, Netherlands*
- *ill Mesirov, Broad Institute of MIT and Harvard, Boston, United States*
- *Roderic Guigo, Center for Genomic Regulation, Barcelona, Spain*
- *Chris Ponting, University of Oxford, United Kingdom*
- *Wolfgang Huber (observer), EMBL, Heidelberg, Germany*

The Gene Ontology Scientific Advisory Board

- *John Hogenesch, University of Pennsylvania, United States*
- *Matthew Hibbs, Trinity University, United States*
- *Gary Bader, University of Toronto, Canada*
- *Andrew Su, Scripps Research Institute, United States*
- *Mike Tyers, University of Montreal, Canada*
- *Simon Tavare, University of Southern California, United States and University of Cambridge, United Kingdom*
- *Weiwei Zhong, Rice University, United States*

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

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

Scientific Advisory Commitees

InterPro/Pfam Scientific Advisory Board

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

Literature Services Scientific Advisory Board

- *Gianni Cesareni, University of Rome, Italy*
- *Wolfram Horstmann, Bodleian Library, Oxford, United Kingdom*
- *Tim Hubbard, King's College London, United Kingdom (Chair)*
- *Larry Hunter, University of Colorado Health Sciences Center, United States*
- *Mark Patterson, eLife, Cambridge, United Kingdom*
- *Carola Tilgmann, Lund University, Sweden*

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

- *Udo Heinemann, Max Delbrück Center for Molecular Medicine, Berlin, Germany*
- *Tomas Lundqvist, AstraZeneca R&D, Mölndal, Sweden*
- *Andrea Mattevi, University of Pavia, Italy*
- *Randy Read, University of Cambridge, United Kingdom (Chair)*
- *Helen Saibil, Birkbeck College London, United Kingdom*
- *Michael Sattler, TUM, Munich, Germany*
- *Torsten Schwede, Swiss Institute of Bioinformatics, Switzerland*
- *Titia Sixma, Netherlands Cancer Institute, Amsterdam, Netherlands*

PRIDE Scientific Advisory Board

- *Ruedi Aebersold, ETH Zurich, Switzerland*
- *Kathryn Lilley, University of Cambridge, United Kingdom*
- *Ioannis Xenarios, SIB Swiss Institute of Bioinformatics, Switzerland*
- *Roz Banks, University of Leeds, United Kingdom*
- *Angus Lamond, University of Dundee, United Kingdom*

Reactome Scientific Advisory Board

- *Russ Altman, Stanford University, United States*
- *Gary Bader, University of Toronto, Canada*
- *Richard Belew, University of California San Diego, United States*
- *John Overington, EMBL-European Bioinformatics Institute, United Kingdom*
- *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*

RNACentral Scientific Advisory Board

- *John Rinn, Harvard University, United States*
- *Sean Eddy, Howard Hughes Janelia Farm Research Campus, United States*
- *Eric Westhof, University of Strasbourg, France*
- *External Training Advisory Group*
- *Alex Bateman, EMBL-EBI, Hinxton, United Kingdom*
- *Bogi Eliasen, 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*

The Universal Protein Resource (UniProt) Scientific Advisory Board

- *Patricia Babbitt, University of California San Francisco, United States*
- *Helen Berman, Rutgers University NJ, United States*
- *Judith Blake, The Jackson Laboratory, ME, United States*
- *Ian Dix, AstraZeneca, Macclesfield, United Kingdom*
- *Takashi Gojobori, National Institute of Genetics, Mishima, Japan*
- *Maricel Kann, University of Maryland, Baltimore, United States*
- *Bernhard Kuester, Technical University Munich, Weihenstephan, Germany*
- *Edward Marcotte, University of Texas, Austin, United States*
- *William Pearson, University of Virginia, Charlottesville, United States*
- *David Searls (Freelancer)*
- *Minoru Kanehisa, Institute for Chemical Research, Kyoto, Japan*
- *Mathias Uhlén Royal Institute of Technology (KTH), Stockholm, Sweden (Chair)*
- *Timothy Wells, Medicines for Malaria Venture, Geneva, Switzerland*

Worldwide Protein Data Bank (wwPDB) Advisory Board

- *Jianpeng 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*
- *Keiichi Namba, Osaka University, Japan*
- *Michael G. Rossmann, Purdue University, United States*
- *Helen Saibil, Birkbeck College London, United Kingdom*
- *Titia Sixma, Netherlands Cancer Institute, Amsterdam, Netherlands*
- *Soichi WakatsUKi, High Energy Accelerator Research Organisation (KEK), Japan*
- *Cynthia Wolberger, Johns Hopkins School of Medicine, United States*
- *Edward N. Baker, University of Auckland, New Zealand (Ex Officio)*
- *R. Andrew Byrd, NIH, United States (Ex Officio)*

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),*
- *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*

UniProt

- *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*
- *SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*
- *Protein Information Resource (PIR), Washington, DC, 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, Alberquerque, 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 (RCSB), Piscataway, New Jersey, United States*

Publications in 2013

The following list represents publications by researchers affiliated with EMBL-EBI in 2013. To generate this list we queried the ScopUS database (Elsevier), which mines author affiliation data, and ORCID, the open registry of unique researcher identifiers. In 2013 all members of EMBL-EBI staff were assigned an ORCID; Group and Team Leaders and others throughout the institute USed it to claim their articles. Many of the papers listed here are the result of work done by individuals who have since moved on from EMBL-EBI.

001. 't Hoen, P.A., Friedländer, M.R., Almlöf, J., et al. (2013) Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. 31, 1015-1022.

002. Abyzov, A., Iskow, R., Gokcumen, et al. (2013) Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 23, 2042-2052.

003. Aghaeepour, N., Finak, G., et al. (2013) Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods 10, 228-238.

004. Aguiar, B., Vieira, J., Cunha, A.E., et al. (2013) Patterns of evolution at the gametophytic self-incompatibility Sorbus aucuparia (Pyrinae) S pollen genes support the non-self recognition by multiple factors model. J. Exp. Bot. 64, 2423-2434.

005. Ajmera, I., Swat, M., Laibe, C., et al. (2013) The impact of mathematical modeling on the understanding of diabetes and related complications. CPT Pharmacometrics Syst Pharmacol 2, e54.

006. Ala, A., Brown, D., Khan, K., et al. (2013) Mucosal addressin cell adhesion molecule (MAdCAM-1) expression is upregulated in the cirrhotic liver and immunolocalises to the peribiliary plexus and lymphoid aggregates. Dig. Dis. Sci. 58, 2528-2541.

007. Alcántara, R., Onwubiko, J., Cao, H., et al. (2013) The EBI enzyme portal. Nucleic Acids Res. 41, D773-80.

008. Altenhoff, A.M., Gil, M., Gonnet, G.H. and Dessimoz C. (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8, e53786.

009. Anderson, J.W., Haas, P.A., Mathieson, L.A., et al. (2013) Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information. Bioinformatics 29, 704-710.

010. Anton, B.P., Chang, Y.C., Brown, P., et al. (2013) The COMBREX project: design, methodology, and initial results. PLoS Biol. 11, e1001638.

011. Anttila, V., Winsvold, B.S., Gormley, P., et al. (2013) Genome-wide meta-analysis identifies new susceptibility loci for migraine. Nat. Genet. 45, 912-917.

012. Araújo, A.R., Reis, M., Rocha, H., et al. (2013) The Drosophila melanogaster methuselah gene: a novel gene with ancient functions. PLoS One 8, e63747.

013. Atanur, S.S., Diaz, A.G., Maratou, K., et al. (2013) Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. Cell 154, 691-703.

014. Azzaoui, K., Jacoby, E., Senger, S., et al. (2013) Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discov. Today 18, 843-852.

015. Balakrishnan, R., Harris, M.A., Huntley, R., et al. (2013) A guide to best practices for Gene Ontology (GO) manual annotation. Database (Oxford) 2013, bat054.

016. Barquist, L., Langridge, G.C., Turner, D.J., et al. (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. Nucleic Acids Res. 41, 4549-4564.

017. Bateman, A., Kelso, J., Mietchen, D., et al. (2013) ISCB computational biology Wikipedia competition. PLoS Comput. Biol. 9, e1003242.

018. Baud, A., Hermsen, R., Guryev, et al. (2013) Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat. Genet. 45, 767-775.

019. Beisken, S., Meinl, T., Wiswedel, B., et al. (2013) KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics 14, 257.

020. Beltrao, P., Bork, P., Krogan, N.J., et al. (2013) Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. 9, 714.

021. Berman, H., Kleywegt, G.J., Nakamura, H., et al. (2013) Comment on the propagation of errors by Jaskolski (2013). Acta Crystallogr. D Biol. Crystallogr. 69, 2297.

