European Bioinformatics Institute

Annual Scientific Report 2011

EMBL-EBI

EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton
Cambridge CB10 1SD
United Kingdom
Tel. +44 (0)1223 494 444, Fax +44 (0)1223 494 468
www.ebi.ac.uk

EMBL Heidelberg
Meyerhofstraße 1
69117 Heidelberg
Germany
Tel. +49 (0)6221 3870, Fax +49 (0)6221 387 8306
www.embl.org
info@embl.org

EMBL Grenoble
6, rue Jules Horowitz, BP181
38042 Grenoble, Cedex 9
France
Tel. +33 (0)476 20 7269, Fax +33 (0)476 20 2199

EMBL Hamburg
c/o DESY
Notkestraße 85
22603 Hamburg
Germany
Tel. +49 (0)4089 902 110, Fax +49 (0)4089 902 149

EMBL Monterotondo
Adriano Buzzati-Traverso Campus
Via Ramarini, 32
00015 Monterotondo (Rome)
Italy
Tel. +39 (0)6900 91402, Fax +39 (0)6900 91406

All texts written by EBI-EMBL Group and Team Leaders. This publication was produced by the EBI's Outreach and Training Programme.

Contents

Introduction

Services

Research

Support

Facts and Figures

Foreword

Janet Thornton

Director

Graham Cameron

Associate Director

Welcome to EMBL-EBI's 2011 Annual Scientific Report!

2011 has been a challenging and good year.

We have continued to provide a safe haven for biological data, adding value by careful curation and annotation and making data easily accessible to all scientists worldwide. We launched several new resources with powerful tools for analysis, for example 'Metabolights' for metabolomic data and a portal for metagenomics data. Our new Enzyme Portal now integrates extensive information on enzymes from several EMBL-EBI resources. The scientific literature remains the greatest repository of information and knowledge and, in the spirit of open access, EMBL-EBI is now leading a collaboration to provide UK PubMedCentral, a free online literature resource for life science researchers.

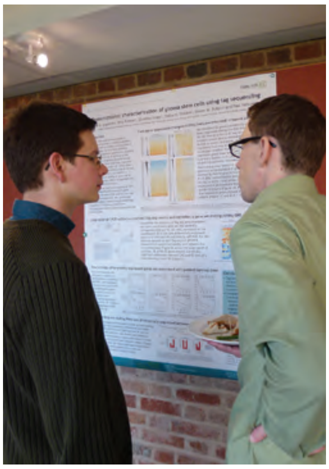
Biological information underpins discovery. Our research scientists are deeply involved in many collaborative projects, developing new algorithms and tools to extract robust, validated knowledge from new and existing data. Our basic research addresses a wide range of biological questions, from the evolution of species to understanding synaptic plasticity and cell differentiation. One study of note this year combined genetic and genomic approaches to reveal how one protein helps protect the genome during duplication. Another shed light on how a large protein complex contributes to stem cell differentiation during embryonic development. Through the development of new models and algorithms, other efforts have elucidated the role of a kinase involved in memory. Our researchers continue to generate powerful resources for exploring biology, for example by helping users create logic-based models of cellular signalling pathways, capture the words used to describe biological processes or track the evolution of enzymes and their ability to evolve new functions.

The flood of data continues to rise on all fronts, and we have to run ever faster just to stand still. We have developed a novel way to compress and store the vast quantities of data pouring out of high-throughput genomic experiments every day. To optimise our delivery of services to the community, we have continued to transfer our resources to data centres in London: a demanding task that has progressed well.

Our resources are increasingly in demand: recently, our website received over 7 million web hits in one day and almost 4 million jobs were run in one month. This reflects an ever-changing landscape of users throughout the world, and an associated rise in the need for bioinformatics training. More and more of our users are translational scientists working in medicine and industry. To supplement our busy hands-on training and roadshow programmes, we launched Train online, a new resource that helps users learn in their own time and at their own pace.

We have continued to grow, and have strengthened our leadership by adding two new Team Leaders to the Protein Data Bank in Europe (PDBe) team; we expect to bolster leadership in our sequence variation resources in 2012. Our new External Relations programme, which got underway in September 2011, has already made substantial progress in building relationships and communications with key stakeholders.

We continue to coordinate the preparatory phase of ELIXIR, the nascent pan-European research infrastructure for biological information in Europe. It is our hope that ELIXIR will provide the facilities necessary for life science researchers to access and



exploit biological data in the coming decades. Following the signing of a Memorandum of Understanding by EMBL and 11 nations, ELIXIR's member states can now work together with EMBL to develop the International Consortium Agreement necessary for the construction of ELIXIR.

From a funding perspective EMBL-EBI's scientists have been very successful in raising funds (€23 million in 2011) for services and research through competitive grants – almost always in collaboration with others in the UK, mainland Europe or internationally. We thank our major funders (the EU, the NIH, the Wellcome Trust and the UK Research Councils). Clearly, without these funds we would not be able to deliver the high-quality services and research for which we are known.

We were delighted when the UK government committed capital funding of €90 million to support EMBL-EBI and the establishment of the ELIXIR infrastructure hub. Over the next eight years this will fund computing equipment for the delivery of all our services, including ELIXIR. In addition it will fund a new building on the Wellcome Trust Genome Campus that will provide accommodation for staff; an Industry and Innovation Suite to promote the use biological information in applications in medicine, biotechnology and the environment; and two new training rooms that will allow us to train more users.

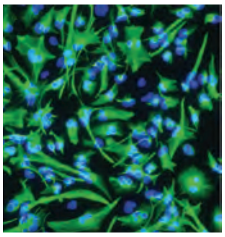
We are very happy to welcome the many visitors who come to EMBL-EBI to work with our scientists, use the data resources, attend training courses and participate in the stimulating scientific environment that characterises our campus. With the emergence of ELIXIR, we expect this number of visitors to increase as we work together to build a distributed bioinformatics infrastructure.

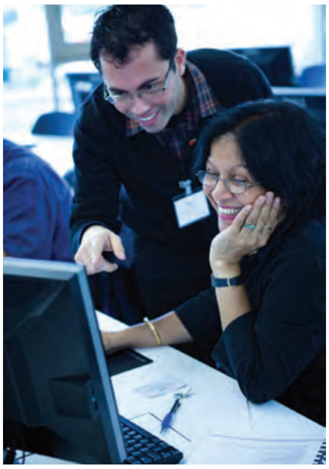
Recent announcements on the next generation of sequencing machines promise to usher in a completely new scale of data production and analysis. The impact of these technologies on medicine and the environment is only just beginning, and requires even closer collaboration with our colleagues in the clinic and in industry.

Such close international collaborations are not new to EMBL-EBI. Indeed, all our efforts rely on such interactions. The deposition of new data, the daily exchange of information between data resources, the joint development of software tools, the sharing of curation tasks and the challenges of collaborative research have allowed us to build an extensive network of colleagues. We look forward to strengthening these links in the coming year and, as always, to creating new collaborations.

Janet Thornton
Director

Graham Cameron
Associate Director





Major achievements 2011

SERVICES

Change has been in the air at EMBL-EBI in the past year, with the announcement that Graham Cameron will retire in March 2012 after 30 years of service at EMBL. Graham was employee number two at the EMBL data library (now the European Nucleotide Archive) and his involvement in EMBL-EBI goes right back to its conceptualisation in the early 90s. It was therefore impossible to conceive of how we could possibly replace him. After an extensive international search, we have appointed Rolf Apweiler and Ewan Birney as joint Associate Directors. The decision to appoint two Associate Directors is testimony to Graham's unique wisdom and perspective. Rolf and Ewan have been jointly running the EMBL-EBI's largest service team, Protein and Nucleotide Databases, since 2007 and since their appointment have been working closely with Graham towards a smooth transition when Graham retires. These changes have not, however, distracted our service teams from enhancing existing resources and creating new services to meet the ever-evolving needs of our users.

Data compression

The EMBL-EBI constantly needs to stay ahead of increasing rates of data submission, and nowhere is this felt more keenly than in submission of DNA and RNA sequences. Data storage costs have become an appreciable proportion of total cost in the creation and analysis of DNA sequence data, and the rate of increase in DNA sequencing is significantly outstripping the price decrements in disk storage. Without some ingenuity, this would lead to ever rising data storage costs. In response, the ENA team has developed the CRAM format, which compresses DNA sequences by storing only differences between aligned and reference sequences with provision for reads that do not map to the reference. Expectations are that, at least for the mid-term, use of the CRAM format will stabilise storage costs despite the increasing data flow rate.

Genomes for all…

In addition to dealing with the threat of overwhelming amounts of data, EMBL-EBI serves a dizzying variety of different user communities. Ensembl, a collaboration with the Wellcome Trust Sanger Institute, has served researchers working on vertebrate genomes for more than ten years now and perhaps the highlight of 2011's newly sequenced genomes was that of our distant cousin the orang-utan (Locke, D.P. *et al.*, 2011). The Ensembl Genomes project, providing a comparable resource for researchers working on plants, fungi, bacteria, protists and non-vertebrate metazoans, was launched much more recently (in 2009) and remarkably now contains 335 species represented across the five kingdoms of life. Presenting accurate and up-to-date information on such a wide variety of organisms requires close interaction with the scientific communities that work on them. These collaborations have spawned other projects based on the Ensembl Genomes structure and data, serving the needs of specific research communities. In December, EMBL-EBI and Rothamsted Research launched a bioinformatics resource for studying plant pathogens: PhytoPath (www.phytopathdb.org). Working with researchers at the University of Cambridge and University College London, the Ensembl Genomes team launched PomBase (www.pombase.org) in November 2011, providing a means for researchers who work on the fission yeast *Schizosaccaromyces pombe* to contribute to the curation of its genome. This may lay the foundation for other community-driven curation efforts.

…and all genomes at once

The needs of metagenomics researchers, who study all the genomes present in a given environment without the need for prior individual identification or amplification, cut across several resources. Throughout 2011 the InterPro, ENA and UniProt teams have worked together to develop an integrated portal for metagenomics researchers (www.ebi.ac.uk/metagenomics). The portal is now accepting raw sequence reads; users can analyse predicted protein-coding sequences using InterPro and automatically have their sequence information archived in the ENA's Sequence Read Archive.

Enzymes and metabolic pathways

Another consolidation effort, involving several groups at EMBL-EBI, has been the development of the Enzyme Portal. Until now, information about enzymes was scattered throughout many different resources. The Enzyme portal (www.ebi.ac.uk/enzymeportal), which will launch early in 2012, mines and displays

Ewan Birney and Rolf Apweiler.

data about proteins with enzymatic activity from public repositories via a single search, and includes biochemical reactions, biological pathways, small molecule chemistry, disease information, 3D protein structures and relevant scientific literature.

Metabolomics researchers also have a new portal: Metabolights (www.ebi.ac.uk/metabolights) covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments. Metabolomics researchers can submit data, and specialist search tools for spectral similarities and chemical structures are available. Metabolights is produced in collaboration with the University of Cambridge.

Open-access literature

In July 2011 EMBL-EBI was awarded a contract for the continuation of UK PubMed Central (http://ukpmc.ac.uk, UKPMC), the free online literature resource for life science researchers. In this contract the EBI will lead the project. Now five years old, UKPMC has grown from a simple mirror of the National Center for Biotechnology Information (NCBI) PubMed Central site to a stand-alone site providing access to a repository of over 2 million full-text biomedical research articles, over 25 million citations from PubMed and Agricola, patents from the European Patent Office, UK treatment guidelines and biomedical PhD theses. Content is discoverable via an integrated full text and abstract search and is semantically enriched by the application of cutting edge text-mining approaches. Over 250 000 articles in this valuable resource are published under open-access licenses, which means that their contents can be freely reused. EMBL-EBI's Literature Services Team is leading the development of the service in partnership with the University of Manchester and the British Library. The goal is to build a gold-standard digital repository for the biomedical literature that benefits life science researchers throughout the world.

Putting the user first

The EBI's large-scale databases provide high-quality data and information curated by biologists, chemists, biochemists and bioinformaticians. We want to make it as easy as possible for researchers to explore and exploit this treasure-trove of publicly accessible data. This not only requires seamless integration of information that is managed by different groups, but also involves the design of intuitive user interfaces.

The need for data integration is now embedded in our culture. In addition to the multi-group collaborations described above, there have been numerous enhancements that make individual resources work more seamlessly with each other. For example, researchers submitting RNAseq data to ArrayExpress have their data automatically archived in the sequence read archive; PDBe users can access schematic views of structural coverage of UniProt entries and browse structures by chemistry, taxonomy or GO classification; and researchers using a range of different -omics-based approaches can cross-reference experiments submitted to different data archives back to the same sample.

In the past, input to our development in terms of user requirements and usability was gathered in a somewhat ad hoc way. Since 2010 we have systematically embraced the concept of user-centred design. It is already clear that by placing the user at the forefront of our minds as we design, test and implement our services, we create more useful and user-friendly resources. One such project produced the new, integrated search functionality, launched in January 2011. EBI Search organises search results around the 'central dogma' of molecular biology (i.e. DNA makes RNA makes protein), displaying an uncluttered results 'dashboard' from which users can explore genes, protein sequences, gene expression, molecular structures and the scientific literature. A species selector allows comparisons of key information for human, mouse, fly and other species, and the literature results include links to free full-text articles. While traversing the central dogma is a very natural route through the data, we believe that other kinds of conceptual connectivity, such as disease or chemistry, might map better on to the mindset of some searchers. Throughout 2011, we have been exploring such possibilities with the aim of extending the search concept in 2012.

RESEARCH

Our research groups are unified by the fact that they all use computational approaches, but differ in that they investigate a remarkably wide range of biological questions. Much of our research involves collaboration with experimental groups, and increasingly bioinformatics drives new experiments, completing a virtuous circle between the wet lab and the dry. Most of our researchers have a foot in both camps. To encourage this trend we have introduced funding for interdisciplinary postdoctoral fellowships involving two supervisors – one at EMBL-EBI and one at the Wellcome Trust Sanger Institute – based on the successful scheme run between different EMBL units. These fellowships have helped to stimulate this holistic approach.

Curiosity-driven research

A map of mammalian genomes: The Flicek and Goldman groups were both involved in the comparative analysis of 29 mammalian genomes in a collaborative study involving 19 organisations (Lindblad-Toh, K. *et al.*, 2011). Paul Flicek's group aligned the genomes, focusing on duplicated segments of DNA; Nick Goldman's group focused on looking at which parts of the genomes were under the most evolutionary pressure. Our genomes are very similar to those of other mammals, and certain parts – such as genes that encode for proteins – change particularly slowly because they are biologically important. The study determined that 4.2% of the human genome is evolutionarily conserved in this way and a potential function for around 60% of these bases can be proposed.

Conservation of translation: Transcription of transfer RNA genes, which requires a specific enzyme known as Pol III, is essential for generating the adaptor molecules that link genetic sequence and protein translation. Alvis Brazma's group, working with collaborators in Cambridge and Glasgow, contributed to a study that mapped Pol III occupancy throughout the genomes of several species to reveal that although Pol III binding to individual transfer RNA genes varies substantially, the combined

Nick Goldman (middle) with two members of his team: Greg Jordan and Sarah Parks.