022. Berman, H., Kleywegt, G.J., Nakamura, H., et al. (2013) Comment on timely deposition of macromolecular structures is necessary for peer review by Joosten et al. (2013). Acta Crystallogr. D Biol. Crystallogr. 69, 2296.

023. Berman, H.M., Kleywegt, G.J., Nakamura, H., et al. (2013) How community has shaped the Protein Data Bank. Structure 21, 1485-1491.

024. Berman, H.M., Kleywegt, G.J., Nakamura, H., et al. (2013) The future of the Protein Data Bank. Biopolymers 99, 218-222.

025. Bezerra, A.R., Simoes J., Lee W., et al. (2013) Reversion of a fungal genetic code alteration links proteome instability with genomic and phenotypic diversification. Proc. Nat. Acad. Sci. U.S.A. 110, 11079-11084.

026. Birney, E. and Pritchard, J.K. (2013) Archaic humans: Four makes a party. Nature 505, 32-34.

027. Blue Mountains Eye Study (BMES), et al. (2013) Genome-wide association study of intraocular pressure identifies the GLCCI1/ICA1 region as a glaucoma susceptibility locus. Hum. Mol. Genet. 22, 4653-4660.

028. Bögershausen, N., Bruford, E., Wollnik, B., et al. (2013) Response to Diaz. Clin. Genet. 83, 296.

029. Bögershausen, N., Bruford, E., Wollnik, B., et al. (2013) Skirting the pitfalls: a clear-cut nomenclature for H3K4 methyltransferases. Clin. Genet. 83, 212-214.

030. Borges, D., Perez-Riverol, Y., Nogueira, F.C.S., et al. (2013) Effectively addressing complex proteomic search spaces with peptide spectrum matching. Bioinformatics 29, 1343-1344.

031. Bradnam, K.R., Fass, J.N., Alexandrov, A., et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2, 10.

032. Brazma, A., Cerans, K., Ruklisa, D., et al. (2013) HSM - a hybrid system based approach for modelling intracellular networks. Gene 518, 70-77.

033. Brennecke, P., Anders, S., Kim, J.K., et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093-1095.

034. Büchel, F., Rodriguez, N., Swainston, N., et al. (2013) Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst Biol 7, 116.

035. Budovsky, A., Craig, T., Wang, J., et al. (2013) LongevityMap: A database of human genetic variants associated with longevity. Trends Genet. 29, 559-560.

036. Buljan, M., Chalancon, G., Dunker, A.K., et al. (2013) Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. Curr. Opin. Struct. Biol. 23, 443-450.

037. Burge, S.W., Daub, J., Eberhardt, R., et al. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41, D226-32.

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

039. Caboche, S. (2013) LeView: Automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions. J. Cheminform. 5, 40.

040. CARDIoGRAMplusC4D Consortium, et al. (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. 45, 25-33.

041. Casado, P., Alcolea, M.P., Iorio, F., et al. (2013) Phosphoproteomics data classify hematological cancer cell lines according to tumor type and sensitivity to kinase inhibitors. Genome Biol. 14, R37.

042. Castelo-Branco, G., Amaral, P.P., Engström, P.G., et al. (2013) The non-coding snRNA 7SK controls transcriptional termination, poising, and bidirectionality in embryonic stem cells. Genome Biol. 14, R98.

043. Chambers, J., Davies, M., Gaulton, A., et al. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform 5, 3.

044. Chang, C.-C., Hsiao, Y.-M., Huang, T.-Y., et al. (2013) Noncanonical expression of caudal during early embryogenesis in the pea aphid Acyrthosiphon pisum: Maternal cad-driven posterior development is not conserved. Insect Mol. Biol. 22, 442-455.

045. Chaouiya, C., Bérenguier, D., Keating, S.M., et al. (2013) SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC Syst Biol 7, 135.

046. Chartier, M., Chenard, T., Barker, J. and Najmanovich R. (2013) Kinome render: A stand-alone and web-accessible tool to annotate the human protein kinome tree. PeerJ 2013, e126.

047. Chelliah, V., Laibe, C., Le Novère, N., et al. (2013) BioModels Database: a repository of mathematical models of biological processes. Methods Mol. Biol. 1021, 189-199.

048. Chen, C.-K., Symmons, O., Uslu, V.V., et al. (2013) TRACER: A resource to study the regulatory architecture of the mouse genome. BMC Genomics 14, 215.

049. Chen, C., Li, Z., Huang, H., et al. (2013) A fast peptide match service for UniProt knowledgebase. Bioinformatics 29, 2808-2809.

050. Chen, Z., Lönnberg, T., Lahesmaa, R., et al. (2013) Holistic systems biology approaches to molecular mechanisms of human helper T cell differentiation to functionally distinct subsets. Scand. J. Immunol. 78, 172-180.

051. Clare, S., John, V., Walker, A.W., et al. (2013) Enhanced susceptibility to Citrobacter rodentium infection in microRNA-155-deficient mice. Infect. Immun. 81, 723-732.

052. Coburn, L., Cerone L., Torney, C., et al. (2013) Tactile interactions lead to coherent motion and enhanced chemotaxis of migrating cells. Phys. Biol. 10, 46002.

053. Cochrane, G., Alako, B., Amid, C., et al. (2013) Facing growth in the European Nucleotide Archive. Nucleic Acids Res. 41, D30-5.

054. Coggill, P., Eberhardt, R.Y., Finn, R.D., et al. (2013) Two Pfam protein families characterized by a crystal structure of protein lpg2210 from Legionella pneumophila. BMC Bioinformatics 14, 265.

055. Cokelaer, T., Pultz, D., Harder, L.M., et al. (2013) BioServices: a common Python package to access biological web services programmatically. Bioinformatics (Oxford, England) 29, 3241-3242.

056. Collier, N., Tran, M.-V., Le, H.-Q., et al. (2013) Learning to recognize phenotype candidates in the auto-immune literature using SVM re-ranking. PLoS One 8, e72965.

057. Conte, N., Varela, I., Grove, C., et al. (2013) Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. Leukemia 27, 1820-1825.

058. Croft, D., et al. (2013) Building models using Reactome pathways as templates. Methods Mol. Biol. 1021, 273-283.

059. Croset, S., Overington, J.P., Rebholz-Schuhmann, D., et al. (2013) Brain: biomedical knowledge manipulation. Bioinformatics 29, 1238-1239.

060. Croset, S., Overington, J.P., Rebholz-Schuhmann, D., et al. (2013) The functional therapeutic chemical classification system. Bioinformatics 30, 876-883.

061. Crossman, L.C., Chen, H., Cerdeno-Tarraga, A.-M., et al. (2013) A small predatory core genome in the divergent marine Bacteriovorax marinus SJ and the terrestrial Bdellovibrio bacteriovorus. ISME J. 7, 148-160.

062. Csordas, A., Wang, R., Ríos, D., et al. (2013) From Peptidome to PRIDE: public proteomics data migration at a large scale. Proteomics 13, 1692-1695.

063. Cvejic, A., Haer-Wigman, L., Stephens, J.C., et al. (2013) SMIM1 underlies the Vel blood group and influences red blood cell traits. Nat. Genet. 45, 542-545.

064. Dalquen, D.A. and Dessimoz, C. (2013) Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol. Evol. 5, 1800-1806.

065. Dalquen, D.A., Altenhoff, A.M., Gonnet, G.H. and Dessimoz, C. (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One 8, e56925.

066. Danovi, D., Folarin, A., Gogolok, S., et al. (2013) A high-content small molecule screen identifies sensitivity of glioblastoma stem cells to inhibition of polo-like kinase 1. PLoS One 8, e77053.

067. Davis, M.P.A., van Dongen, S., Abreu-Goodger, C., et al. (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods 63, 41-49.

068. Dawson, D.A., Ball, A.D., Spurgin, L.G., et al. (2013) High-utility conserved avian microsatellite markers enable parentage and population studies across a wide range of species. BMC Genomics 14, 176.

069. De Beer, T.A.P., Laskowski, R.A., Duban, M.-E., et al. (2013) LigSearch: A knowledge-based web server to identify likely ligands for a protein target. Acta Crystallogr. D: Biol. Crystallogr. 69, 2395-2402.

070. de Beer, T.A.P., Laskowski, R.A., Parks, S.L., et al. (2013) Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes Project dataset. PLoS Comp. Biol. 9, e1003382.

071. de Matos, P., Cham, J.A., Cao, H., et al. (2013) The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. BMC Bioinformatics 14, 103.

Publications in 2013

072. del-Toro, N., Dumousseau, M., Orchard, S., et al. (2013) A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. 41, W601-6.

073. Dessimoz, C., Skunca, N. and Thomas, P.D. (2013) CAFA and the Open World of protein function predictions. Trends Genet. 29, 609-610.

074. Dhanoa, B.S., Cogliati, T., Satish, A.G., et al. (2013) Update on the Kelch-like (KLHL) gene family. Hum. Genomics 7, 13.

075. Di Giacomo, M., Comazzetto, S., Saini, H., et al. (2013) Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. Mol. Cell 50, 601-608.

076. Doelken, S.C., Kohler, S., Mungall, C.J., et al. (2013) Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. DMM Disease Models and Mechanisms 6, 358-372.

077. Doğan, T., Karaçalı, B., et al. (2013) Automatic identification of highly conserved family regions and relationships in genome wide datasets including remote protein sequences. PLoS One 8, e75458.

078. Drewe, P., Stegle, O., Hartmann, L., et al. (2013) Accurate detection of differential RNA processing. Nucleic Acids Res. 41, 5189-5198.

079. Dunham, I. (2013) ENCODE-ing the future. Genet. Eng. Biotech. News 33, 38-39.

080. Dutta, S., Dimitropoulos, D., Feng, Z., et al. (2013) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. Biopolymers; doi: 10.1002/bip.22434

081. Dvinge, H., Git, A., Graf, S., et al. (2013) The shaping and functional consequences of the microRNA landscape in breast cancer. Nature 497, 378-382.

082. Eberhardt, R.Y., Bartholdson, S.J., Punta, M., et al. (2013) The SHOCT domain: a widespread domain under-represented in model organisms. PLoS One 8, e57848.

083. Eberhardt, R.Y., Chang, Y., Bateman, A., et al. (2013) Filling out the structural map of the NTF2-like superfamily. BMC Bioinformatics 14, 327.

084. Emwas, A.-H.M., Salek, R.M., Griffin, J.L. and Merzaban, J. (2013) NMR-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. Metabolomics 9, 1048-1072.

085. Engström, P.G., Steijger, T., Sipos, B., et al. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods 10, 1185-1191.

086. Espinoza-Moraga, M., Njuguna, N.M., Mugumbate, G., et al. (2013) In silico comparison of antimycobacterial natural products with known antituberculosis drugs. J. Chem. Inf. Model. 53, 649-660.

087. Evans, J.D., Brown, S.J., Hackett, K.J.J., et al. (2013) The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. 104, 595-600.

088. Fabbro, A., Sucapane, A., Toma, F.M., et al. (2013) Adhesion to carbon nanotube conductive scaffolds forces action-potential appearance in immature rat spinal neurons. PLoS One 8, e73621.

089. Fechner, N., Papadatos, G., Evans, D., et al. (2013) ChEMBLSpace-a graphical explorer of the chemogenomic space covered by the ChEMBL database. Bioinformatics 29, 523-524.

090. Feig, C., Jones, J.O., Kraman, M., et al. (2013) Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. Proc. Nat. Acad. Sci. U.S.A. 110, 20212-20217.

091. Ferreir os-Vidal, I., Carroll, T., Taylor, B., et al. (2013) Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. Blood 121, 1769-1782.

092. Ferreira, J.D., Hastings, J., Couto, F.M., et al. (2013) Exploiting disjointness axioms to improve semantic similarity measures. Bioinformatics 29, 2781-2787.

093. Fidalgo, S., Ivanov, D.K. and Wood, S.H. (2013) Serotonin: From top to bottom. Biogerontology 14, 21-45.

094. Flicek, P. (2013) Evolutionary biology: The handiwork of tinkering. Nature 500, 158-159.

095. Flicek, P., Ahmed, I., Amode, M.R., et al. (2013) Ensembl 2013. Nucleic Acids Res. 41, D48-55.

096. Fonseca, N.A., Morales-Hojas, R., Reis, M., et al. (2013) Drosophila americana as a model species for comparative studies on the molecular basis of phenotypic variation. Genome Biol Evol 5, 661-679.

097. Foster, J.M., Moreno, P., Fabregat, A., et al. (2013) LipidHome: a database of theoretical lipids optimized for high throughput mass spectrometry lipidomics. PLoS One 8, e61951.

098. Funnell, A.P., Wilson, M.D., Ballester, B., et al. (2013) A CpG mutational hotspot in a ONECUT binding site accounts for the prevalent variant of hemophilia B Leyden. Am. J. Hum. Genet. 92, 460-467.

099. Furnham, N., Laskowski, R.A., Thornton, J.M., et al. (2013) Abstracting knowledge from the Protein Data Bank. Biopolymers 99, 183-188.

100. Fusi, N., Lippert, C., Borgwardt, K., et al. (2013) Detecting regulatory gene-environment interactions with unmeasured environmental factors. Bioinformatics 29, 1382-1389.

101. Gagneur, J., Stegle, O., Zhu, C., et al. (2013) Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet. 9, e1003803.

102. Gane, P.J., Chan, A.W., et al. (2013) Molecular fields in ligand discovery. Methods Mol. Biol. 1008, 479-499.

103. Gaudet, P., Munoz-Torres, M., Robinson-Rechavi, M., et al. (2013) Database, the Journal of Biological Databases and Curation, is now the official journal of the International Society for Biocuration. Database (Oxford) 2013, bat077.

104. Gendrel, A.V., Tang, Y.A., Suzuki, M., et al. (2013) Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. Mol. Cell. Biol. 33, 3150-3165.

105. Gene Ontology Consortium, et al. (2013) Gene Ontology annotations and resources. Nucleic Acids Res. 41, D530-5.

106. Ghali, F., Krishna, R., Lukasse, P., et al. (2013) Tools (viewer, library and validator) that facilitate use of the peptide and protein identification standard format, termed mzidentML. Mol. Cell. Proteomics 12, 3026-3035.

107. Gokcumen, O., Tischler, V., Tica, J., et al. (2013) Primate genome architecture influences structural variation mechanisms and functional consequences. Proc. Nat. Acad. Sci. U.S.A. 110, 15764-15769.

108. Goldman, N., Bertone, P., Chen, S., et al. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77-80.

109. Gómez, J., García, L.J., Salazar, G.A., et al. (2013) BioJS: an open source JavaScript framework for biological data visualization. Bioinformatics 29, 1103-1104.

110. Goncalves, E., and Saez-Rodriguez, J. (2013) An interface from Cytoscape to R that provides a user interface to R packages. F1000 Res. 2, 192.

111. Gonçalves, E., Bucher, J., Ryll, A., et al. (2013) Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. Mol Biosyst 9, 1576-1583.

112. Gonçalves, E., van Iersel, M., Saez-Rodriguez, J., et al. (2013) CySBGN: a Cytoscape plug-in to integrate SBGN maps. BMC Bioinformatics 14, 17.

113. Gonzalez-Perez, A., Mustonen, V., Reva, B., et al. (2013) Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods 10, 723-729.

114. Gonzàlez-Porta, M., Frankish, A., Rung, J., et al. (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 14, R70.

115. Gravel, S., Zakharia, F., Moreno-Estrada, A., et al. (2013) Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. 9, e1004023.

116. Gray, K.A., Daugherty, L.C., Gordon, S.M., et al. (2013) Genenames.org: the HGNC resources in 2013. Nucleic Acids Res. 41, D545-52.

117. Grenon, P. and de Bono, B. (2013) Eliciting candidate anatomical routes for protein interactions: A scenario from endocrine physiology. BMC Bioinform. 14, 131.

118. Griss, J., Foster, J.M., Hermjakob, H., et al. (2013) PRIDE Cluster: building a consensus of proteomics data. Nat. Methods 10, 95-96.

119. Groza, T., Oellrich, A. and Collier, N. (2013) using silver and semi-gold standard corpora to compare open named entity recognisers. Proc. 2013 IEEE Int. Conf. Bioinf. Biomed. (IEEE BIBM) 6732541, 481-485.

120. Gutmanas, A., Oldfield, T.J., Patwardhan, A., et al. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. Acta Crystallogr. D Biol. Crystallogr. 69, 710-721.

121. Guziolowski, C., Videla, S., Eduati, F., et al. (2013) Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. Bioinformatics 29, 2320-2326.

122. Hakenberg, J., Nenadic, G., Rebholz-Schuhmann, D. and Kim, J.-D. (2013) Literature mining solutions for life science research. Adv. Bioinform. 2013, 320436.

123. Hall Z., Hernandez, H., Marsh, J.A., et al. (2013) The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure 21, 1325-1337.

124. Hamilton-Williams, E.E., Rainbow, D.B., Cheung, J., et al. (2013) Fine mapping of type 1 diabetes regions Idd9.1 and Idd9.2 reveals genetic complexity. Mamm. Genome 24, 358-375.

125. Hanning, J.E., Saini, H.K., Murray, M.J., et al. (2013) Depletion of HPV16 early genes induces autophagy and senescence in a cervical carcinogenesis model, regardless of viral physical state. J. Pathol. 231, 354-366.