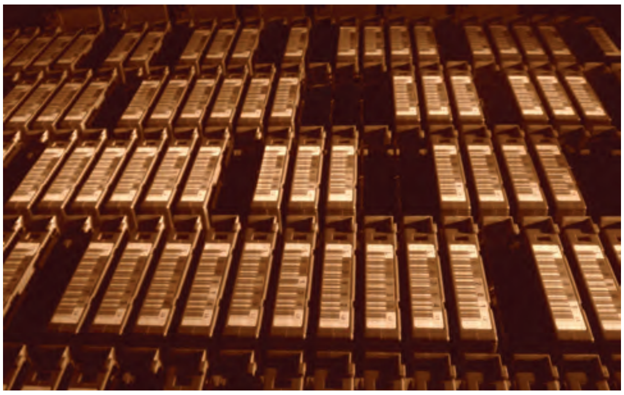
synthesis of transfer RNAs responsible for the transfer of each amino acid is highly constrained (Kutter, C. *et al.,* 2011).

Genetics of platelet count: Henning Hermjakob's team contributed to a study that identified regions on the genome associated with platelet count and volume, which affect the formation of blood clots. The team helped to construct a network that included most of the novel genes in the study, improving our understanding of the function of the molecular players involved, and pointing to potential new targets for the treatment of blood clotting disorders (Geiger, C. *et al.*, 2011).

A protein that prevents the genome from self-destructing: When genetic material is being duplicated, sequences of DNA called transposons copy and paste themselves with abandon in the genome. If left unchecked, this can cause serious damage to the genome. To avoid this, cells make sure these transposons are silenced, but how this works is poorly understood. A widely held theory is that two small RNA molecules, Miwi2 and Mili, act in tandem, each generating products that the other uses in a cycle that ensures transposons are silenced. Using a combination of genetic engineering, next-generation sequencing and bioinformatics, Anton Enright, working with Dónal O'Carroll's group at EMBL Monterotondo, found that only one of the proteins, Mili, is responsible for transposon silencing during embryonic development (De Fazio, S. *et al.*, 2011).

Directing gene expression during development: Pluripotent stem cells can differentiate into any cell type. Working with Brian Hendrich's group at the Wellcome Trust Centre for Stem Cell Research and MRC Centre for Stem Cell Biology and Regenerative Medicine, University of Cambridge, Paul Bertone's group elucidated the mechanism by which the nucleosome remodelling and deacetylation complex (NuRD) contributes to lineage commitment in pluripotent stem cells by precisely directing the expression of genes critical for embryonic development (Reynolds, N. *et al.,* 2011).

Modelling pathways involved in synaptic plasticity: Synaptic plasticity, the ability of nerve synapses to change in response to how much they are used, is a vital contributor to learning and memory. Nicolas Le Novère's group has built computational models explaining the behaviour of calmodulin-dependent protein kinase II, which is thought to play an important role in synaptic plasticity. The latest model (Stefan, M. *et al.*, 2012) predicts that binding of calmodulin alone is not sufficient to activate calmodulin-dependent protein kinase II, although high-affinity binding of calmodulin is.



Data tapes in the Hinxton data centre.

Resources for research

Evolution of enzyme-catalysed reactions: Working with colleagues at University College London, Janet Thornton's group has developed FunTree (www.ebi.ac.uk/thorton-srv/databases/FunTree), a new resource that brings together sequence, structure, phylogenetic, chemical and mechanistic information for structurally defined enzyme superfamilies. Central to the resource are trees generated from structurally informed multiple sequence alignments. These trees are decorated with functional information. Gathering together this range of data into a single resource allows users to investigate the evolutionary relationships between enzyme-catalysed reactions (Furnham, N. *et al.*, 2011).

Modelling signalling pathways: Julio Saez-Rodriguez and colleagues have developed CellNOpt, a toolbox for creating logic-based models of cellular signalling pathways and training them against data generated by high-throughput experiments.

Lexicon of life: Research is producing papers faster than anyone could ever hope to read them. Biologists therefore require increasingly sophisticated and efficient systems to help them to search for relevant information. Such systems need to deal with the fact that the authors of papers describe the same entity in different ways. Working with colleagues at the University of Manchester and other text-mining experts around the world, Dietrich Rebholz-Schuhmann's group developed the BioLexicon (Thompson, P. *et al.*, 2011). The BioLexicon gathers together different types of terms from several existing data resources into a single, unified repository, and augments them with new term variants automatically extracted from biomedical literature. It can then be used to improve the performance of text mining, making it easier to extract relevant papers from the huge body of biomedical literature.

Research leading to development of resources

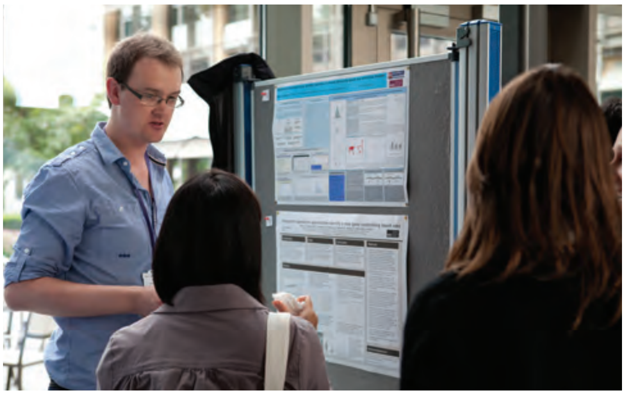
Reference-based compression: Markus Fritz, a PhD student in Ewan Birney's group, developed a new reference-based compression mechanism that enables the efficient storage of DNA sequence data. This work led to the creation of the CRAM toolkit (see page 4), which is enabling the European Nucleotide Archive to cope with unprecedented demand on storage capacity for DNA sequence data (Fritz, M. *et al.*, 2011).

Phylogeny-aware alignment tools: Nick Goldman's group expanded their phylogeny-aware alignment method PAGAN to perform extension of existing alignments with new data. They also published an analysis of several popular multiple alignment tools to quantify the false positives and false negatives introduced by alignment error and the ability of alignment filters to improve performance (Jordon, G. and Goldman, N., in press).

OTHER MAJOR ACHIEVEMENTS

A major ongoing effort is the migration of EMBL-EBI's databases to our new London Data Centre: our Systems and Networking team managed the migration of 18 databases to the London Data Centre throughout 2011. The London Data Centre has a geographically distributed topology to protect against local disaster, provides high-level security and is sited very close to high-bandwidth internet connections.

In September 2011 the Outreach and Training Team launched the beta version of Train online (www.ebi.ac.uk/training/online), its new web-based user-training resource that helps users worldwide learn about EMBL-EBI data resources in their own time and at



Poaster session, Summer School in Bioinformatics (EBI-Wellcome Trust course). Image courtesy of Samuel Kerrien.

their own pace. By the end of 2011 Train online had been accessed by 8000 unique users.

The Industry Programme welcomed two new major pharmaceutical companies as partners: Novartis and UCB.

All of these achievements were collaborative efforts, involving personnel from multiple teams.

COORDINATION OF MAJOR EUROPEAN PROJECTS

EMBL-EBI coordinates several EU-funded projects and is a partner in many others. Two large projects coordinated by EMBL-EBI reached successful completion in 2011: ENFIN and IMPACT.

ENFIN

The European Network of Excellence ENFIN was formed in November 2005 to provide Europe-wide integration of computational approaches in systems biology. The network focused on the development and critical assessment of computational approaches in this area, and uniquely brought together a variety of backgrounds and laboratory contexts ranging from investigative computer science through to traditional bench-based molecular biology. It comprised 20 laboratories across 11 different European countries plus Israel.

The computational work included the development of ENCORE, a database infrastructure for small laboratories, and ENSUITE, a toolbox of analysis methods including Bayesian networks, text mining, metabolite flux modelling and correlations of protein modifications to pathways. The network emphasised the study of mammalian intracellular signalling associated with the cell cycle.

ENFIN contributed actively to the systems biology community through the yearly organisation and support of international conferences and training courses, in particular the European School of Bioinformatics in collaboration with the EU-funded project BioSapiens.

ENFIN helped to foster close collaborations between experimental and computational researchers. One unique component of the project was a large internal fund to cover the cost of experimental validation of the computational predictions arising from the data analysis and modelling. This gave rise to 26 scientific projects throughout the course of ENFIN. By the end of the project in May 2011, ENFIN had been acknowledged in more than 110 publications in the scientific literature.

ENFIN's products and documentation are available on the website: www.enfin.org.

IMPACT

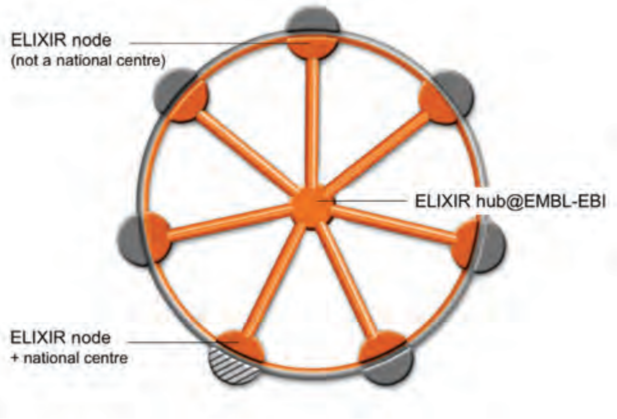
Protein families and domains are invaluable pointers that help biologists find distantly related proteins and predict their functions. A daunting array of resources, each with different strengths and weaknesses, is available to search genomes and proteomes for 'protein signatures' – diagnostic entities that are used to recognise a particular domain or protein family. To make it easier for life scientists to use these signatures effectively in their research, EMBL-EBI coordinates their amalgamation into a single resource called InterPro. The continued generation of these signatures, their annotation and integration into InterPro and the provision of software and databases to serve them to the public is therefore of great importance to the life science community.

The IMPACT (IMproving Protein Annotation through Coordination and Technology) project was completed in June 2011 and involved a consortium of nine experienced partners, some of whom have been collaborating for almost ten years in the field of protein family and domain prediction. These include the InterPro, PRINTS, Prosite, SuperFamily, SMART, Pfam, Gene3D and ProDom databases.

The quantity and quality of predictive signatures produced by the consortium increased significantly over the project's lifespan and, correspondingly, the number of proteins characterised by these signatures increased more than 2.5-fold, from 5 million to 13 million sequences. The consortium developed new interfaces to these data, including a redesign of the InterPro website, additional sources of structural and alignment data, new web services for end-users and a complete re-write of InterProScan – a powerful software tool that scans sequences for known signatures. A new XML format was developed by the consortium, which is used for information exchange among consortium members and by external users. Project developments were disseminated to end users through a training and outreach programme.

ELIXIR

ELIXIR, Europe's emerging infrastructure for biological information, entered the fifth and final year of its preparatory phase in November 2011. ELIXIR is a pan-European initiative to safeguard and foster data generated in life-science experiments. Its core objective is to ensure that Europe can continue to handle a rapidly growing volume and variety of data from high-throughput experiments such as DNA sequencing. Proper management of this information promotes knowledge-based economic growth, and facilitates the translation of research into innovations that meet global challenges in many key areas including food security, energy and health. ELIXIR will be coordinated from its hub hosted by EMBL-EBI and its nodes will be sited at appropriate centres in participating countries throughout Europe.



ELIXIR organisation: the 'Hub-and-Nodes' model.

Students in one of our hands-on training courses. Image courtesy of Samuel Kerrien.

2011 was an exciting and fast-paced year in which several milestones en route to ELIXIR's construction and operation were reached.

The completion and publication of the ELIXIR Business Case in early 2011 defined the process for ELIXIR's construction and operation.

During the spring and summer of 2011 we worked with our stakeholders throughout Europe to gain support for ELIXIR. By September 2011, five countries plus EMBL had signed a Memorandum of Understanding to catalyse the implementation and construction of ELIXIR. The Memorandum is a first formal – yet non-binding – step towards the implementation and construction of ELIXIR.

This enabled us to convene ELIXIR's Interim Board, the main body for negotiating the final legal and governance structure of ELIXIR. The first Interim Board meeting was held in in London in November 2011, during which Søren Brunak of the Technical University of Denmark was welcomed to his new role as elected Chair. By then a total of ten countries had signed the memorandum: Denmark, Estonia, Finland, the Netherlands, Norway, Slovenia, Spain, Sweden, Switzerland the European Molecular Biology Laboratory (EMBL). Since the meeting, Israel has also signed the memorandum. An important role of the Interim Board will be to establish an international consortium agreement and decide how ELIXIR will be governed and funded in the future.

By the end of 2011, funding bodies from several member states had committed a total of €117 million to the construction of both the Hub and Nodes of ELIXIR. A significant proportion of this – some £75 million (€90 million) – comes from the UK's Department for Business, Innovation and Skills' Large Facilities Capital Fund (LFCF) as a commitment to EMBL-EBI. This funding will allow the construction of facilities for ELIXIR's central Hub at EMBL-EBI on the Wellcome Trust Genome Campus in Hinxton, Cambridge. The hub will be a nerve centre for bioinformatics in Europe, helping to coordinate the delivery of services and user training from centres of excellence Europe-wide. The hub will also establish a robust computing infrastructure that can handle the rising tide of life science data. Other significant financial contributions towards the construction of ELIXIR nodes throughout Europe have been made by Denmark, Finland, Norway, Spain, Sweden and Switzerland.

Selected publications

De Fazio, S., *et al.* (2011) The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* 480, 259-63.

Fritz, M., *et al.* (2011) Efficient storage of high throughput sequencing data using reference-based compression. *Genome Res.* 21, 734-40.

Furnham, N. *et al.* (2012) FunTree: A resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.* 40, D776–82.

Gieger, C. *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature* 480, 201-8.

Jordan, G. and Goldman, N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* Article in press.

Kutter, C., *et al.* (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43, 948-55.

Lindblad-Toh, K., *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–82.

Locke, D.P., *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469, 529-33.

Reynolds, N., *et al.* (2011) NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression. *EMBO J.* 31, 593-605.

Stefan, M., Marshall, D. and Le Novère, N. (2012) Structural analysis and stochastic modelling suggest a mechanism for calmodulin trapping by CaMKII *PLoS ONE* 7, e29406

Thompson, P., *et al.* (2011) The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinform.* 12, 397.

Services

Protein and nucleotide data

Rolf Apweiler

PhD 1994, University of Heidelberg. At EMBL since 1987, at EMBL-EBI since 1994.

Ewan Birney

PhD 2000, Sanger Institute. At EMBL since 2000.

DESCRIPTION OF SERVICES AND RESEARCH

The Protein and Nucleotide Data group was created in June 2007 by merging the former Ensembl (Birney) and Sequence Database (Apweiler) groups. The group focuses on the production of protein sequence, protein family and nucleotide sequence databases at EMBL-EBI. We maintain and host the European Nucleotide Archive (see page 14), the Ensembl and Ensembl Genomes resources (see pages 16 and 18), the UniProt protein resource (see pages 26-29), the InterPro domain resource (see page 24) and a range of other biomolecular databases. These efforts can be divided into three major groups: nucleotides, proteins and chemoinformatics and metabolism. Substantial training and outreach efforts are part of these activities. Both the Birney and Apweiler groups have complementary research components.