126. Hanning, J.E., Saini, H.K., Murray, M.J., et al. (2013) Lack of correlation between predicted and actual off-target effects of short-interfering RNAs targeting the human papillomavirus type 16 E7 oncogene. Br. J. Cancer 108, 450-460.

127. Hannula-Jouppi, K., Massinen, S., Siljander, T., et al. (2013) Genetic susceptibility to non-necrotizing erysipelas/cellulitis. PLoS One 8, e56225.

128. Hardisty, A., Roberts, D., the Biodiversity Informatics Community, et al. (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecol. 13, 16.

129. Hardman, M., Brooksbank, C., Johnson, C., et al. (2013) LifeTrain: Towards a European framework for continuing professional development in biomedical sciences. Nat. Rev. Drug Discov. 12, 407-408.

130. Harrow, I., Filsell, W., Woollard, P., et al. (2013) Towards Virtual Knowledge Broker services for semantic integration of life science literature and data sources. Drug Discov. Today 18, 428-434.

131. Hartley, D.M., Nelson, N.P., Arthur, R.R., et al. (2013) An overview of internet biosurveillance. Clin. Microbiol. Infection 19, 1006-1013.

132. Hastings, J., Conesa, P., Dekker, A., et al. (2013) Expanding natural product chemistry resources at the EBI. Journal of Cheminformatics 5, P43-P43.

133. Hastings, J., de Matos, P., Dekker, A., et al. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 41, D456-63.

134. Haug, K., Salek, R.M., Conesa, P., et al. (2013) MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. 41, D781-6.

135. Hawkins, R.D., Larjo, A., Tripathi, S.K., et al. (2013) Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. Immunity 38, 1271-1284.

136. Hay, S.I., Battle, K.E., Pigott, D.M., et al. (2013) Global mapping of infectious disease. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 368, 20120250.

137. Hediger, M.A., Clémençon, B., Burrier, R.E., et al. (2013) The ABCs of membrane transporters in health and disease (SLC series): introduction. Mol. Aspects Med. 34, 95-107.

138. Heit, C., Jackson, B.C., McAndrews, M., et al. (2013) Update of the human and mouse SERPIN gene superfamily. Hum. Genomics 7, 22.

139. Hendrickx, P.M., Gutmanas, A., Kleywegt, G.J., et al. (2013) Vivaldi: visualization and validation of biomacromolecular NMR structures from the PDB. Proteins 81, 583-591.

140. Hermjakob, H., Apweiler, R., et al. (2013) Maximising proteomics data for the scientific community European Pharmaceutical Review 9, 23-29.

141. Hickey, G., Paten, B., Earl, D., et al. (2013) HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29, 1341-1342.

142. Hill, D.P., Adams, N., Bada, M., et al. (2013) Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics 14, 513.

143. Hinks, A., Cobb, J., Marion, M.C., et al. (2013) Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat. Genet. 45, 664-669.

144. Hirani, N., Westenberg, M., Gami, M.S., et al. (2013) A simplified counter-selection recombineering protocol for creating fluorescent protein reporter constructs directly from C. elegans fosmid genomic clones. BMC Biotechnol. 13, 1.

145. Hoffman, M.M., Ernst, J., Wilder, S.P., et al. (2013) Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827-841.

146. Howe, K., Clark, M.D., Torroja, C.F., et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496, 498-503.

147. Hu, Y.J., Berndt, S.I., Gustafsson, S., et al. (2013) Meta-analysis of gene-level associations for rare variants based on single-variant statistics. Am. J. Hum. Genet. 93, 236-248.

148. Huang, Y., Li, Y., Burt, D.W., et al. (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nat. Genet. 45, 776-783.

149. Hughes, J.R., Lower, K.M., Dunham, I., et al. (2013) High-resolution analysis of cis-acting regulatory networks at the α-globin locus. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 368, 20120361.

Publications in 2013

150. Hulstaert, N., Reisinger, F., Rameseder, J., et al. (2013) Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. J Proteomics 95, 89-92.

151. Hussain, A., Saraiva, L.R., Ferrero, D.M., et al. (2013) High-affinity olfactory receptor for the death-associated odor cadaverine. Proc. Nat. Acad. Sci. U.S.A. 110, 19579-19584.

152. Hwang, W.C., Bakolitsa, C., Punta, M., et al. (2013) LUD, a new protein domain associated with lactate utilization. BMC Bioinformatics 14, 341.

153. Ilik, I., Quinn, J., Georgiev, P., et al. (2013) Tandem stem-loops in roX RNAs act together to mediate X Chromosome dosage compensation in Drosophila. Mol. Cell 51, 156-173.

154. Ilsley, G.R., Fisher, J., Apweiler, R., et al. (2013) Cellular resolution models for even skipped regulation in the entire Drosophila embryo. Elife 2, e00522.

155. International Multiple Sclerosis Genetics Consortium (IMSGC), et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet. 45, 1353-1360.

156. Iorio, F., Rittman, T., Ge, H., et al. (2013) Transcriptional data: a new gateway to drug repositioning? Drug Discov. Today 18, 350-357.

157. Iorio, F., Saez-Rodriguez, J. and Bernardo, D.D. (2013) Network based elucidation of drug response: From modulators to targets. BMC Sys. Biol. 7, 139.

158. Ison, J., Kalas, M., Jonassen, I., et al. (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325-1332.

159. Ivanov, D.K., Papatheodorou, I., Ziehm, M., et al. (2013) Transcriptional feedback in the insulin signalling pathway modulates ageing in both Caenorhabditis elegans and Drosophila melanogaster. Mol Biosyst 9, 1756-1764.

160. Jimenez, R.C., Albar, J.P., Bhak, J., et al. (2013) iAnn: an event sharing platform for the life sciences. Bioinformatics 29, 1919-1921.

161. Jimenez, R.C., Corpas, M., et al. (2013) Bioinformatics workflows and web services in systems biology made easy for experimentalists. Methods Mol. Biol. 1021, 299-310.

162. Jiménez, R.C., Vizcaíno, J.A., et al. (2013) Proteomics data exchange and storage: the need for common standards and public repositories. Methods Mol. Biol. 1007, 317-333.

163. Jolma, A., Yan, J., Whitington, T., et al. (2013) DNA-binding specificities of human transcription factors. Cell 152, 327-339.

164. Joosten, R.P., Chinea G., Kleywegt G.J., and Vriend G. (2013) Protein three-dimensional structure validation. In: Reedijk, J., Ed. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Waltham, MA: Elsevier. doi: 10.1016/B978-0-12-409547-2.02534-8.

165. Juty, N., Laibe, C., Novère, N.L., et al. (2013) Controlled annotations for systems biology. Methods Mol. Biol. 1021, 227-245.

166. Juty, N., Le Novère, N., Hermjakob, H., et al. (2013) Towards the collaborative curation of the registry underlying Identifiers.org. Database (Oxford) 2013, bat017.

167. Kafkas, S., Kim, J.H., McEntyre, J.R., et al. (2013) Database citation in full text biomedical articles. PLoS One 8, e63184.

168. Keating, S.M., Le Novère, N., et al. (2013) Supporting SBML as a model exchange format in software applications. Methods Mol. Biol. 1021, 201-225.

169. Keller, R., Dorr, A., Tabira, A., et al. (2013) The systems biology simulation core algorithm. BMC Sys. Biol. 7, 55.

170. Khurana, E., Fu, Y., Colonna, V., et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342, 1235587.

171. Kim, J.K., Marioni, J.C., et al. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 14, R7.

172. Kingsley, R.A., Kay, S., Connor, T., et al. (2013) Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted salmonella enterica Serovar Typhimurium pathovar. mBio 4, e00565-13.

173. Kogej, T., Blomberg, N., Greasley, P.J., et al. (2013) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. Drug Discov. Today 18, 1014-1024.

174. Lagerstedt, I., Moore, W.J., Patwardhan, A., et al. (2013) Web-based visualisation and analysis of 3D electron-microscopy data from EMDB and PDB. J. Struct. Biol. 184, 173-181.

175. Lahti, L., Torrente, A., Elo, L.L., et al. (2013) A fully scalable online pre-processing algorithm for short oligonucleotide microarray atlases. Nucleic Acids Res. 41, e110.

176. Laing, R., Kikuchi, T., Martinelli, A., et al. (2013) The genome and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery. Genome Biol. 14, R88.

177. Lappalainen, I., Lopez, J., Skipper, L., et al. (2013) DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 41, D936-41.

178. Lappalainen, T., Sammeth, M., Friedländer, M.R., et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.

179. Laskowski, R.A., Furnham, N., Thornton, J.M., et al. (2013) The Ramachandran plot and protein structure validation. In: Biomolecular Forms and Functions: pp. 62-75; doi: 10.1142/9789814449144_0005.