SUMMARY OF PROGRESS

- Handled an ever-growing amount of data in all of our databases;
- Established a method for reference-based compression and released the CRAM toolkit (see page 14);
- Released spreadsheet-based, interactive submission system for next generation data into the Sequence Read Archive (SRA Webin; see page 14);
- Launched PomBase, a new resource for the model yeast Schizosaccharomyces pombe, based partly on the use of Ensembl technology (see page 18);
- Adapted data submission tools and created pipelines for metagenomics data analysis;
- Launched the Metagenomics portal (in beta; see page 24);
- Published five releases of Ensembl encompassing new species, new genome assemblies and extensive expansions of regulatory and variation data.

MAJOR ACHIEVEMENTS

Nucleotide resources

Ewan Birney provides strategic oversight across all nucleotide resource groups. The three main branches of activity are: vertebrate genomics (Ensembl; see Paul Flicek, page 16); non-vertebrate genomics (Ensembl Genomes; see Paul Kersey, page 18) and nucleotide sequences (European Nucleotide Archive; see Guy Cochrane, page 14). The HUGO Gene Nomenclature Committee (HGNC), a smaller group coordinated by Elspeth Bruford, is presented here. The key organising principal across all these groups is to best coordinate resources for each genome sequence.

DNA sequence remains at the heart of molecular biology and bioinformatics. In 2011 the use of next-generation sequencing technologies became nearly universal, which impacted all EMBL-EBI teams in this area. Our European Nucleotide Archive Team launched a new compression toolkit, CRAM, which saves storage costs through the intelligent compression of very large data sets. Ensembl expanded to include more species as well as regulatory and variation data, and Ensembl Genomes grew to represent many more species, including plants and nematodes.

A key difference between the groups is the provenance of the data. Datasets in the European Nucleotide Archive are determined by the submitter (any added value information is provided as additional resources), which means they may be redundant or conflicting; however, they represent the foundational DNA dataset on which all genomic (and nearly all protein sequence) is based. This dataset is coordinated worldwide, in partnership with NCBI and DDBJ, as the International Nucleotide Sequence Database Collaboration (INSDC), and forms a single, worldwide coordinated set of information. In contrast, Ensembl and Ensembl Genomes are community led and we aim to present a single, non-redundant view of each species' genomic sequence. For these resources, interactions with the community help determine the way information is represented to users. For example, the unambiguous assignment of gene symbols allows researchers to use memorable names for genes in scientific communications; for the human genome, these assignments are provided by the HGNC group.

HUGO Gene Nomenclature Committee

Elspeth Bruford, Louise Daugherty, Susan Gordon, Michael Lush, Ruth Seal, Matt Wright

The HUGO Gene Nomenclature Committee (HGNC) is the only worldwide authority that assigns standardised human gene nomenclature, and remains an essential component of human gene and genome management. The HGNC has two overriding goals: to provide a unique name and symbol for every human gene; and to ensure this information is freely available, widely disseminated and universally used. Achieving these goals involves:

- Bioinformatic analysis of nucleotide and protein sequences;
- Curation of www.genenames.org, a database comprising individual gene records containing the gene name, symbol and relevant information (e.g. cDNA sequence, chromosomal location, key publications, links to other databases);
- Communication, including: consultation with researchers; collaboration with nomenclature groups for other species in order to coordinate naming of orthologous genes; exchanging data with numerous databases; and raising awareness of the resource within the scientific community through publications, digital media and attendance of conferences and meetings.

In 2011 gene naming focused on the consensus genes identified by the CCDS project, with approximately 60 of the CCDS genes in the current release of 18 472 in need of an HGNC-approved gene symbol. Over the course of the year the total number of approved gene symbols increased by 2700, and now stands at 32 574. This was largely thanks to the increased assignment of names for non-protein-coding RNA genes and to the naming of pseudogenes based on the pseudogene.org dataset.

Members of the team co-authored several academic papers in 2011 and attended international meetings in Canada (ICHG/ASHG), Japan (RNA 2011), the Netherlands (ASHG), the United Arab Emirates (HGM2011), and the United Kingdom (Abcam 2011 and Quest for Orthologs). HGNC lost two key team members in 2011: Susan Gordon returned to Canada and, after 11 years with the Committee, Michael Lush took a new position at Cancer Research UK.

Protein resources

Rolf Apweiler provides strategic oversight across all protein resource groups. The main branches of activity are: UniProt (see Maria-Jesus Martin, page 26, and Claire O'Donovan, page 28), GOA (see Jane Lomax, page 20), InterPro (see Sarah Hunter, page 24) and Proteomics Services (see Henning Hermjakob, page 22). The RESID project, overseen by John Garavelli, is presented here.

EMBL-EBI protein resources provide public access to all known protein sequences and functional information about these proteins. UniProt is at the centre of these activities. Most UniProt sequence data is derived from the translation of nucleotide sequences, which are provided by the European Nucleotide Archive and Ensembl. All UniProt data undergoes annotation with Gene Ontology terms (GO) and uses the classification into protein families and domains provided by InterPro. The UniProt team adds information that its curators extract from the scientific literature as well as from curator-evaluated computational analyses. UniProt makes use of automatic annotation to add information to the sequence data without adding experimental functional data; this is based on InterPro and literature annotation approaches. The IntAct protein–protein interaction database and the Protein Identification (PRIDE) database provide protein interaction and identification data to UniProt.

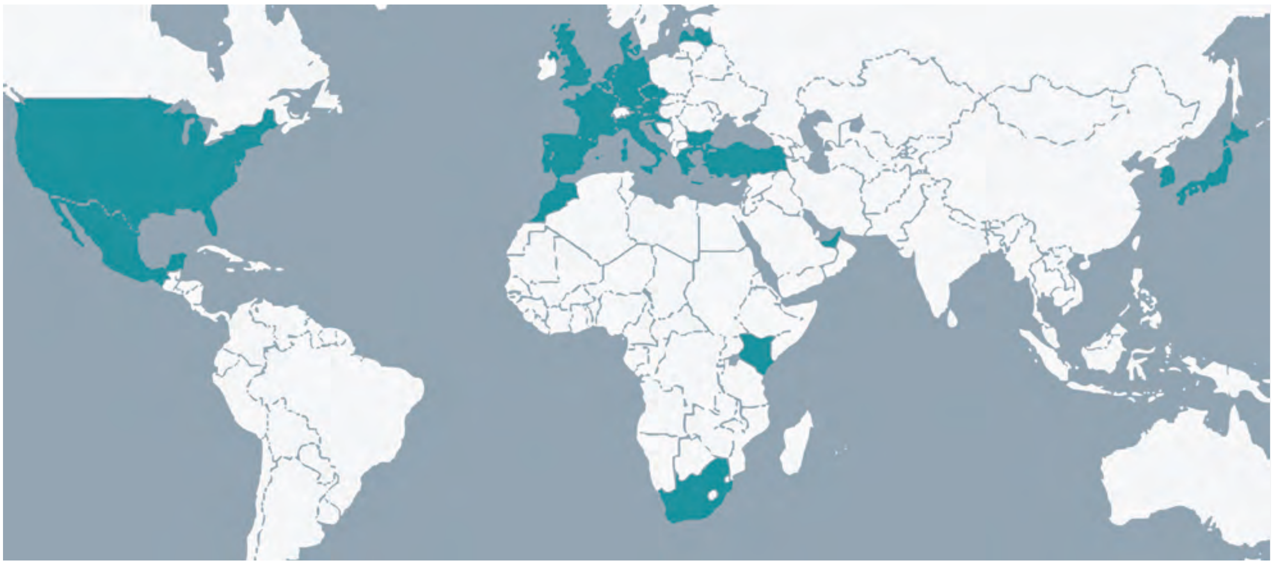


Figure. Locations that hosted PANDA training activities in 2011.

The RESID Database of Protein Modifications

John S. Garavelli

RESID (Garavelli, 2004; www.ebi.ac.uk/RESID) is a comprehensive collection of annotations and structures for protein modifications and cross-links. It provides systematic and alternate names, atomic formulas and masses as well as enzymatic activities that generate the modifications, keywords, literature citations and cross-references to GO, ChEBI, PSI-MOD, the Protein Data Bank, structure diagrams and molecular models. As of December 2011, RESID contained 572 entries for chemically unique protein modifications.

RESID documents the controlled vocabulary for natural protein modifications in the feature table annotations of the UniProt Knowledge Base (UniProtKB), and supplements the modification descriptions with more detailed information. It was used during the initial phase of the UniProt project to merge the feature annotations of Swiss-Prot and the Protein Information Resource (PIR), and to design new standard annotations. RESID provides original reports for new types of modification and for modifications newly found in additional proteins. This information contributes to the annotation of UniProtKB by describing the newly discovered modifications, producing standard feature annotations for them and predicting their occurrence in other entries through automated annotation. Researchers working in high-throughput proteomics can search RESID to find monoisotopic masses and mass differences and to identify known and predicted protein modifications. RESID will also suggest the modified sequences from alternative isobaric peptides. RESID contributes to the Proteomics Standards Initiative ontology of protein modifications (PSI-MOD), which is maintained by John Garavelli for the Proteomics Standards Initiative and PRIDE (see page 22).

OUTREACH AND TRAINING

Jeff Almeida-King, Denise Carvalho-Silva, Bert Overduin, Michael Schuster, Giulietta Spudich

Delivering quality workshops and maintaining our worldwide outreach is a priority for the group. In addition to running our own workshops, we contribute regularly to the EMBL-EBI hands-on courses and Bioinformatics Roadshow (see Cath Brooksbank, page 80). In 2011 we extended the reach of our programme, with 26 countries hosting our training events (see Table). Highlights included a series of workshops throughout Japan and Korea, as well as countries targeted by EMBL-EBI's SLING grant: Bulgaria, Croatia, Slovenia and Turkey. Thirteen Ensembl workshops focused on the Perl API in 2011, reflecting an increased request for this type of developer-oriented workshop (see Figure; countries that hosted an Ensembl browser or API workshop in 2011 are shown in green).

Subject	Courses
Ensembl (+ Ensembl Genomes) browser or API	93
Protein databases (UniProt, InterPro)	28
Proteomics (IntAct, PRIDE)	24
Sequence databases (ENA, GOA)	6
Pathways (ChEBI, Reactome)	15

Table. Training events run by the Protein and Nucleotide Data group. Our outreach team also participates in the EBI Bioinformatics Roadshow (see page 80).

RESEARCH

We run an active trainee programme in which undergraduates and PhD students (usually Marie Curie fellows) join our group for a period of three months to a year, applying their theoretical knowledge to practical problems. In 2011 the group hosted 30 trainees and X visitors, who worked on a broad range of projects.

Apweiler research

PhD student Joe Foster is currently working under the supervision of Rolf Apweiler.

Joe Foster

Investigating the application of peptide retention time for improved transition selection in Single Reaction Monitoring

The field of lipidomics is undergoing rapid expansion as high-throughput identification and quantification instrumentation are adopted by an increasing number of laboratories. The field of lipidomics has much to gain by implementing bioinformatics principles that are common in proteomics. Protein databases form the basis of the majority of proteomics experiments, acting as reference databases for peptide identification and protein inference; sources of protein-specific metadata; and a centralised repository of literature references. A comprehensive lipid database on a par with UniProtKB in proteomics has yet to emerge; although some resources exist, they lack a foundation to support modern, high-throughput methods. LipidHome is a database of theoretical lipid species organised in a novel structural hierarchy, using the most recent lipid nomenclature standards. It is a comprehensive dictionary of lipid species that can relate lipid identifications from all technologies, regardless of structural resolution. It provides easy access to the raw data and a selection of tools and related metadata through an intuitive web application and simple web services.

Birney research

Ewan Birney's research group focuses on algorithmic methods for genome analysis and the use of genetic association techniques to understand basic biology. Three PhD students are currently working under the supervision of Ewan Birney: Markus Hsi-Yang Fritz, Dace Ruklisa and Sander Timmer. Mikhail Spivakov, a joint EIPOD postdoc, also works in the group.

Markus Hsi-Yang Fritz

Hominid segmental duplications and repeat evolution

We have created a robust, scalable segmental-duplication pipeline that can find duplication regions reliably and in reasonable time. Using diagnostic subsequences from this discovery pipeline and other resources, we can probe large-scale short read data to understand the distribution of such duplications in the absence of assemblies. This allows the use of low-coverage data such as the 1000 Genomes Project data as well as other sources such as Neanderthal information.

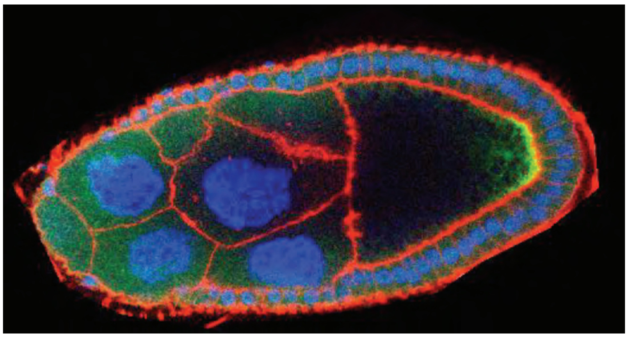


Figure. A Drosophila egg chamber (oocyte) present in a near-isogenic wild Drosophila line. The upper large area is the egg. The cells below with large nuclei, stained blue, are the nurse cells. We are able to find genetic variations associated with the size and shape of the nurse cells.

Dace Ruklisa

Large-scale association studies: from inference framework to effects

We use involved statistical models to explore the relationship between genotype and phenotypes. Using *Drosophila* oocytes, we converted in situ images of 70 different isogenic lines into a complex array of 92 different phenotypes. This allows us to use phenotypes with known individual genotypes to find associations. By carefully separating the components of variance specific to an individual from a strain, we can discover a variety of robust genotype-to-phenotype associations. In human data, we used statistical subsampling techniques to generate richer, multi-SNP models for associations.

Sander Timmer

Association studies in development of Drosophila *and Human*

We are looking at genetic variation in fundamental processes of *Drosophila* development, in particular at gene expression in the early fly embryo. We are also looking at gene expression during human skeletal development. In the former case we are working with the Furlong laboratory to perform a large-scale eQTL study focused on early development. In the latter we are using human MRI scans to provide robust measurements of human skeletal phenotypes.

Mikhail Spivakov

EIPOD project: Drosophila *mesoderm development*

Medaka fish is a promising model for population studies for a number of reasons, including the high survival rates of inbred lines, a relatively small genome and a vast amount of accumulated knowledge on the husbandry, genetics and phenotypic diversity of the species. In collaboration with the groups of Prof Jochen Wittbrodt (University of Heidelberg), Prof Kiyoshi Naruse (NIBB, Okazaki, Japan) and Dr Felix Loosli (Karlsruhe Institute of Technology) we have initiated an effort to create a resource for medaka population genomics. The ultimate goal is to make available a collection of several hundred genotyped medaka inbred lines derived from the same population. We are involved in the overall coordination of the project, with specific responsibilities for the sequence analysis of the founder population and the future inbred lines.