180. Lawler, K., Hammond-Kosack, K., Brazma, A., et al. (2013) Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus Fusarium graminearum. BMC Syst Biol 7, 52.

181. Lawson C., Patwardhan A., Pintilie G., et al. (2013) EMDataBank: Unified Data Resource for 3DEM". Biophys. J. 104, 351a.

182. Le Novere, N. and Endler, L. (2013) using chemical kinetics to model biochemical pathways. Methods Mol. Biol. 1021, 147-167.

183. Leprevost, F.V., Lima, D.B., Crestani, J., et al. (2013) Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for proteomics. J. Proteomics 89, 176-182.

184. Lewis, T.E., Sillitoe, I., Andreeva, A., et al. (2013) Genome3D: a United Kingdom collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic Acids Res. 41, D499-507.

185. Li, C., Jimeno-Yepes, A., Arregui, M., et al. (2013) PCorral--interactive mining of protein interactions from MEDLINE. Database (Oxford) 2013, bat030.

186. Li, C., Liakata, M., Rebholz-Schuhmann, D., et al. (2013) Biological network extraction from scientific literature: state of the art and challenges. Brief. Bioinform.; doi: 10.1093/bib/bbt006.

187. Li, J.W., Bolser, D., Manske, M., et al. (2013) The NGS WikiBook: a dynamic collaborative online training effort with long-term sustainability. Brief. Bioinform. 14, 548-555.

188. Li, W., Kondratowicz, B., McWilliam, H., et al. (2013) The annotation-enriched non-redundant patent sequence databases. Database (Oxford) 2013, bat005.

189. Li, X., Wilmanns, M., Thornton, J., et al. (2013) Elucidating human phosphatase-substrate networks. Sci Signal 6, rs10.

190. Liebisch, G., Vizcaíno, J.A., Köfeler, H., et al. (2013) Shorthand notation for lipid structures derived from mass spectrometry. J. Lipid Res. 54, 1523-1530.

191. Listgarten, J., Stegle, O., Morris, Q., et al. (2013) Personalized medicine: from genotypes and molecular phenotypes towards therapy-session introduction. Pacific Symposium on Biocomputing 19, 224–228.

192. Liu, S., Yamada, M., Collier, N., et al. (2013) Change-point detection in time-series data by relative density-ratio estimation. Neural Netw 43, 72-83.

193. Lönnberg, T., Chen, Z., Lahesmaa, R., et al. (2013) From a gene-centric to whole-proteome view of differentiation of T helper cell subsets. Brief Funct Genomics 12, 471-482.

194. Lonnberg, T., Yetukuri, L., Seppanen-Laakso, T., et al. (2013) T-cell activation induces selective changes of cellular lipidome. Front Biosci (Elite Ed) 5, 558-573.

195. Lotz, C., Lin, A.J., Black, C.M., et al. (2013) Characterization, design, and function of the mitochondrial proteome: from organs to organisms. J. Proteome Res.; doi: 10.1021/pr400539j.

196. Louis, A., Muffato, M., Roest Crollius, H., et al. (2013) Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 41, D700-5.

197. Loureiro, T., Camacho, R., Vieira, J. and Fonseca, N.A. (2013) Boosting the detection of transposable elements using machine learning. Adv. Intel. Sys. Computing 222, 85-91.

198. Loureiro, T., Camacho, R., Vieira, J., et al. (2013) Improving the performance of Transposable Elements detection tools. J. Integr. Bioinform. 10, 231.

199. Lupini, L., Bassi, C., Ferracin, M., et al. (2013) miR-221 affects multiple cancer pathways by modulating the level of hundreds messenger RNAs. Front Genet 4, 64.

200. MacNamara, A., Henriques, D., Saez-Rodriguez, J., et al. (2013) Modeling signaling networks with different formalisms: a preview. Methods Mol. Biol. 1021, 89-105.

201. Majewski, I.J., Mittempergher, L., Davidson, N.M., et al. (2013) Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. J. Pathol. 230, 270-276.

202. Malone, J. and Stevens, R. (2013) Measuring the level of activity in community built bio-ontologies. J. Biomedical Inform. 46, 5-14.

203. Malouf, C. Bartonicek, N., Bacon, W., et al. (2013) Deciphering the role of micrornas in early stages of haematopoiesis. Exp. Hematol. 41, S38–S38.

204. Mancini, C., Roncaglia, P., Brussino, A., et al. (2013) Genome-wide expression profiling and functional characterization of SCA28 lymphoblastoid cell lines reveal impairment in cell growth and activation of apoptotic pathways. BMC Med Genomics 6, 22.

205. Maree, F.F., Blignaut, B., de Beer, T.A.P. and Rieder, E. (2013) Analysis of SAT type foot-and-mouth disease virus capsid proteins and the identification of putative amino acid residues affecting virus stability. PLoS One 8, e61612.

206. Marsh, J. A., Hernández H., Hall Z., et al. (2013) Structural and evolutionary dynamics facilitate ordered protein complex assembly. Biophys. J. 104, 391a.

207. Marsh, J.A. (2013) Buried and accessible surface area control intrinsic protein flexibility. J. Mol. Biol. 425, 3250-3263.

208. Marsh, J.A., Hernandez, H., Hall, Z., et al. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153, 461-470.

209. Martello, G., Bertone, P., Smith, A., et al. (2013) Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. EMBO J. 32, 2561-2574.

210. Martin, M.J., Clare, S., Goulding, D., et al. (2013) The agr locus regulates virulence and colonization genes in clostridium difficile 027. J. Bacteriol. 195, 3672-3681.

211. Martincorena, I. and Luscombe, N.M. (2013) Non-random mutation: The evolution of targeted hypermutation and hypomutation. BioEssays 35, 123-130.

212. Martínez-Jiménez, F., Papadatos, G., Yang, L., et al. (2013) Target prediction for an open access set of compounds active against Mycobacterium tuberculosis. PLoS Comput. Biol. 9, e1003253.

213. Mattioni, M. and Le Novere, N. (2013) Integration of biochemical and electrical signaling-multiscale model of the medium spiny neuron of the striatum. PLoS One 8, e66811.

214. May, J.W., James, A.G. and Steinbeck, C. (2013) Metingear: A development environment for annotating genome-scale metabolic models. Bioinformatics 29, 2213-2215.

215. Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., et al. (2013) The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. Database (Oxford) 2013, bat009.

216. McWilliam, H., Li, W., Uludag, M., et al. (2013) Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41, W597-600.

217. Meehan, T.F., Vasilevsky, N.A., Mungall, C.J., et al. (2013) Ontology based molecular signatures for immune cell types via gene expression analysis. BMC Bioinform. 14, 263.

218. Mellars, P., Gori, K.C., Carr, M., et al. (2013) Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc. Natl. Acad. Sci. U.S.A. 110, 10699-10704.

219. Menden, M.P., Iorio, F., Garnett, M., et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One 8, e61318.

220. Millán, P.P., et al. (2013) Visualization and analysis of biological networks. Methods Mol. Biol. 1021, 63-88.

221. Milne, J.L.S., Borgnia, M.J., Bartesaghi, A., et al. (2013) Cryo-electron microscopy - A primer for the non-microscopist. FEBS J. 280, 28-45.

222. Mistry, J., Coggill, P., Eberhardt, R.Y., et al. (2013) The challenge of increasing Pfam coverage of the human proteome. Database: The Journal of Biological Databases and Curation 2013, bat023.

223. Mistry, J., Finn, R.D., Eddy, S.R., et al. (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121.

224. Mistry, J., Kloppmann, E., Rost, B., et al. (2013) An estimated 5% of new protein structures solved today represent a new Pfam family. Acta Crystallogr. D Biol. Crystallogr. 69, 2186-2193.

225. Montelione, G.T., Nilges, M., Bax, A., et al. (2013) Recommendations of the wwPDB NMR Validation Task Force. Structure 21, 1563-1570.

226. Montgomery, S.B., Goode, D.L., Kvikstad, E., et al. (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 23, 749-761.

227. Moretti, R., Fleishman, S.J., Agius, R., et al. (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. Proteins 81, 1980-1987.

228. Morris, M.K., Melas, I., Saez-Rodriguez, J., et al. (2013) Construction of cell type-specific logic models of signaling networks using CellNOpt. Methods Mol. Biol. 930, 179-214.

229. Morris, Q., Brenner, S.E., Listgarten, J., et al. (2013) The future of genome-based medicine. Pac Symp Biocomput 456-458.

Publications in 2013

230. Moulos, P., Klein, J., Jupp, S., et al. (2013) The KUPNetViz: a biological network viewer for multiple -omics datasets in kidney diseases. BMC Bioinformatics 14, 235.

231. Mugumbate, G., Newton, A.S., Rosenthal, P.J., et al. (2013) Novel anti-plasmodial hits identified by virtual screening of the ZINC database. J. Comput. Aided Mol. Des. 27, 859-871.