Cited reference: Garavelli, J.S. (2004). The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics*, 4, 1527-33.

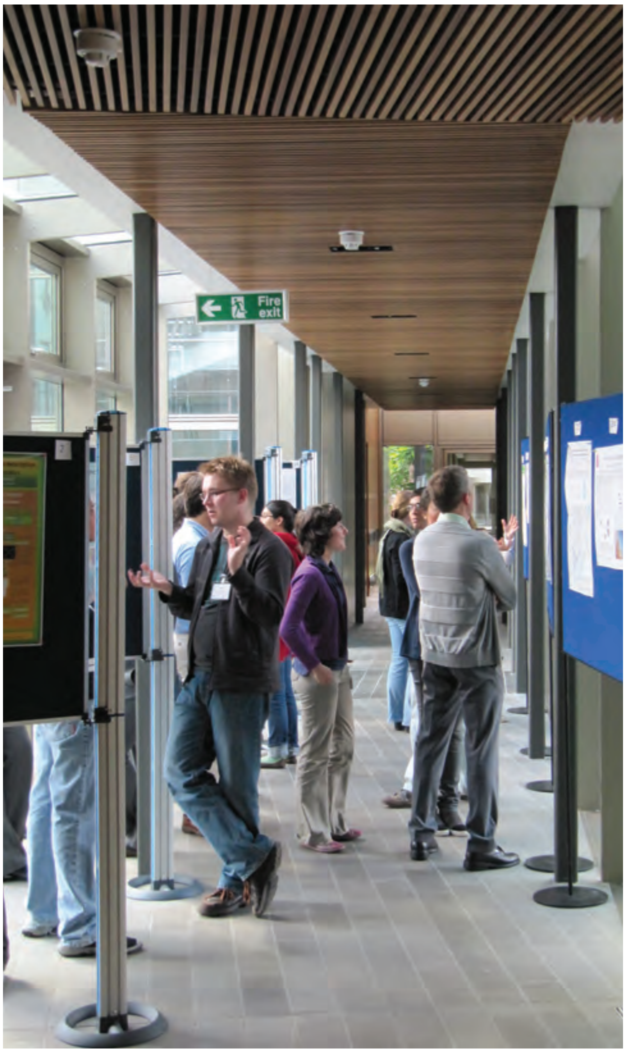
Selected publications

Alam-Faruque, Y., Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

O'Donovan, C. and Apweiler, R. (2011) A guide to UniProt for protein scientists. *Methods Mol. Biol.* 694, 25-35.

Birney, E. (2011) Chromatin and heritability: how epigenetic studies can complement genetic approaches. *Trends Genet*. 27 (5), 172-6.

Fritz, M.H., Leinonen, R., *et al.* (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40.



The European Nucleotide Archive

Guy Cochrane

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

DESCRIPTION OF SERVICES

The European Nucleotide Archive (ENA) provides globally comprehensive primary data repositories for nucleotide sequencing information. ENA content spans the spectrum of data from raw sequence reads through assembly and alignment information to functional annotation of assembled sequences and genomes. Services for data providers include interactive and programmatic submission tools and curation support. Data consumers are offered a palette of services – sequence similarity search, text search, browsing, rich integration with data resources beyond ENA – all provided over the web and through an increasingly sophisticated programmatic interface. All ENA services are supported with a helpdesk and a growing training programme. These services are for users who approach ENA data and services directly, and those who provide secondary services (e.g. UniProt, Ensembl, Ensembl Genomes, ArrayExpress) that build on ENA content. Reflecting the centrality of nucleotide sequencing in the life sciences and the emerging importance of the technologies in applied areas such as healthcare, environmental and food sciences, ENA data and services form a core foundation upon which scientific understanding of biological systems has been assembled and our exploitation of these systems will 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.

SUMMARY OF PROGRESS

- Enabled the capture, processing and presentation of major new raw read, read alignment, assembly and annotation data sets;
- Improved data submission pipelines for next-generation data from diverse platforms in a variety of formats;
- Further developed the ArrayExpress data-brokering scheme to allow metadata components of direct submissions from existing large-scale submitters (e.g. the Wellcome Trust Sanger Institute) to be routed to ArrayExpress curators;
- Deployed new similarity-search algorithms and datasets in ENA Sequence Search;
- Enabled community-informed development of a progressive compression strategy for sustainable ENA growth;
- Engineered and tested the CRAM Toolkit: a compression codebase;
- Significantly improved the ENA browser and text search, and delivered these improvements in a stepwise manner;
- Supported the ENA community by developing training courses and online training materials, and by delivering workshops.

MAJOR ACHIEVEMENTS

The ENA team launched nine new submission templates into our Webin system in 2011, bringing template coverage to the majority of curated submissions of assembled and annotated sequences. We also launched SRA Webin, a spreadsheet-based submission tool for interactive next generation sequence data submissions.

The development and launch of the ENA Taxon Portal and underlying data warehouse in 2011 has resulted in backend improvements and now provides richer real-time access to the massive data sets that we store.

We were key partners in a research project leading to the publication of a proof-of-principle paper on reference-based sequence compression. The method we proposed has significant implications for the storage of raw sequence data, and accordingly for the reliability of the scientific record.

It is crucial that we are able to adapt quickly to changes in sequencing technology and to user requirements: accordingly, we lead a community-facing sequence read-compression initiative called CRAM. During 2011 we released the CRAM toolkit (in beta), a robustly engineered implementation of reference-based compression.

A major area of prototyping in 2011 was around the representation of assembly information. While ENA has traditionally captured and presented this information, the internal data structures and the public presentation of the content in these structures has been based upon concepts derived from flat file structure. As the complexity of assembly information grows, such as with the use of assembly patches and alternative allele components, our traditional data structure has struggled to accommodate the storage of granular information of maximum utility to our consumers. Our prototype, a relational schema based on modelling of the assembly process itself and parallel work being carried out by our colleagues at the National Center for Biotechnology Information (NCBI), underwent testing with the UniProt, Ensembl and Ensembl Genomes teams and rounds of iterative improvement are in progress.

FUTURE PLANS

In 2012 we will focus on the consolidation of data submission services around the secure, dropbox-based model that has proven successful and flexible for SRA data submissions. Early in the year, we will deploy project registration services under this model and later development will be the addition of assembly submission services. In addition, we will prototype under our submissions services brokering to the EBI BioSamples Database.

Given the volume and dynamic nature of ENA content, a major challenge is the provision of data warehouses that can serve data to our interfaces in rapid and robust ways. In 2012 we will revisit our indexing and warehousing infrastructure and investigate the potential of new technologies to support rapid and frequent warehouse rebuilding (to support the latest content) and flexible interactive queries (to support such functionalities as faceted search and Boolean operations on search results).

In 2012, we expect our prototype for assembly information representation to reach maturity. We will then turn our attention to the ongoing curation of assembly information and the delivery of an easily consumed assembly data product.

Selected publications

Amid, C., Birney, E., Bower, L., *et al.* 2012. Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res.* 40, D43-7.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., *et al.* 2011. Efficient storage of high-throughput DNA sequencing data using reference-based compression. *Genome Res.* 5, 734-40.

Karsch-Mizrachi I., Nakamura, Y., Cochrane, G. 2012. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40, D33-D37.

Kodama Y., Shumway, M., Leionen, R., 2012. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40, D54-6.

We will continue to work on sustainable data archiving. In 2012, the CRAM toolkit is expected to mature from beta into production-ready code and we expect SRA submission pipelines, core data storage and presentation layers to use the CRAM format and toolkit to some extent. Our approach to CRAM toolkit development will continue to reflect a strategy that recognises the value of deep integration of reference-based compression in existing community tools and workflows but draws on the great potential that 'lossless' and 'lossy' compression offer.

In the context of CRAM, we will also focus on development and implementation of policy around the application of compression, specifically the nature of lossy compression models and the extent of data reduction applied under these models. Progressive compression, in which greater data reduction is applied to more reproducible sequence data (those data from highly available or reproducible samples) and greater data retention is applied to less reproducible sequence data (those data from rare and irreplaceable samples), is central to sustainable data archiving.

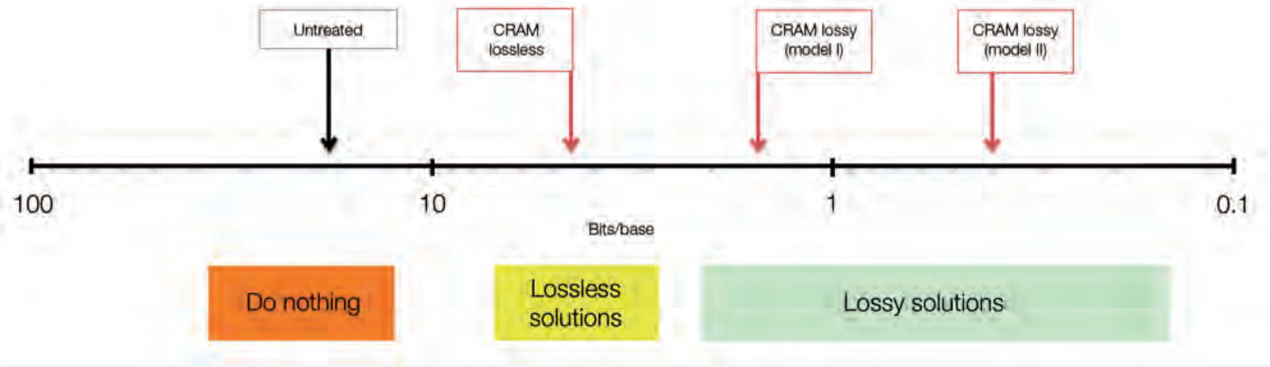


Figure. Reference-based compression scale: levels of compression in bits per base that can be achieved using CRAM reference-based compression. Current practice ('Untreated') is to use binary file formats, such as fastq and BAM. Lossless models under CRAM achieve around four-fold improvements in compression. Two lossy models are shown (models I and II) in which 10- and 50-fold compression can be achieved. Under more aggressive models, greater levels of compression can be achieved (not shown).

Vertebrate genomics

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.

DESCRIPTION OF SERVICES

Our team creates and maintains the genomic resources of the Ensembl project, a joint project with the Wellcome Trust Sanger Institute, and is responsible for data management for a number of large-scale international projects, including the 1000 Genomes Project and, in collaboration with the Functional Genomics team, the International Mouse Phenotyping Consortium. We also maintain and develop EMBL-EBI's major variation databases, including the European Genome-phenome Archive (EGA) and the DGVa database of copy number and structural variation. All of these resources are made publicly available and are widely used by the scientific community and by the team itself as part of our research into evolution, epigenetics and transcriptional regulation.

Our specific research projects, largely done in collaboration with Duncan Odom's group at the University of Cambridge, focus on the evolution of transcriptional regulation. Recently we have expanded 'comparative regulatory genomics' techniques including mapping the same DNA–protein interactions in matched tissues in multiple species to understand how gene regulation has evolved while the tissue-level functions are largely conserved. We are also interested in the role of chromatin conformation in tissue-specific gene regulation and have investigated both the CTCF and cohesin complex in this context.

SUMMARY OF PROGRESS

- Issued five releases of Ensembl encompassing extensive expansions of human variation and regulatory data in addition to new species, new genome assemblies and other new features;
- Launched the International Mouse Phenotyping Consortium (IMPC);
- Published 27 scientific papers in peer-reviewed journal articles, representing all of the group's major activities.

MAJOR ACHIEVEMENTS

The five major releases of Ensembl in 2011 featured extensive updates to the core genomic resources provided for human, mouse, rat and zebrafish. They also included five new supported species, including turkey and the endangered gibbon and Tasmanian devil genomes. We continued to support updates to the human reference assembly, including the recent 'patch version' releases. Our variation data resources such as the Ensembl Variant Effect Predictor (VEP) support a large, integrated collection of sequence variation associated with disease as well as the reference data included in the most recent version of dbSNP, which comprises the comprehensive, world-wide variation data created by the 1000 Genomes Project. Our participation in the ENCODE project included both extensive analysis of whole-genome functional data as well as the incorporation of the most important ENCODE datasets and results into Ensembl. The Ensembl outreach team has conducted nearly 100 hands-on training courses across Europe and around the world.

The European Genome-phenome Archive (EGA) nearly doubled the number of available studies in 2011 to approximately 150 and introduced a new web interface that makes it much easier for users to find and submit data. The DGVa database of structural and copy number variation now contains nearly every available CNV/SV dataset. Streamlined data-submission and -exchange procedures were introduced with its peer database, dbVar, at the NCBI and the project's main collaborator at the Database of Genomic Variants in Toronto.

In collaboration with Helen Parkinson in the Functional Genomics team (see page 38), the mouse informatics team will lead a five-year effort to provide the data management and coordination for the NIH-funded Knockout Mouse Phenotyping Program (KOMP2). The KOMP2 effort, our on-going work with the European Mouse Mutant Archive (EMMA) and the Infrafrontier European infrastructure are all key components of the International Mouse Phenotyping Consortium, which was formally launched in September 2011.

Data from the 1000 Genomes Project was accessed by a steadily growing number of users, with a considerable increase following the October 2010 initial project publication. We released the full phase 1 data set and developed several new tools to enable community access to the data. We also started the process of collecting data from phase 2 of the project.

Our research into the comparative regulatory genomics of the DNA-binding protein CTCF, led at EMBL-EBI by PhD student Petra Schwalie, demonstrated that over evolutionary time CTCF's binding profile in mammalian genomes has been dramatically affected by waves of retrotransposon insertions. Specific retrotransposon repeats containing the

CTCF binding site have spread though the mammalian genome multiple times, including within the past 20-30 million years in the mouse and rat genomes, leaving behind a conserved hierarchical signature. In addition, the same data demonstrated that CTCF has an evolutionary conserved 34 bp binding site, which is approximately twice as long as previously known.

FUTURE PLANS

The rapidly growing volume and diversity of data across the scope of genomics research is the major challenge for projects like Ensembl. We see an ever-increasing number of whole-genome sequences as well as comprehensive variation, regulatory, disease and phenotype data in human and other species. We have created a number of analysis and visualisation methods to summarise and present dense and complex regulation data (see Figure) and will continue to expand these to other species and disease states. At the same time, the EU Blueprint project and the NIH-funded KOMP2 project are both expected to produce their first major data sets in 2012. This means that we will continue to play an end-to-end role in major genomics projects from raw-data management for the project to summary-data presentation to the wider scientific community.

Selected publications

Flicek, P., Amode, M.R., *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.* 39 (Database issue), D800-6.

Lindblad-Toh, K., Garber, M., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

Locke, D.P., Hillier, L.W., *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469 (7331), 529-33.

Marth, G.T., Yu, F., *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.* 12 (9), R84.

Renfree, M.B., Papenfuss, A.T., et al. (2011) Genome sequence of an Australian kangaroo, Macropus eugenii, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.* 12 (8), R81.

Our research projects are expanding in number of species, tissues and specific DNA–protein interactions. We will also focus on understanding the differentiation process and components of cell- and tissue-specific regulation. These questions will be addressed both in the context of our established collaborative projects with the Odom group and as part of other collaborations, including larger EU-funded projects.

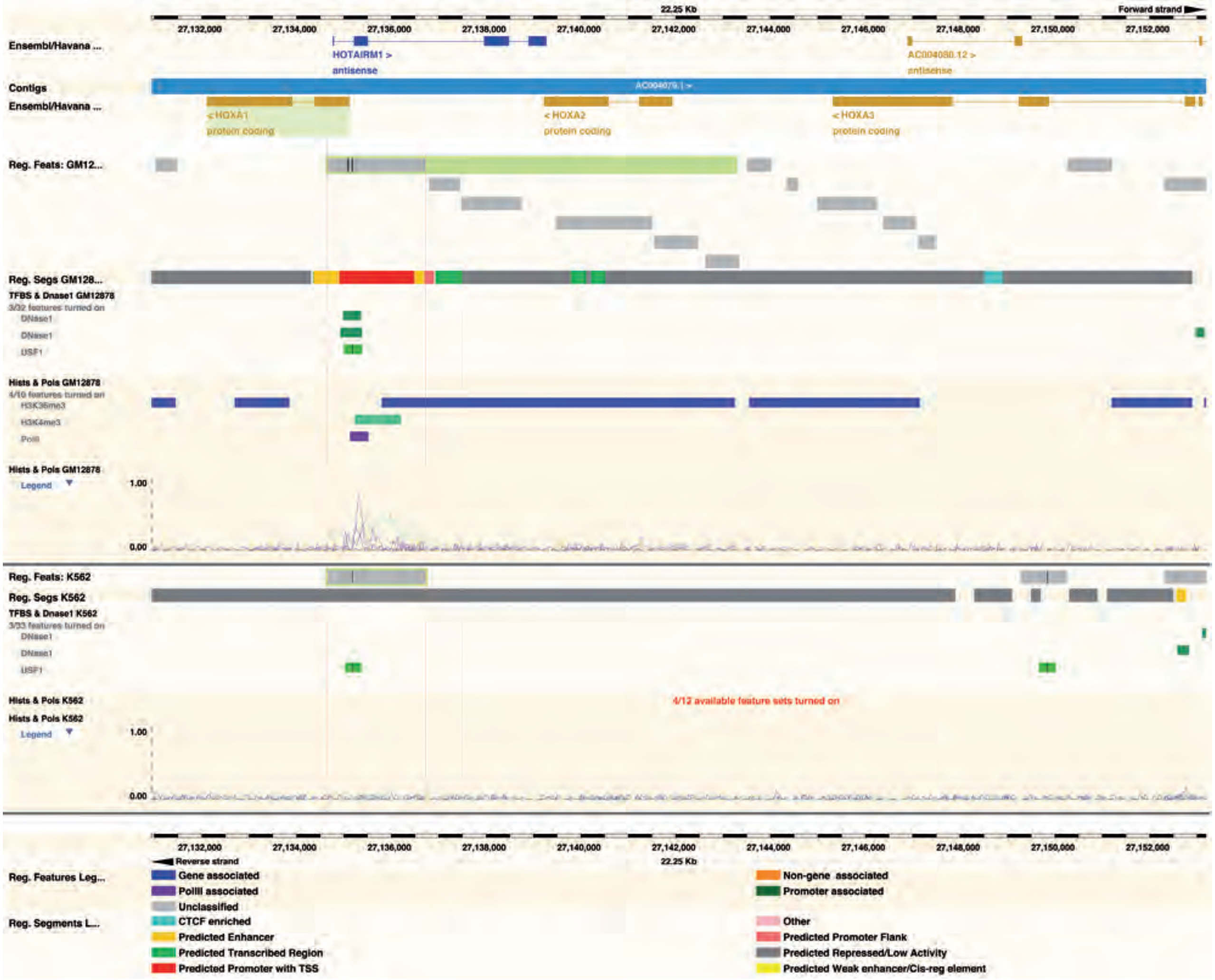


Figure. Raw, processed and summarised data from two human cell lines in region around the HOXA1 gene on chromosome 7 as presented in Ensembl. Lymphoblastoid (GM12878) and myelogenous leukemia (K562) cell lines from the ENCODE project show differences in raw data (wiggle tracks), processed data (coloured rectangular boxes in the tracks labelled DNase1) and summary level multi-coloured segmentation tracks demonstrating the difference at the chromatin and regulatory level between the active gene in GM12878 cells and the inactive gene in K562 cells.

Ensembl genomes

Paul Kersey

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

DESCRIPTION OF SERVICES

The Ensembl Genomes team provides services based on the genomes of non-vertebrate species. The falling costs of DNA sequencing have led to an explosion of reference genome sequences and genome-wide measurements and interpretation. Ensembl Genomes (Kersey *et al.*, 2012) provides portals for bacteria, protists, fungi, plants and invertebrate metazoa, offering access to these 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.

Even small communities with little informatics infrastructure can now perform highly complex and data generative experiments, once the sole domain of large, internationally coordinated sequencing projects. Through collaborating with the EBI and re-using our established toolset, such small communities can store, analyse and disseminate data more cheaply and powerfully than if they develop their own tools. Our leading collaborators 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), focused on the fission yeast *Schizosaccharomyces pombe*. In the plant domain, we collaborate closely with Gramene in the US and with a range of European groups in the transPLANT project (see page X). 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. We have developed a new portal for plant pathogen data, PhytoPath (launched in early 2012), and are involved in the development of Microme, a new resource for bacterial metabolic pathways.

SUMMARY OF PROGRESS

- Launched PomBase, a new resource for the model yeast *Schizosaccharomyces pombe*, based partly on the use of Ensembl technology;
- Initiated the transPLANT project, a new European collaboration to develop a trans-national infrastructure to support plant genomic science;
- Became a partner in the WormBase consortium, which has provided a genome-centric resource for the nematode research community since 2000;
- Significantly expanded the coverage of all domains of life within Ensembl Genomes, with particular emphasis on invertebrate metazoa, plants, and plant and oomycete plant pathogens (with 330 species now represented across the five divisions);
- Issued four public releases of Ensembl Genomes.

MAJOR ACHIEVEMENTS

In the course of 2011 we made four public releases of Ensembl Genomes, considerably expanding our coverage of the taxonomic space. Particular efforts focused on incorporating new genomes of invertebrate metazoans, plants and plant pathogens (fungal and oomycete). The latter efforts are principally funded as part of PhytoPath, a BBSRC project with Rothamsted Research which integrates data about pathogenic phenotypes with the genomic data made available in Ensembl (launched in early 2012).

In 2011 we launched PomBase, a database to meet the needs of the community working on fission yeast, an important model species. PomBase combines a major new effort focused on community collaboration of the relevant scientific literature with the use of the Ensembl database infrastructure to support integration of high-throughput data and more powerful comparative analyses with other species. The team additionally joined the WormBase consortium which has been serving similar needs for the nematode research community since 2000, and will contribute to the future development of this project.

The ongoing development of variation resources for *Arabidopsis thaliana*, which is serving as a model for our presentation of this data, has resulted in the production of a resource that now contains data from over 1600 strains, integrating the first published results from the “1001 genomes” initiatve to catalogue the variation present in this species through whole genome resequencing with earlier, array-based genotyping data. A variation resource has also been published for the plant pathogen *Giberella zeae*.

FUTURE PLANS

In 2012 we expect to further increase the number of genomes included in Ensembl Genomes and plug the remaining gaps in our taxonomic coverage. We have been funded to join with leading groups engaged in wheat and barley research who are attempting to assemble and annotate the complex genomes of these cereal species. Wheat in particular has a large, repetitive and hexaploid genome and has long been considered a particular challenge in genomics. These efforts will occur in the contexts of our existing and new collaborations in the plant domain, with the goal of unifying global efforts and establishing a reference, international resource for plant genomics. The genomes of bacteria are less well served by our current models of data organisation than those of eukaryotes and a major restructuring of our services for these will occur soon; this will result in significantly increased coverage of this kingdom within our resources.

Selected publications

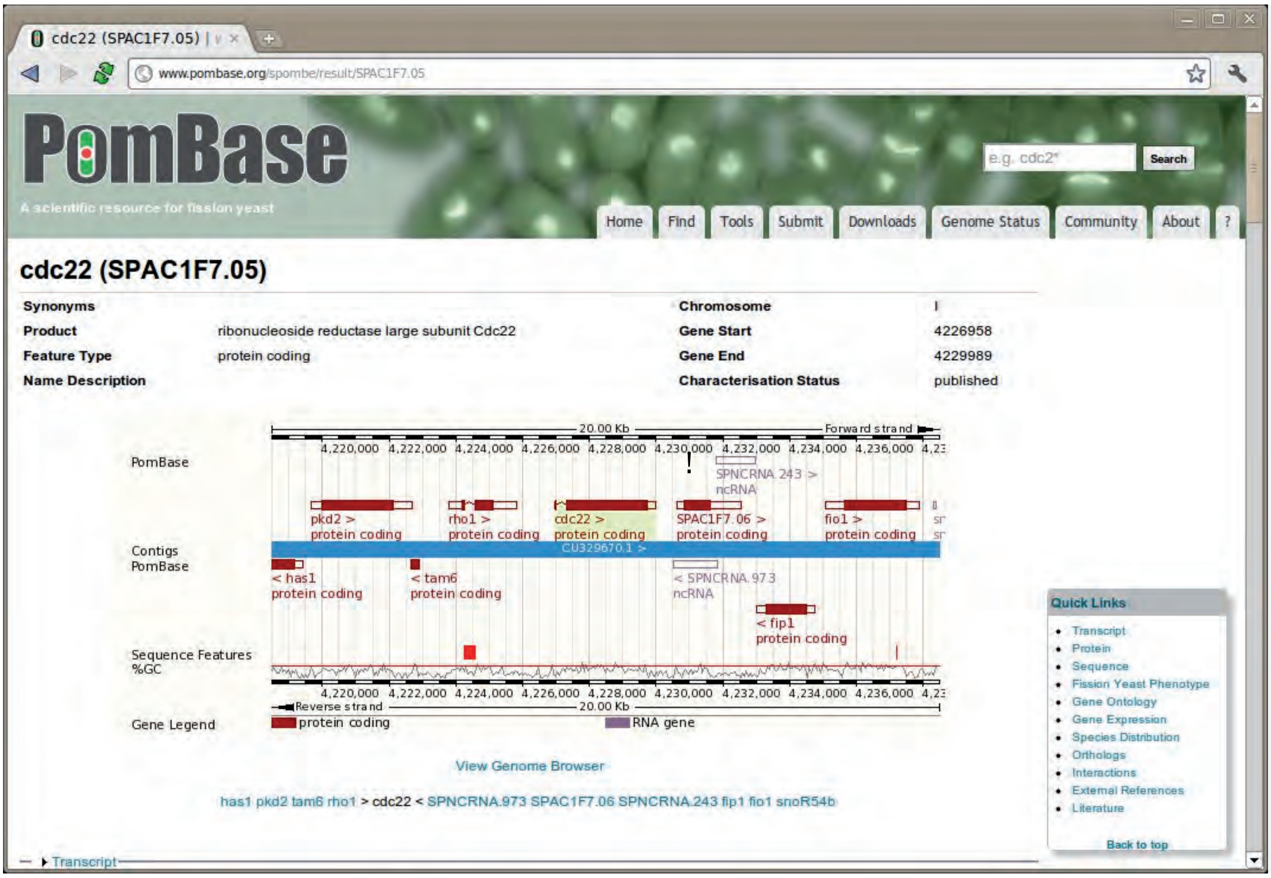
Bateman, A., Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

Earl, D., Bradnam, K., *et al.* (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res.* 21 (12), 2224-41.

Gan, X., Stegle, O., *et al.* (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* 477 (7365), 419-23.

Kinsella, R.J., Kahari, A., *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011, bar030.

Youens-Clark, K., Buckler, E., *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* 39 (Database issue), D1085-94.



cdc22 (SPAC1F7.05)

Synonyms		Chromosome	I
Product	ribonucleoside reductase large subunit Cdc22	Gene Start	4226958
Feature Type	protein coding	Gene End	4229989
Name Description		Characterisation Status	published

has1 pkd2 tam6 rho1 > cdc22 < SPNCRNA.973 SPAC1F7.06 SPNCRNA.243 fip1 fio1 snoR54b

Figure. PomBase, launched in 2011, provides access to genome-scale data for the community working on the fission yeast *Schizoscaahromyces pombe*. Large-scale data is stored and visualised using Ensembl technology, and is integrated within a portal that displays literature-curated data about gene function. A new initiative that we expect to launch in 2012 encourages members of the fission yeast community to curate their own research in the database, with the aid of new online tools already in development. The PomBase data is mirrored within the Ensembl Fungi site.

The Gene Ontology office

Jane Lomax

GO Curation Coordinator. PhD in parasite population genetics, University of Cambridge, 2002. At EMBL-EBI since 2002.

DESCRIPTION OF SERVICES

The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene-product attributes across all species. The aims of the Gene Ontology project are threefold: to maintain and further develop its ontologies of gene and gene product attributes; to annotate genes and gene products, and to assimilate and disseminate annotation data; to provide tools to facilitate access to all aspects of the data provided by the Gene Ontology project. The GO ontologies cover three key biological domains that are shared by all species: the cellular component (the parts of a cell or its extracellular environment); molecular function (the elemental activities of a gene product at the molecular level, e.g. binding or catalysis); and biological process (operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units, e.g. cells, tissues, organs, and organisms).

Groups participating in the GO Consortium include major model organism databases and other bioinformatics resource centres. At EMBL-EBI, the editors play a key role in managing the distributed task of developing and improving the GO ontologies.

SUMMARY OF PROGRESS

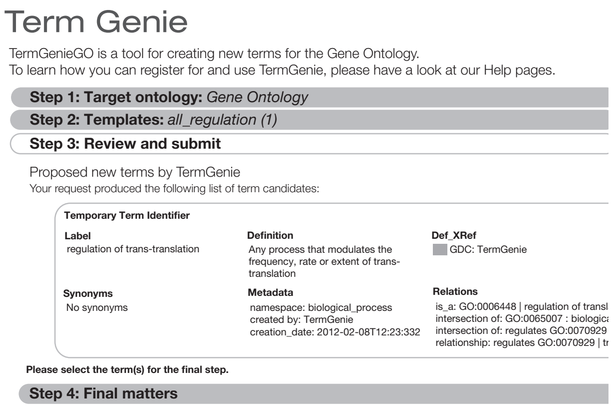
- Refined and improved the logical content of the ontologies;
- Released the TermGenie tool;
- Developed ontology content in the areas of apoptosis, transcription and cardiac conduction.