232. Murray, M.J., Saini, H.K., Siegler, C.A., et al. (2013) LIN28 Expression in malignant germ cell tumors downregulates let-7 and increases oncogene levels. Cancer Res. 73, 4872-4884.

233. Mutowo-Meullenet, P., Huntley, R.P., Dimmer, E.C., et al. (2013) use of Gene Ontology Annotation to understand the peroxisome proteome in humans. Database (Oxford) 2013, bas062.

234. Nakamura, Y., Cochrane, G., Karsch-Mizrachi, I., et al. (2013) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 41, D21-4.

235. Nayduch, D., Cohnstaedt, L.W., Saski, C., et al. (2013) Studying Culicoides vectors of BTV in the post-genomic era: resources, bottlenecks to progress and future directions. Virus Res. 182, 43-49.

236. Neafsey, D.E., Christophides, G.K., Collins, F.H., et al. (2013) The evolution of the Anopheles 16 genomes project. G3 (Bethesda) 3, 1191-1194.

237. Nishi, H., Fong, J.H., Chang, C., et al. (2013) Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: Analysis of human protein complexes. Mol. BioSystems 9, 1620-1626.

238. Ochoa, D., García-Gutiérrez, P., Juan, D., et al. (2013) Incorporating information on predicted solvent accessibility to the co-evolution-based study of protein interactions. Mol Biosyst 9, 70-76.

239. Oetting, W.S., Robinson, P.N., Greenblatt, M.S., et al. (2013) Getting ready for the Human Phenome Project: The 2012 Forum of the Human Variome Project. Hum. Mutat. 34, 661-666.

240. Orchard, S., Binz, P.A., Jones, A.R., et al. (2013) Preparing to work with big data in proteomics - a report on the HUPO-PSI Spring Workshop: April 15-17, 2013, Liverpool, United Kingdom. Proteomics 13, 2931-2937.

241. Pacini, C., Iorio, F., Gonçalves, E., et al. (2013) DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. Bioinformatics 29, 132-134.

242. Pande, S., Merker, H., Bohl, K., et al. (2013) Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. ISME J.; doi: 10.1038/ismej.2013.211.

243. Pedruzzi, I., Rivoire, C., Auchincloss, A.H., et al. (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 41, D584-9.

244. Perez-Riverol, Y., Hermjakob, H., Kohlbacher, O., et al. (2013) Computational proteomics pitfalls and challenges: HavanaBioinfo 2012 workshop report. J Proteomics 87, 134-138.

245. Perez-Riverol, Y., Sánchez, A., Noda, J., et al. (2013) HI-bone: a scoring system for identifying phenylisothiocyanate-derivatized peptides based on precursor mass and high intensity fragment ions. Anal. Chem. 85, 3515-3520.

246. Petrov, A.I., Zirbel, C.L., Leontis, N.B., et al. (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA 19, 1327-1340.

247. Pettit, J.B., Marioni, J.C., et al. (2013) bioWeb3D: an online webGL 3D data visualisation tool. BMC Bioinformatics 14, 185.

248. Pickard, D., Kingsley, R.A., Hale, C., et al. (2013) A genomewide mutagenesis screen identifies multiple genes contributing to Vi capsular expression in Salmonella enterica serovar typhi. J. Bacteriol. 195, 1320-1326.

249. Radivojac, P., Clark, W.T., Oron, T.R., et al. (2013) A large-scale evaluation of computational protein function prediction. Nat. Methods 10, 221-227.

250. Rahrig, R.R., Petrov, A.I., Leontis, N.B., et al. (2013) R3D Align web server for global nucleotide to nucleotide alignments of RNA 3D structures. Nucleic Acids Res. 41, W15-21.

251. Rakitsch, B., Lipper, C,. Borgwardt, K. and Stegle, O. (2013) It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In: Burges, C., et al. (Eds.) Advances in Neural Information Processing Systems 26 (NIPS 2013), pp. 1466-1474.

252. Rakitsch, B., Lippert, C., Stegle, O., et al. (2013) A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 29, 206-214.

253. Rawlings, N.D. (2013) Evolution of the thermopsin peptidase family (A5). PLoS One 8, e78998.

254. Rawlings, N.D. (2013) Identification and prioritization of novel uncharacterized peptidases for biochemical characterization. Database 2013, bat022.

255. Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., et al. (2013) Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. Lecture Notes in Computer Science 8138 LNCS, 353-367.

256. Rebholz-Schuhmann, D., Grabmuller, C., Kavaliauskas, S., et al. (2013) A case study: semantic integration of gene-disease associations for type 2 diabetes mellitus from literature and biomedical data resources. Drug Discov. Today; doi: 10.1016/j.drudis.2013.10.024.

257. Rebholz-Schuhmann, D., Kafkas, S., Kim, J.H., et al. (2013) Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. J Biomed Semantics 4, 28.

258. Rebholz-Schuhmann, D., Kafkas, S., Kim, J.H., et al. (2013) Monitoring named entity recognition: the League Table. J Biomed Semantics 4, 19.

259. Rebholz-Schuhmann, D., Kim, J.H., Yan, Y., et al. (2013) Evaluation and cross-comparison of lexical entities of biological interest (LexEBI). PLoS One 8, e75185.

260. Rezola, A., Pey, J., de Figueiredo, L.F., et al. (2013) Selection of human tissue-specific elementary flux modes using gene expression data. Bioinformatics 29, 2009-2016.

261. Ripke, S., O'Dushlaine, C., Chambert, K., et al. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. 45, 1150-1159.

262. Robinson, J., Halliwell, J.A., McWilliam, H., et al. (2013) IPD--the Immuno Polymorphism Database. Nucleic Acids Res. 41, D1234-40.

263. Robinson, J., Halliwell, J.A., McWilliam, H., et al. (2013) The IMGT/HLA database. Nucleic Acids Res. 41, D1222-7.

264. Roncaglia, P., Martone, M.E., Hill, D.P., et al. (2013) The Gene Ontology (GO) Cellular Component Ontology: integration with SAO (Subcellular Anatomy Ontology) and other recent developments. J Biomed Semantics 4, 20.

265. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., et al. (2013) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 41, D56-63.

266. Rung, J., Brazma, A., et al. (2013) Reuse of public genome-wide gene expression data. Nat. Rev. Genet. 14, 89-99.

267. Rustici, G., Kolesnikov, N., Brandizi, M., et al. (2013) ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987-D990.

268. Salek, R.M., Haug, K., Conesa, P., et al. (2013) The MetaboLights repository: curation challenges in metabolomics. Database (Oxford) 2013, bat029.

269. Salek, R.M., Haug, K., Steinbeck, C., et al. (2013) Dissemination of metabolomics results: role of MetaboLights and COSMOS. Gigascience 2, 8.

270. Sanges, R., Hadzhiev, Y., Gueroult-Bellone, M., et al. (2013) Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. Nucleic Acids Res. 41, 3600-3618.

271. Sazanov, L.A., Baradaran, R., Efremov, R.G., et al. (2013) A long road towards the structure of respiratory complex I, a giant molecular proton pump. Biochem. Soc. Trans. 41, 1265-1271.

272. Schauer, T., Schwalie, P.C., Handley, A., et al. (2013) CAST-ChIP maps cell-type-specific chromatin states in the Drosophila central nervous system. Cell Rep 5, 271-282.

273. Schneider, L., Pellegatta, S., Favaro, R., et al. (2013) DNA damage in mammalian neural stem cells leads to astrocytic differentiation mediated by BMP2 signaling through JAK-STAT. Stem Cell Reports 1, 123-138.

274. Schwalie, P.C., Ward, M.C., Cain, C.E., et al. (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. Genome Biol. 14, R148.

275. Seal, R.L., Wright, M.W., Gray, K.A., et al. (2013) Vive la difference: naming structural variants in the human reference genome. Hum. Genomics 7, 12.

276. Seitan, V.C., Faure, A.J., Zhan, Y., et al. (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. 23, 2066-2077.

277. Sillitoe, I., Cuff, A.L., Dessailly, B.H., et al. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. 41, D490-D498.

278. Smith, J.J., Kuraku, S., Holt, C., et al. (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat. Genet. 45, 415-421, 421e1-2.

279. Soler, J.J., Martín-Gálvez, D., de Neve, L., et al. (2013) Brood parasitism correlates with the strength of spatial autocorrelation of life history and defensive traits in Magpies. Ecology 94, 1338-1346.

280. Soler, M., Ruiz-Castellano, C., Carra, L.G., et al. (2013) Do first-time breeding females imprint on their own eggs? Proc. Biol. Sci. 280, 20122518.

281. Somel, M., Wilson Sayres, M. A., Jordan, G., et al. (2013) A scan for human-specific relaxation of negative selection reveals unexpected polymorphism in proteasome genes. Mol. Biol. Evol. 30, 1808-1815.