MAJOR ACHIEVEMENTS

Significant changes introduced to GO in 2011 affect both biological and logical aspects of the ontologies. We have continued to create logical definitions for GO, including those to external ontologies such as chemicals. Our tool for automatic term addition has had an official release, and several term templates are available for public use. We developed ontology content in the areas of apoptosis, transcription and cardiac conduction, and developed a high-level subset of the GO that can be used for all species.

Many GO classes can be defined as being intersections of other terms both within and external to GO. For example, 'glucose metabolic process' is the intersection of GO 'metabolic process' and the ChEBI ontology term 'glucose' (Mungall *et al.*, 2011). We made considerable progress in adding this type of logical definition to GO terms in 2011. This gives us the ability to reason over the ontologies, which has major benefits in terms of maintaining consistency and accuracy and saves us time by allowing our users to automatically add terms to the ontology that conform to certain templates.

In 2011 we started to develop TermGenie, a tool that allows users to add new GO terms that conform to a cross-product template directly to the ontologies. Terms are automatically placed correctly within the ontology, and textual definitions and synonyms are automatically generated. TermGenie was rewritten from the development version and officially released to our user community in 2011 (see Figure 1). It makes available several GO term templates including those for 'regulation', 'involved in' and 'occurs in'.



Term Genie

TermGenieGO is a tool for creating new terms for the Gene Ontology.
To learn how you can register for and use TermGenie, please have a look at our Help pages.

Step 1: Target ontology: *Gene Ontology*

Step 2: Templates: *all_regulation (1)*

Step 3: Review and submit

Proposed new terms by TermGenie
Your request produced the following list of term candidates:

Temporary Term Identifier

Label
regulation of trans-translation

Definition
Any process that modulates the frequency, rate or extent of trans-translation

Def_XRef
GDC: TermGenie

Synonyms
No synonyms

Metadata
namespace: biological_process
created by: TermGenie
creation_date: 2012-02-08T12:23:332

Relations
is_a: GO:0006448 | regulation of transl
intersection of: GO:0065007 : biologica
intersection of: regulates GO:0070929
relationship: regulates GO:0070929 | tr

Please select the term(s) for the final step.

Step 4: Final matters

Figure 1. The TermGenie tool for automatic GO term addition.

The past year has also seen major improvements to the biological content of several areas of the ontologies; apoptosis; transcription and transcription factors; and cardiac conduction. The changes in these areas were developed in collaboration with biological experts. Of particular note was the work relating to apoptosis, which was carried out in collaboration with the APO-SYS Consortium and which culminated in a face-to-face meeting at EMBL-EBI in September 2011.

One of the most powerful uses of GO is to give a high-level view of function for large-scale analyses. To this end, we released a new version of the generic GO slim: a cut-down version of the GO that gives a broad overview for any species. Several more tailored GO slims are also available, including a GO slim specifically for metagenomics analysis.

FUTURE PLANS

In 2012 we expect to make major changes the underlying data-format of our ontologies. Currently, GO is primarily stored and edited in OBO format but this format has proved limiting for some purposes so we hope to move over to using the Web Ontology Language (OWL) for some of our data. We will also be using the OWL ontology editor Protégé 4 to perform some elements of our editing such as reasoning. Improvements begun or continued in 2011 on apoptosis and cardiac conduction and other topics will therefore continue, and we intend to start developing terms in the area of the cell cycle and translation and complete our ongoing project on viral processes.

Selected publications

Gaudet, P., Livstone, M. S., *et al*. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.* 12 (5), 449-62.

Leonelli, S., Diehl, A.D., *et al.* (2011) How the gene ontology evolves. *BMC Bioinformatics* 12, 325.

Mungall, C.J., Bada, M., *et al.* (2011) Cross-product extensions of the Gene Ontology. *J. Biomed. Inform.* 44 (1), 80-6.

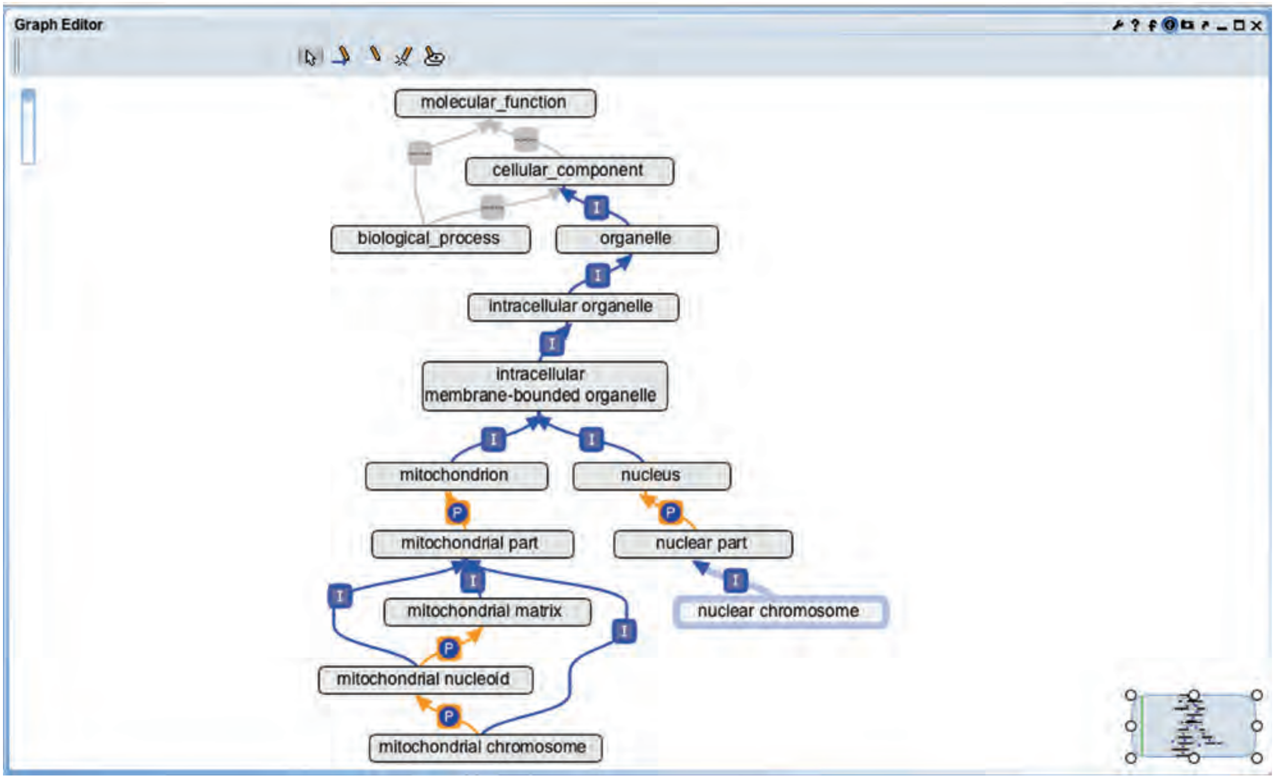


Figure 2. Structure of the Gene Ontology, shown in OBO Edit.

Proteomics services

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.

DESCRIPTION OF SERVICES

The Proteomics Services team develops tools and resources for the representation, deposition, distribution and analysis of proteomics and related data. We follow an open-source, open-data approach: all resources we develop are freely available. The team is a major contributor to the Proteomics Standards Initiative (PSI) of the international Human Proteome Organization (HUPO). We provide reference implementations for the PSI community standards, in particular the PRIDE proteomics identifications database and the IntAct molecular interaction database. We provide the Reactome pathway database in collaboration with New York University and the Ontario Institute for Cancer Research. In the context of the EU RICORDO project, we contribute to the development of an interoperability framework that bridges physiology and molecular biology.

As a result of long-term engagement with the proteomics community, journal editors and funding organisations, proteomics data deposition in PSI-compliant data resources such as IntAct and PRIDE is becoming a strongly recommended part of the publishing process. This has resulted in a rapid increase in the data content of our resources. In addition, the proteomics 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 the Distributed Annotation System (DAS) and web services across a range of projects, including EU Apo-Sys, LipidomicNet, SLING, and the University of California Los Angeles (UCLA)–National Heart, Lung, and Blood Institute (NHLBI/NIH) Proteomics Center.

SUMMARY OF PROGRESS

- Co-ordinated the implementation of the PSICQUIC molecular interaction query interface by more than 20 organisations (Aranda *et al.*, 2011);
- Contributed with both IntAct and Reactome to a high-impact study on human platelet formation (Gieger *et al.*, 2011);
- Contributed to community standard documents on mass spectrometry (Martens *et al.*, 2011) and bioactive entity (Orchard *et al.*, 2011) data representation;
- Oversaw the growth of the Proteomics Identifications Database (PRIDE) to more than 250 million mass spectra;
- Published a series of open-source components for complete DAS infrastructure (Villaveces *et al.*, 2011).

MAJOR ACHIEVEMENTS

The PSI Molecular Interactions workgroup collaborates with several key molecular interaction data providers to synchronise their curation efforts and provide non-redundant datasets that are curated to common standards. IntAct is an active member of the International Molecular Exchange consortium (IMEX), which started full production mode and released a common website in February 2010. Other members of the consortium include DIP (University of California Los Angeles, US), MINT (University of Rome, Italy), MatrixDB (University of Lyon, France), Molecular Connections Inc. and MPIDB (J. Craig Venter Institute, US).

Based on the PSI molecular interaction standards, we developed the PSI Common Query Interface (PSICQUIC), a common query API for molecular interaction databases. PSICQUIC was published in 2011 and has seen rapid community adoption, currently providing access to more than 30 million binary interaction evidences from 24 different sources, including protein-protein interactions, protein–small molecule interactions, and simplified pathway data (Aranda *et al.*, 2011).

Members of the IntAct and Reactome projects contributed to a high-impact meta-analysis of several genome-wide association studies (GWAS) comprising approximately 67 000 individuals that identified 53 new genomic loci reliably associated with human platelet count and volume. We contributed advice and metadata leading to a protein-interaction network based on Reactome-derived interactions extended with interactions from IntAct to include most of the novel genes. Functional experiments in zebrafish and fruitflies show that 11 of the newly identified genes encode novel regulators of blood cell formation. The results serve as an excellent example of how GWAS results can be combined with pathways to understand function and suggest new targets for the treatment of haematological disorders (Gieger *et al.*, 2011).

In collaboration with major proteomics data providers (e.g. PeptideAtlas, Peptidome, UniProt, University of Ghent, University of Liverpool, ETH Zurich, University of Michigan, Wiley-VCH) we developed a concept for regular proteomics data exchange between key repositories. The resulting ProteomeXchange EU grant started in January 2011. PRIDE, as one of the major resources of the ProteomeXchange consortium, more than doubled in size in 2011, growing to 270 million spectra. A new tool, PRIDE Inspector, was released, providing a much improved interface to PRIDE data for data submitters, curators, reviewers and the scientific community at large.

FUTURE PLANS

After rapid development and achievement of major milestones in the molecular interaction domain, we will consolidate our achievements, selectively open the IMEX collaboration to new partners and develop advanced tools to take advantage of detailed IMEX curation and the integrative PSICQUIC interface. A major challenge will be the complete redevelopment of the PRIDE database, which is necessary if we are to cope with the rapid increase in data content and will transform PRIDE from a publication-centric repository to a key source for protein-expression information. Beyond the technical challenges of data quantity, the two major conceptual challenges are to capture the very diverse quantitative proteomics data and to develop quality criteria to enable the selective export of high-confidence PRIDE data to other resources like UniProt, Reactome or integrative data-analysis tools. We plan to intensify data integration within and beyond the projects of the Proteomics Services team, in particular using web services and DAS. We will also continue to integrate Reactome pathways and IntAct molecular interactions, as well as integrating PRIDE and IntAct, to enable efficient data deposition and navigation between molecular interactions and underlying mass spectrometry data. We will continue our successful collaboration with all PSI partners, in particular with journals and editors, to encourage data producers to make their data available to the community through public databases by utilising community-supported standards.

Selected publications

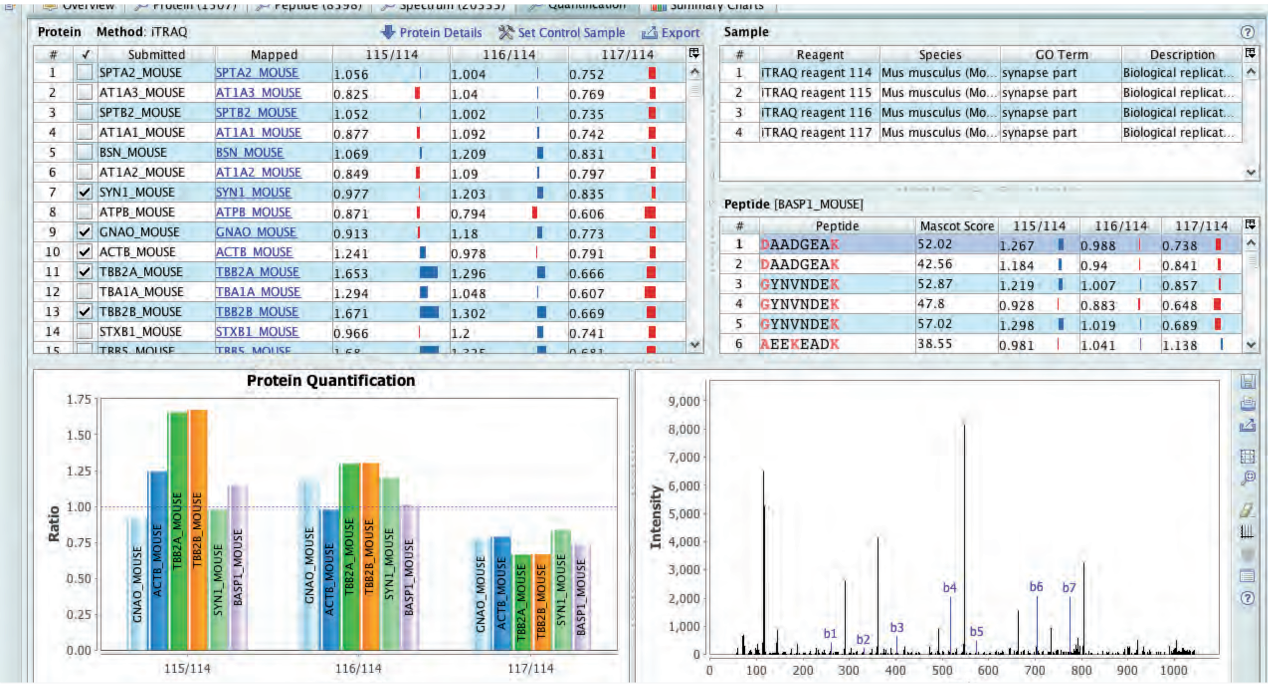
Aranda, B., Blankenburg, H., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8 (7), 528-9.

Gieger, C., Radhakrishnan, A., *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature* 480, 201–208.

Martens, L., Chambers, M., *et al.* (2011) mzML - A community standard for mass spectrometry data. *Mol. Cell. Proteomics* 10 (1), R110.000133.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

Villaveces, J.M., Jimenez, R.C., *et al.* (2011) Dasty3, a WEB Framework for DAS. *Bioinformatics* 27 (18), 2616-7.