282. Sottoriva, A., Spiteri, I., Piccirillo, S.G., et al. (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc. Natl. Acad. Sci. U.S.A. 110, 4009-4014.

283. Stefflova, K., Thybert, D., Wilson, M.D., et al. (2013) Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 154, 530-540.

284. Steijger, T., Abril, J.F., Engström, P.G., et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10, 1177-1184.

285. Stricker, S.H., Feber, A., Engström, P.G., et al. (2013) Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. Genes Dev. 27, 654-669.

286. Stubbington, M.J., Corcoran, A.E., et al. (2013) Non-coding transcription and large-scale nuclear organisation of immunoglobulin recombination. Curr. Opin. Genet. Dev. 23, 81-88.

287. Sunnaker, M., Busetto, A.G., Numminen, E., et al. (2013) Approximate Bayesian computation. PLoS Comp. Biol. 9, e1002803.

288. Suter, B., Fontaine, J.F., Yildirimman, R., et al. (2013) Development and application of a DNA microarray-based yeast two-hybrid system. Nucleic Acids Res. 41, 1496-1507.

289. Swaney, D.L., Beltrao, P., Starita, L., et al. (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. Nat. Methods 10, 676-682.

290. Tachmazidou, I., Dedoussis, G., Southam, L., et al. (2013) A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nat. Comm. 4, 2872.

291. Teichmann, S. A. (2013) Gene expression genomics. Biophysical J. 104, 534a.

292. Teichmann, S.A. (2013) Immunology meets genomics. Brief. Funct. Genomics 12, 469-470.

293. Thiele, I., Swainston, N., Fleming, R.M., et al. (2013) A community-driven global reconstruction of human metabolism. Nat. Biotechnol. 31, 419-425.

294. Tiikkainen, P., Bellis, L., Light, Y. and Franke, L. (2013) Estimating error rates in bioactivity databases. J. Chem. Inf. Modeling 53, 2499-2505.

295. Torrente, A., Lopez-Pintado, S. and Romo, J. (2013) DepthTools: An R package for a robust analysis of gene expression data. BMC Bioinform. 14, 237.

296. Touloumis, A., Agresti, A. and Kateri, M. (2013) GEE for multinomial responses using a local odds ratios parameterization. Biometrics 69, 633-640.

297. Traylor, M., Bevan, S., Rothwell, P.M., et al. (2013) using phenotypic heterogeneity to increase the power of genome-wide association studies: application to age at onset of ischaemic stroke subphenotypes. Genet. Epidemiol. 37, 495-503.

298. Trewhella, J., Hendrickson, W.A., Kleywegt, G.J., et al. (2013) Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure 21, 875-881.

299. Tripathi, S., Christie, K.R., Balakrishnan, R., et al. (2013) Gene Ontology annotation of sequence-specific DNA binding transcription factors: Setting the stage for a large-scale curation effort. Database 2013, bat062.

300. Tudose, I., Hastings, J., Muthukrishnan, V., et al. (2013) OntoQuery: easy-to-use web-based OWL querying. Bioinformatics 29, 2955-2957.

301. Tukulula, M., Njoroge, M., Mugumbate, G.C., et al. (2013) Tetrazole-based deoxyamodiaquines: synthesis, ADME/PK profiling and pharmacological evaluation as potential antimalarial agents. Bioorg. Med. Chem. 21, 4904-4913.

302. UniProt Consortium, et al. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 41, D43-D47.

303. Ur-Rehman, S., Gao, Q., Mitsopoulos, C., et al. (2013) ROCK: a resource for integrative breast cancer data analysis. Breast Cancer Res. Treat. 139, 907-921.

304. Van Roey, K., Orchard, S., Kerrien, S., et al. (2013) Capturing cooperative interactions with the PSI-MI format. Database (Oxford) 2013, bat066.

305. van Westen, G.J., Hendriks, A., Wegner, J.K., et al. (2013) Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. PLoS Comput. Biol. 9, e1002899.

306. van Westen, G.J., Overington, J.P., et al. (2013) A ligand's-eye view of protein similarity. Nat. Methods 10, 116-117.

307. van Westen, G.J., Swier, R.F., Wegner, J.K., et al. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J. Cheminform. 5, 41.

308. van Westen, G.J., Swier, R.F., Cortes-Ciriano, I., et al. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. J Cheminform 5, 42.

309. Vandermarliere, E., Mueller, M. and Martens, L. (2013) Getting intimate with trypsin, the leading protease in proteomics. Mass Spec. Rev. 32, 453-465.

310. Velankar, S., Dana, J.M., Jacobsen, J., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 41, D483-9.

311. Vera, R., Perez-Riverol, Y., Perez, S., et al. (2013) JBioWH: An open-source Java framework for bioinformatics data integration. Database 2013, bat051.

312. Veuthey, A.L., Bridge, A., Gobeill, J., et al. (2013) Application of text-mining for updating protein post-translational modification annotation in UniProtKB. BMC Bioinformatics 14, 104.

313. Via, A., Blicher, T., Bongcam-Rudloff, E., et al. (2013) Best practices in bioinformatics training for life scientists. Brief. Bioinformatics 14, 528-537.

314. Vizcaíno, J.A., Côté, R.G., Csordas, A., et al. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 41, D1063-D1069.

315. Vuister, G.W., Fogh, R.H., Hendrickx, P.M., et al. (2013) An overview of tools for the validation of protein NMR structures. J. Biomol. NMR.; doi: 10.1007/s10858-013-9750-x.

316. Waller-Evans, H., Hue, C., Fearnside, J., et al. (2013) Nutrigenomics of high fat diet induced obesity in mice suggests relationships between susceptibility to fatty liver disease and the proteasome. PLoS One 8, e82825.

317. Walzer, M., Qi, D., Mayer, G., et al. (2013) The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol. Cell Proteomics 12, 2332-2340.

318. Wang, J., Zhuang, J., Iyer, S., et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 41, D171-6.

319. Wang, Z., Pascual-Anaya, J., Zadissa, A., et al. (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. 45, 701-706.

320. Ward, M.C., Wilson, M.D., Barbosa-Morais, N.L., et al. (2013) Latent regulatory potential of human-specific repetitive elements. Mol. Cell 49, 262-272.

321. Weirauch, M.T., Cote, A., Norel, R., et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. 31, 126-134.

322. Whitaker, H.C., Patel, D., Howat, W.J., et al. (2013) Peroxiredoxin-3 is overexpressed in prostate cancer and promotes cancer cell survival by protecting cells from oxidative stress. Br. J. Cancer 109, 983-993.

323. White, J.K., Gerdin, A.K., Karp, N.A., et al. (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell 154, 452-464.

324. Willighagen, E.L., Waagmeester, A., Spjuth, O., et al. (2013) The ChEMBL database as linked open data. J Cheminform 5, 23.

325. Wood, L. and Gebhardt, P. (2013) Bioinformatics goes to school-new avenues for teaching contemporary biology. PLoS Comp. Biol. 9, e1003089.

326. Wright, C.F., Middleton, A., Burton, H., et al. (2013) Policy challenges of clinical genome sequencing. BMJ 347, f6845.

327. Zarnack, K., Konig, J., Tajnik, M., et al. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell 152, 453-466.

328. Zhang, X., Perica, T. and Teichmann, S.A. (2013) Evolution of protein structures and interactions from the perspective of residue contact networks. Curr. Opin. Struct. Biol. 23, 954-963.

329. Ziehm, M., Piper, M.D., Thornton, J.M., et al. (2013) Analysing variation in Drosophila aging across independent experimental studies: a meta-analysis of survival data. Aging Cell 12, 917-922.

330. Ziehm, M., Thornton, J.M., et al. (2013) Unlocking the potential of survival data for model organisms through a new database and online analysis platform: SurvCurv. Aging Cell 12, 910-916.

331. Zong, N.C., Li, H., Li, H., et al. (2013) Integration of cardiac proteome biology and medicine by a specialized knowledgebase. Circ. Res. 113, 1043-1053.

Organisation of EMBL-EBI Leadership

RESEARCH GROUPS

Ewan Birney group

Pedro Beltrao group

Paul Bertone group

Anton Enright group

Nick Goldman group

John Marioni group

Julio Saez-Rodriguez

Oliver Stegle group

Sarah Teichmann group

Janet Thornton group

Proteomics services
Henning Hermjakob

Admininstration
Mark Green

External relations
Lindsey Crosswell

Janet Thornton
Director

Ewan Birney
Associate director

Rolf Apweiler
Associate director

SERVICE TEAMS

Functional genomics
Alvis Brazma

Functional genomics development
Ugis Sarkans

Protein Databank in Europe (PDBe)
Gerard Kleywegt

PDBe Databases & services
Tom Oldfield

PDBe content & intergration
Sameer Velankar

Protein resources
Alex Bateman

UniProt content
Claire O'Donovan

UniProt development
Maria-Jesus Martin

InterPro
Sarah Hunter

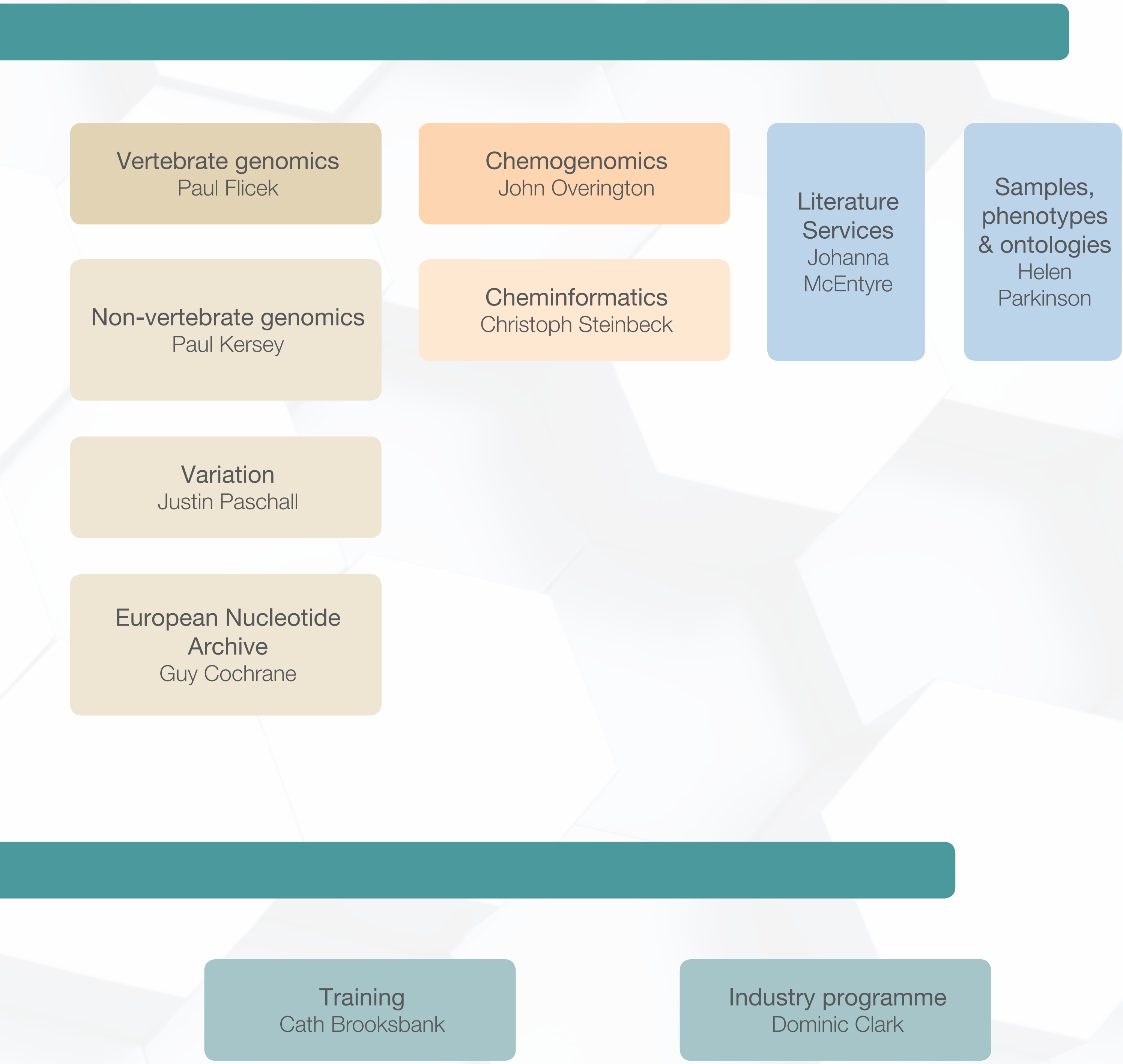
Protein families
Rob Finn (2014)

EXTERNAL-FACING ACTIVITIES

Web development
Brendan Vaughan

Web production
Rodrigo Lopez

Systems and networking
Petteri Jokinen



Vertebrate genomics
Paul Flicek

Non-vertebrate genomics
Paul Kersey

Variation
Justin Paschall

European Nucleotide Archive
Guy Cochrane

Chemogenomics
John Overington

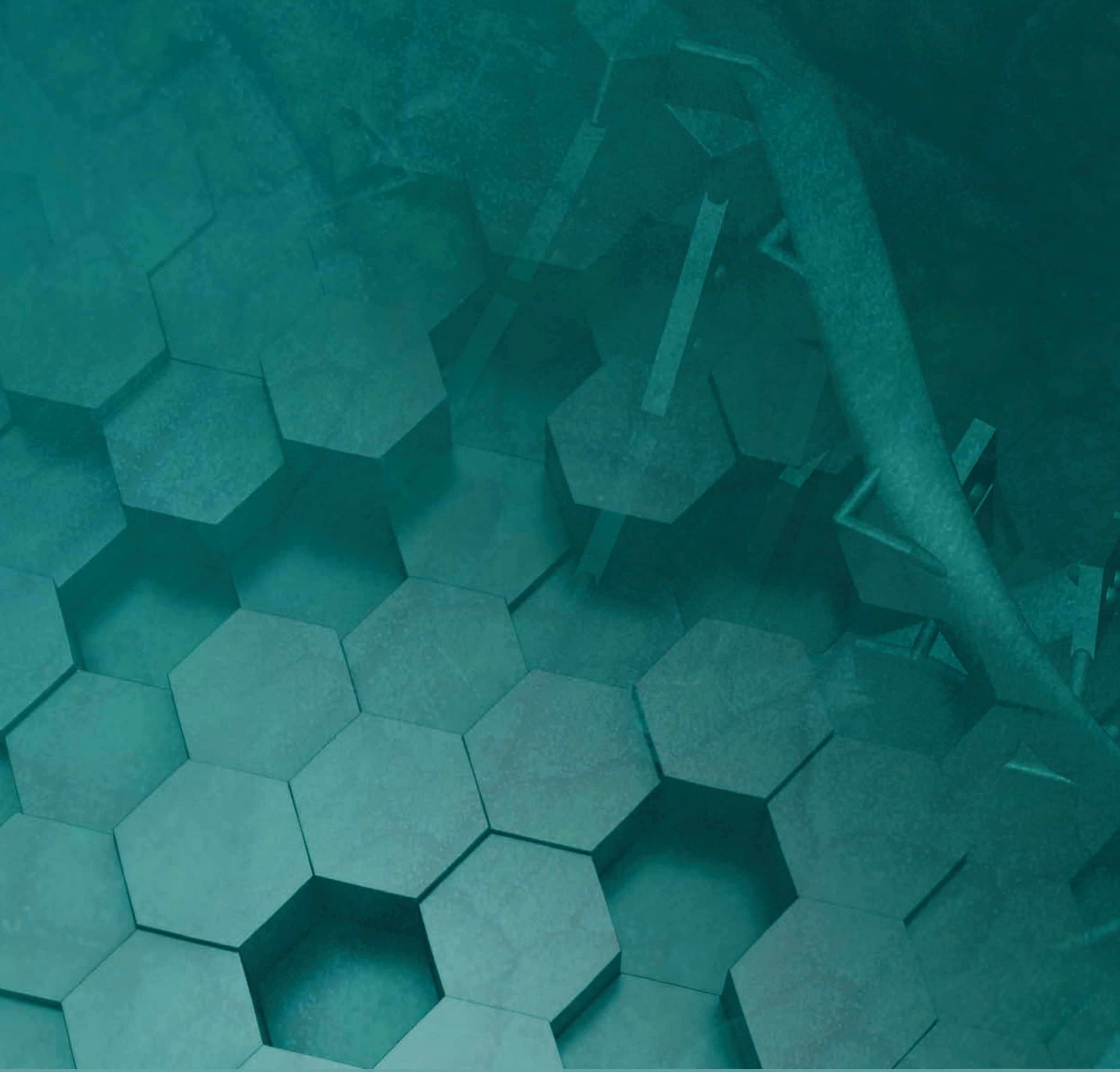
Cheminformatics
Christoph Steinbeck

Literature Services
Johanna McEntyre

Samples, phenotypes & ontologies
Helen Parkinson

Training
Cath Brooksbank

Industry programme
Dominic Clark



EMBL-European Bioinformatics Institute
Wellcome Trust 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

@emblebi
/EMBLEBI
/EMBLEBI

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

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

www.ebi.ac.uk/about/brochures