Protein **Method**: iTRAQ

#	✓	Submitted	Mapped	115/114	116/114	117/114
1	□	SPTA2_MOUSE	SPTA2_MOUSE	1.056	1.004	0.752
2	□	AT1A3_MOUSE	AT1A3_MOUSE	0.825	1.04	0.769
3	□	SPTB2_MOUSE	SPTB2_MOUSE	1.052	1.002	0.735
4	□	AT1A1_MOUSE	AT1A1_MOUSE	0.877	1.092	0.742
5	□	BSN_MOUSE	BSN_MOUSE	1.069	1.209	0.831
6	□	AT1A2_MOUSE	AT1A2_MOUSE	0.849	1.09	0.797
7	☑	SYN1_MOUSE	SYN1_MOUSE	0.977	1.203	0.835
8	□	ATPB_MOUSE	ATPB_MOUSE	0.871	0.794	0.606
9	☑	GNAO_MOUSE	GNAO_MOUSE	0.913	1.18	0.773
10	☑	ACTB_MOUSE	ACTB_MOUSE	1.241	0.978	0.791
11	☑	TBB2A_MOUSE	TBB2A_MOUSE	1.653	1.296	0.666
12	□	TBA1A_MOUSE	TBA1A_MOUSE	1.294	1.048	0.607
13	☑	TBB2B_MOUSE	TBB2B_MOUSE	1.671	1.302	0.669
14	□	STXB1_MOUSE	STXB1_MOUSE	0.966	1.2	0.741
15	□	TBB5_MOUSE	TBB5_MOUSE	1.68	1.325	0.681

Sample

#	Reagent	Species	GO Term	Description
1	iTRAQ reagent 114	Mus musculus (Mo...	synapse part	Biological replicat...
2	iTRAQ reagent 115	Mus musculus (Mo...	synapse part	Biological replicat...
3	iTRAQ reagent 116	Mus musculus (Mo...	synapse part	Biological replicat...
4	iTRAQ reagent 117	Mus musculus (Mo...	synapse part	Biological replicat...

Peptide [BASP1_MOUSE]

#	Peptide	Mascot Score	115/114	116/114	117/114
1	DAADGEAK	52.02	1.267	0.988	0.738
2	DAADGEAK	42.56	1.184	0.94	0.841
3	GYNVNDEK	52.87	1.219	1.007	0.857
4	GYNVNDEK	47.8	0.928	0.883	0.648
5	GYNVNDEK	57.02	1.298	1.019	0.689
6	AEEKEADK	38.55	0.981	1.041	1.138

Figure. Protein quantification view in PRIDE Inspector.

InterPro

Sarah Hunter

*MSc University of Manchester, 1998. Pharmaceutical and Biotech Industry (Sweden), 1999-2004.
At EMBL-EBI since 2005.*

DESCRIPTION OF SERVICES

Our team coordinates the InterPro and Metagenomics projects at EMBL-EBI.

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) into a single resource. 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. Additional biological annotation is included, together with links to external databases such as GO, PDB, SCOP and CATH. InterPro 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.

Metagenomics is the study of the sum genetic material found in an environmental sample or host species, typically using next-generation sequencing (NGS) technology. The Metagenomics Portal, 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 and the results generated are visualised via a web interface.

SUMMARY OF PROGRESS

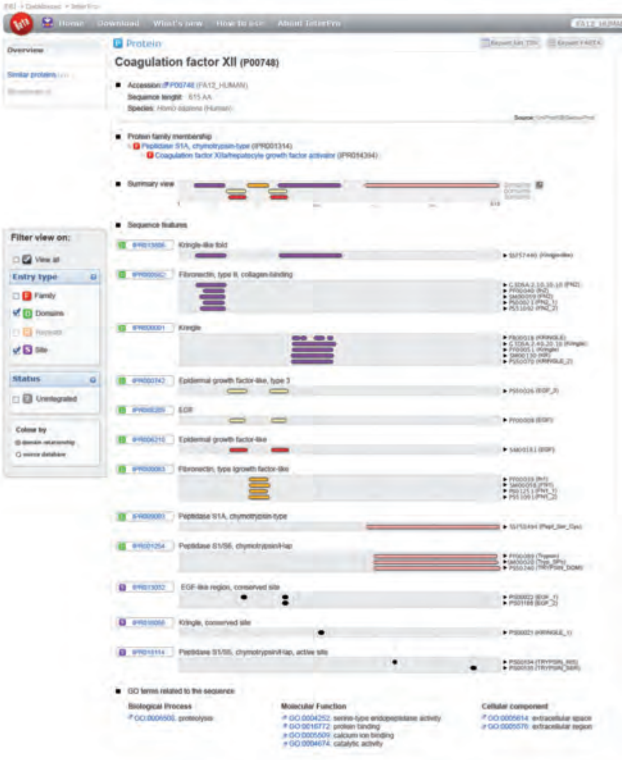
- Issued five major releases of the InterPro database: created 2122 new entries and integrated 1176 signatures;
- Issued two new releases of InterProScan, a Perl-based signature-scanning software;
- Migrated InterPro to Hmmer3 for TIGRFAMs and Gene3D. Upgraded Gene3D results processing to use DomainFinder3 algorithm;
- Redesigned the InterPro website and released in beta;
- Launched the Metagenomics Portal in beta;
- Adapted Metagenomics data-submission tools and created pipelines for data analysis;
- Retired the CluSTr resource, which was previously maintained by the InterPro team, in June 2011.

MAJOR ACHIEVEMENTS

During 2011 development of the InterPro beta website continued, with particular focus on the representation of matches of InterPro's signatures to protein sequences. Previously, the graphical output from InterProScan differed markedly from the InterPro website's view of the same data. The new display now looks identical, regardless of the source, giving users a more consistent experience. Other features on the site include a cleaner aesthetic and design; a new home page; better delineated entry pages that are split into sub-sections; and updated documentation to reflect these changes.

InterPro curators continued to add content to the database, and integrated over 1000 signatures during 2011. The most recent release (v35.0) has seen an improvement in coverage: 95.4% of UniProtKB/Swiss-Prot proteins match at least one InterPro entry, 79.2% of UniProtKB/TrEMBL and 79.7% overall for UniProtKB (Swiss-Prot and TrEMBL). A major focus of curation work during 2011 was improving the process for Gene Ontology term mapping to InterPro. Links to pathway databases such as KEGG and Reactome were added automatically where appropriate.

Two additional InterPro member databases, TIGRFAMs and Gene3D, were updated to use HMMER3, following Pfam's adoption of the algorithm in 2010. As with Pfam, it was necessary to check the validity of existing integrations as HMMER3 can more sensitively detect remote homologs compared to its predecessor. The algorithm was updated in InterProScan for these two resources.



Coagulation factor XII (P00748)

- Accession: P00748 (FA12_HUMAN)
  Sequence length: 615 AA
  Species: Homo sapiens (Human)
- Protein family membership
  Peptidase S1A, chymotrypsin-type (IPR001314)
  Coagulation factor XIIa/hepatocyte growth factor activator (IPR014394)
- Summary view
- Sequence features
  Kringle-like fold
  Fibronectin, type II, collagen-binding
  Kringle
  Epidermal growth factor-like, type 3
  EGF
  Epidermal growth factor-like
  Fibronectin, type I/growth factor-like
  Peptidase S1A, chymotrypsin-type
  Peptidase S1/S6, chymotrypsin/Hap
  EGF-like region, conserved site
  Kringle, conserved site
  Peptidase S1/S6, chymotrypsin/Hap, active site
- GO terms related to the sequence
  Biological Process
  GO:0006508 proteolysis
  Molecular Function
  GO:0004252 serine-type endopeptidase activity
  GO:0016772 protein binding
  GO:0005509 calcium ion binding
  GO:0004674 catalytic activity
  Cellular component
  GO:0005614 extracellular space
  GO:0005576 extracellular region

Figure 1. View of the domains and sites in a protein (FA12_HUMAN) on the new InterPro beta site.

InterProScan5 was developed to utilise 'best-in-breed' Java technologies to improve the robustness of the software compared to the previous version. InterProScan5 outputs a descriptive XML format, which can then be transformed into a graphical, tab-delimited or GFF3 representation of results. If a user searches using a nucleotide sequence, InterProScan will translate the sequence into six open reading frames. The software then maps protein domain and family results back to the original DNA sequence once they have been calculated. Users may also request a look-up of the Gene Ontology terms and pathways associated with the InterPro entries matching their sequence(s). InterProScan5 contains an additional algorithm called Phobius, which is used to predict the presence of signal peptides and transmembrane regions in proteins.

Selected publications

Hunter, C., *et al.* (2011) The EBI metagenomics archive, integration and analysis resource. In: Frans J. de Bruijn, Ed, *Handbook of Molecular Microbial Ecology I: Metagenomics and complementary approaches*. Wiley-Blackwell.

Hunter, S., Jones, P., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40 (Database issue), D306-12.

Jones, P., Binns, D., *et al.* (2011) The InterPro BioMart: federated query and web service access to the InterPro Resource. *Database* (Oxford) 2011, bar033.

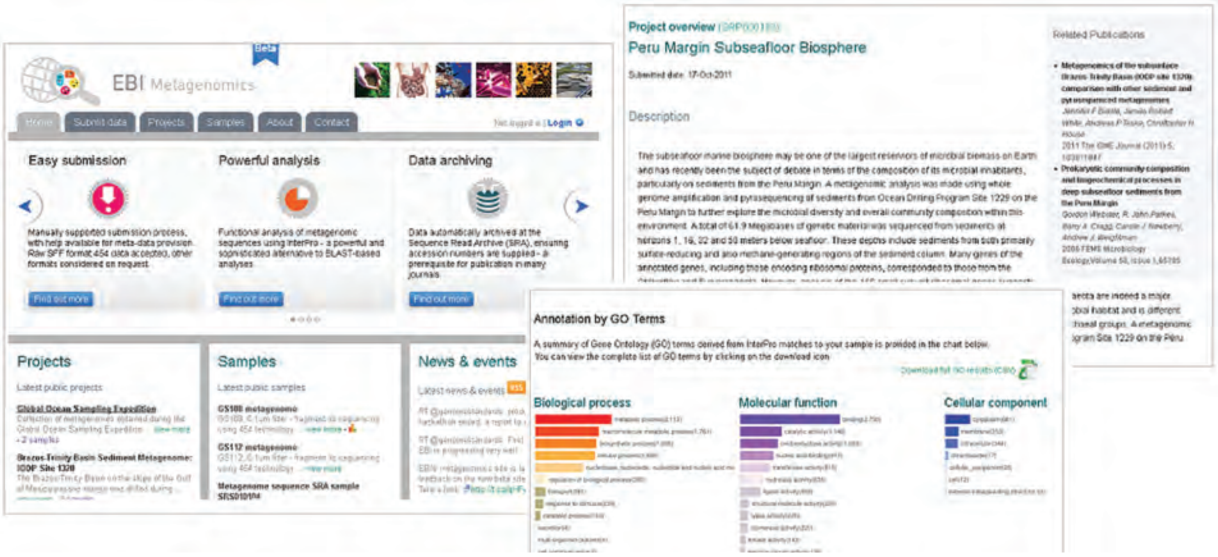
McDowall, J. and Hunter, S. (2011) InterPro Protein Classification. *Methods Mol. Biol.* 694, 37-47.

The Metagenomics Portal was launched in beta in late 2011, offering a number of public datasets for browsing. The initial version was designed to assist researchers in submitting, organising and analysing their metagenomic datasets. We built an analysis pipeline comprising quality control, clustering and filtering steps, followed by an InterPro-based functional characterisation step (which includes the association of GO terms). We also created an interface that enables users to submit their raw sequence and sample metadata to the European Nucleotide Archive (ENA) and retrieve any subsequent analysis results. Data may be held privately prior to publication, although policy dictates that all data must eventually be made available in the public domain.

FUTURE PLANS

The InterPro website (currently in beta) will be launched to the public in early 2012; InterProScan5 will be released concurrently. InterPro is expected to be served from EMBL-EBI's London data centres by the end of 2012.

Further developments are planned for the Metagenomics Portal. Existing data pipelines will be enhanced by the addition of 16S marker gene analysis software so that the taxonomic diversity of metagenomics samples can be estimated. Submission tools will similarly be improved to increase the ease with which data can be uploaded for analysis and archiving. New visualisation tools will be created that allow researchers to compare the functional and taxonomic profiles of different metagenomics samples in an intuitive way. We intend to make these improvements by working closely with target users and by extensive usability testing.



Easy submission

Manually supported submission process, with help available for meta-data provision. Raw SFF format 454 data accepted, other formats considered on request.

Powerful analysis

Functional analysis of metagenomic sequences using InterPro - a powerful and sophisticated alternative to BLAST-based analyses.

Data archiving

Data automatically archived at the Sequence Read Archive (SRA), ensuring accession numbers are supplied - a prerequisite for publication in many journals.

Projects

Latest public projects

Global Ocean Sampling Expedition

Brazos-Trinity Basin Sediment Metagenome: IODP Site 1320

Samples

Latest public samples

GS108 metagenome

GS112 metagenome

Metagenome sequence SRA sample SRS010104

News & events

Latest news & events

Project overview

Peru Margin Subseafloor Biosphere

Submitted date: 17-Oct-2011

Description

The subseafloor marine biosphere may be one of the largest reservoirs of microbial biomass on Earth and has recently been the subject of debate in terms of the composition of its microbial inhabitants, particularly on sediments from the Peru Margin. A metagenomic analysis was made using whole genome amplification and pyrosequencing of sediments from Ocean Drilling Program Site 1229 on the Peru Margin to further explore the microbial diversity and overall community composition within this environment. A total of 61.9 Megabases of genetic material was sequenced from sediments at horizons 1, 16, 32 and 50 meters below seafloor. These depths include sediments from both primarily sulfate-reducing and also methane-generating regions of the sediment column. Many genes of the annotated genes, including those encoding ribosomal proteins, corresponded to those from the

Related Publications

- Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes
- Prokaryotic community composition and biogeochemical processes in deep subseafloor sediments from the Peru Margin

Annotation by GO Terms

A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the chart below. You can view the complete list of GO terms by clicking on the download icon.

Biological process

Molecular function

Cellular component

Figure 2. The Metagenomics Portal home page, an example of a page describing a metagenomics project and a summary of the annotation of a sample's sequences using the Gene Ontology.

The Universal Protein Resource (UniProt): Development

Maria Martin

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

DESCRIPTION OF SERVICES

Our team provides the bioinformatics infrastructure for the protein-related databases and services of UniProt. We are also responsible for the maintenance and development of tools for UniProt curation. UniProt provides the scientific community with a centralised repository of protein sequences and functional annotation.

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.

SUMMARY OF PROGRESS

- Analysed the needs of UniProt users who access our resource in a variety of contexts, and used this information to create a strategy for optimising our services;
- Offered many new data sets of interest to our user community. Some of these were reference proteomes and datasets for species that were previously distributed via the International Protein Index (IPI);
- Implemented annotation tools (UniRule, Gene Ontology, proteome editors) to support curation of the resources;
- In collaboration with Ensembl and PDBe, extended the data-import infrastructure to achieve consensus sequence annotation;
- Consolidated software and extended the databases to accommodate a rapidly growing volume of data.

MAJOR ACHIEVEMENTS

Leadership of the UniProt Development team was strengthened in 2011 by the appointment of Jie Luo as Project Leader of Database Operations.

The UniProt website facilitates the search, identification and analysis of gene products. To evaluate its usability and understand users' needs, we organised two interactive workshops and a number of website reviews. User-experience (UX) testing involved gathering users' views about what they expect from UniProt as well as observing and quantifying their interactions with the website. Analysis of these data revealed a number of areas for improvement; the team accordingly created mock-ups and prototypes to take back to the users. Several iterations of this exercise have resulted in substantial improvements to the usability and data representation of UniProt.

In the interests of facilitating communication with users, we explored social media (primarily Twitter and Facebook) and set out an engagement strategy to ensure that the information in UniProt is useful to a broad research community with diverse scientific interests, requirements and levels of computational expertise.

Following the closure of the International Protein Index (IPI), UniProt pledged to provide complete proteome datasets for all species covered by IPI that are particularly relevant for the proteomics community. This required new pipelines for the generation and distribution of frequently requested eukaryotic genomes (e.g. *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Canis familiaris*, *Danio rerio*, *Gallus gallus*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*).

The number of complete genomes sequenced rose dramatically in 2011, and the team worked hard to organise and represent over 450 reference proteomes. The species selected provide broad coverage of the 'Tree of Life' as well as well studied model organisms and other proteomes of interest.

In 2011 the team also update the interface of QuickGO, a fast, web-based browser that provides access to all information about Gene Ontology (GO) terms and the GO annotations released by the UniProtKB-GOA team. It is now more user-friendly and intuitive, and displays additional information such as a 'change log' for individual GO terms (and the ontology as a whole); a display of taxon constraints that apply to a GO term; and links to guidelines for use in annotation.

We extended the production pipeline for UniRule, a system for automatic annotation of a large volume of uncharacterised proteins. UniRule is of paramount importance to the UniProt Consortium; the pipeline now integrates automatic annotation systems from the three consortium members. This has already led to a higher number of automatic annotations in UniProtKB. We also implemented new interfaces for the UniRule curation tool that make it easier for curators to manage prediction rules and perform statistical assessment of existing and new rules.

We developed a proteome annotation platform that helps curators monitor completely sequenced genomes – and handle annotations of their encoded proteins – much more efficiently. We started the process of redesigning the production pipelines and database back-end in order to cope with the critical tracking and annotation of a growing number of genome/proteomes. We also developed a proteome editor, which acts as an annotation platform for the organisation and management of information related to a set of proteins encoded by completely sequenced genomes.

The team develops the web-based Protein2GO tool, which UniProt curators use to contribute annotations to the GOA project. We extended this tool in 2011 to ensure that it is in line with current GO Consortium guidelines.

Selected publications

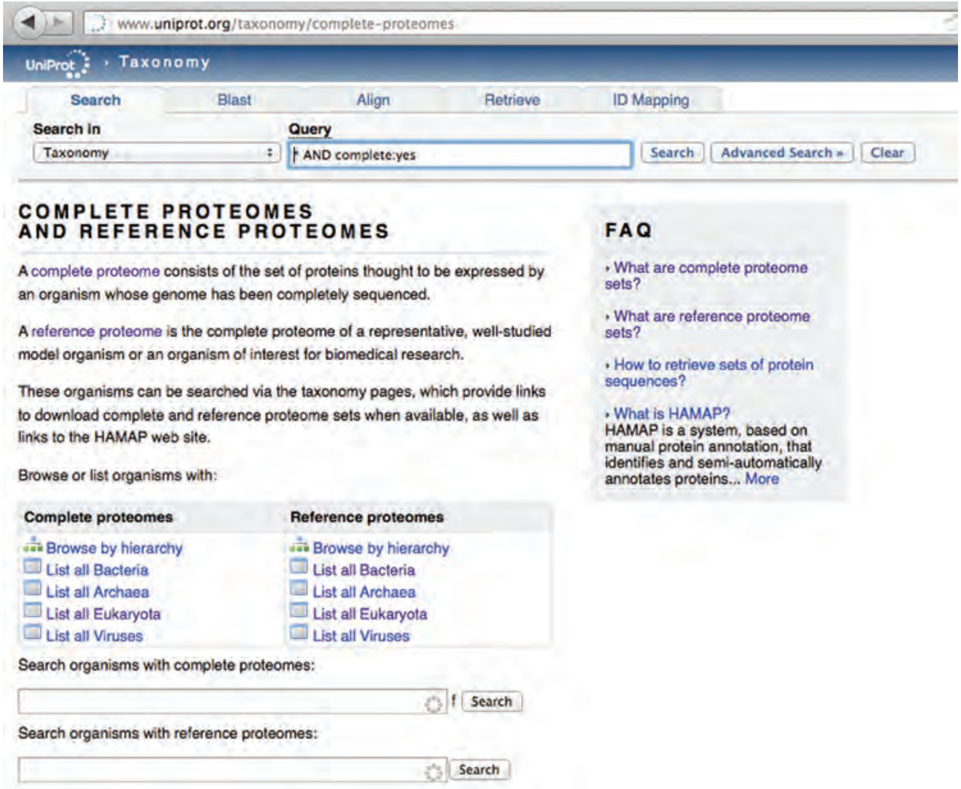
Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the international protein index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics* 11 (22), 4434–4438.

Magrane, M. and UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, bar009.

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39 (Database issue), D214-9.

FUTURE PLANS

We will extend the rule-annotation tool in to allow the prediction of sequence-related features in 2012. We will continue to explore data-exchange mechanisms, and will generate a UniRule XML format for public distribution to assist the annotation communities in applying annotation rules; this should also provide the means to annotate within UniProt. We will develop new interfaces for the UniProt website according to the needs of our users, and will facilitate easy access to reference and complete proteomes. We will continue to focus on usability issues and engage with our users to ensure we maintain a global genome/proteome- and gene-product-centric view of the sequence space, and 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 cooperate with diverse data providers (e.g. Ensembl, RefSeq, PRIDE) to complete protocols for the exchange and integration of relevant information in UniProt.



Search in: Taxonomy

Query: AND complete:yes

COMPLETE PROTEOMES AND REFERENCE PROTEOMES

A complete proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

A reference proteome is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP web site.

Browse or list organisms with:

Complete proteomes	Reference proteomes
Browse by hierarchy	Browse by hierarchy
List all Bacteria	List all Bacteria
List all Archaea	List all Archaea
List all Eukaryota	List all Eukaryota
List all Viruses	List all Viruses

Search organisms with complete proteomes:

Search organisms with reference proteomes:

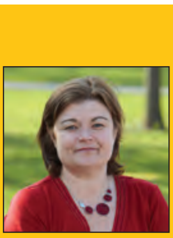
FAQ

- What are complete proteome sets?
- What are reference proteome sets?
- How to retrieve sets of protein sequences?
- What is HAMAP?

HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins... More

Figure. The UniProt Taxonomy homepage.

The Universal Protein Resource (UniProt): Content



Claire O'Donovan

BSc (Hons) in Biochemistry, 1992, University College Cork, Ireland. Diploma in Computer Science, 1993, University College Cork, Ireland. At EMBL since 1993, at EMBL-EBI since 1994. Team Leader since 2009.

DESCRIPTION OF SERVICES

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

The work of our 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.

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 data sets. Accurate and comprehensive representation of biological knowledge, as well as easy access to this data for working scientists and a basis for computational analysis, are primary goals of biocuration.

UniProt manual curation: Manual curation involves a critical review of experimental and predicted data for each protein, as well as manual verification of each protein sequence. The curation methods we apply to UniProtKB/Swiss-Prot include manual extraction and structuring of information from the literature, manual verification of results from computational analyses, mining and integration of large-scale data sets and continuous updating as new information becomes available.

UniProt automatic annotation: UniProt has developed two complementary approaches in order to automatically annotate 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 InterPro classification and the manually curated data in UniProtKB/Swiss-Prot.

UniProt GO annotation (GOA): The UniProt GO annotation (GOA) program aims to add high-quality GO annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt biocuration. We supplement UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups. This ensures that users have a comprehensive GO annotation dataset. UniProt-GOA is a member of the GO consortium.

SUMMARY OF PROGRESS

- Continued to manually annotate UniProtKB, with a particular focus on the human and other reference proteomes. Collaborated closely with other resources to ensure comprehensiveness and mutually beneficial exchange of data;
- Substantially progressed automatic annotation efforts, leading the consolidation of the rule systems of all participating consortium members and undertaking the extension of these efforts to external collaborating groups;
- Increased manual and electronic GO annotation efforts: as of December 2011 there were 108 million GO annotations for 12.6 million UniProtKB entries, covering more than 396 000 taxonomic groups.

www.uniprot.org/program | www.uniprot.org/program/automatic_annotation | www.ebi.ac.uk/GOA

MAJOR ACHIEVEMENTS

As a core contributor to the Consensus CDS (CCDS) project, UniProt now has 18 109 manually curated human entries of the total 20 242 records that are in synch with the RefSeq annotation group (National Center for Biotechnology Information, NCBI) and the Ensembl and HAVANA teams (EMBL-EBI and the Wellcome Trust Sanger Institute). This effort was extended to ensure a curated and complete synchronisation with the HUGO Gene Nomenclature Committee (HGNC), which has assigned unique gene symbols and names to more than 32 000 human loci (over 19 000 of these are listed as coding for proteins).

We played a major role in the establishment of 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 world-wide These standards have contributed significantly to the annotation of complete genomes and proteomes, and are helping the scientific community exploit these data to their full potential. Our group continues to develop and promote these standards widely.

Our team aims to provide a 'gold standard' dataset that will enable users to identify all experimental data for a given protein from a particular strain of a given organism, as well as all experimentally characterised annotations/ proteomes from a proteome or protein family. To that end, we have developed annotation standards for the demerging of existing UniProtKB/Swiss-Prot entries and extended the scope of the evidence-attribution system. This work was done in collaboration with model organism databases and the Evidence Code Ontology (ECO).

EMBL-EBI completed the consolidation of the consortium members' three automatic annotation systems in 2011, with the exception of the sequence-feature annotation (expected to be completed in summer 2012). This enabled the UniProt automatic annotation system to respond successfully to the exponential growth of UniProtKB/TrEMBL, and to achieve a widening of the taxonomic and annotation depth. Our curators are key members of the GO Consortium Reference Genomes Initiative for the human proteome and provide high-quality annotations to human proteins. The GOA renal project has been very successful, providing 2204 proteins with 34 078 annotations and resulting in the creation of 479 new GO terms (1.34% of the whole of GO). The electronic GO annotation pipeline was reviewed and improved, with particular focus on other UniProt controlled vocabularies such as subcellular location, taxonomic range and InterPro member databases.

Selected publications

Magrane, M. and UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, bar009.

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39 (Database issue), D214-9.

Alam-Faruque, Y., Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB 'complete proteome' sets. *Proteomics* 11 (22), 4434-8.

Klimke, W., O'Donovan, C., *et al.* (2011) Solving the problem: Genome annotation standards before the data deluge. *Stand. Genomic Sci.* 5 (1), 168-93.

O'Donovan, C. and Apweiler, R. (2011) A guide to UniProt for protein scientists. *Methods Mol. Biol.* 694 25-35.

FUTURE PLANS

In 2012 we will work to provide a 'gold standard' data set. In the past, UniProtKB/Swiss-Prot merged 100% identical protein sequences from different genes in the same species into a single record; however, with the recent exponential increase in the the availability and usage of genomic information, UniProtKB has modified its merging policy. We will demerge entries containing multiple individual genes coding for 100% identical protein sequences into individual UniProtKB/Swiss-Prot. This will give a gene-centric view of protein space, and will allow a cleaner and more logical mapping of gene and genomic resources to UniProtKB. We also plan to extend our nomenclature collaborations to include higher-level organisms. We will continue to prioritise the extraction of experimental data from the literature and expect to extend our use of data mining methods to identify scientific literature of particular interest with regards to our annotation priorities. We are committed to expanding UniRule by adding feature annotation and extending the number and range of rules with additional curator resources. We will be in a position to provide a resource that will allow our external affiliates to contribute to the UniRule effort, and that allows our rules to be used in their pipelines. The SAAS approach will be fully reviewed in 2012 in order to incorporate more recent developments in this field.

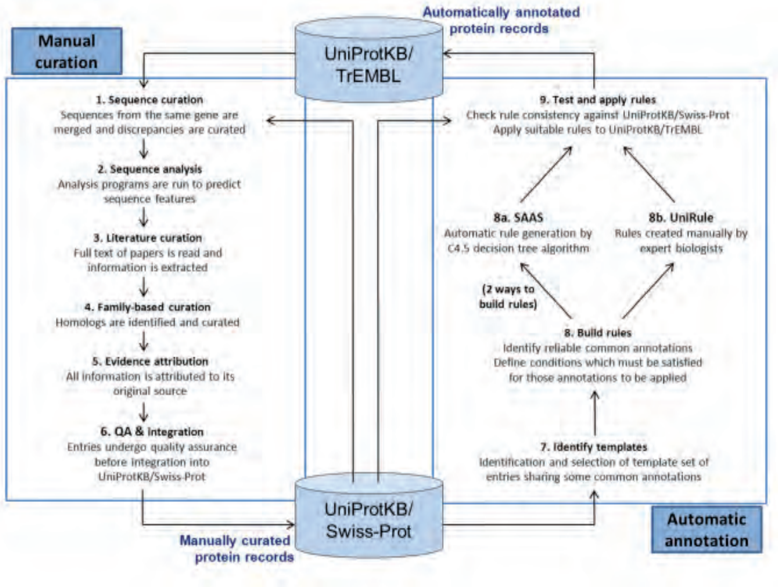

Automatically annotated protein records
Manual curation
UniProtKB/ TrEMBL
1. Sequence curation
Sequences from the same gene are merged and discrepancies are curated
2. Sequence analysis
Analysis programs are run to predict sequence features
3. Literature curation
Full text of papers is read and information is 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
8a. SAAS
Automatic rule generation by C4.5 decision tree algorithm
8b. UniRule
Rules created manually by expert biologists
(2 ways to build rules)
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
UniProtKB/ Swiss-Prot
Manually curated protein records
Automatic annotation

Figure. Organisation of data and information in the Universal Protein Resource.

ChEMBL: a database of bioactive drug-like small molecules

John Overington

PhD in Crystallography, Birkbeck College, London, 1991. Postdoctoral research, ICRF, 1990-1992. Pfizer 1992-2000. Inpharmatica 2000-2008. At EMBL-EBI since 2008.

DESCRIPTION OF SERVICES AND RESEARCH

The ChEMBL group develops and manages EMBL-EBI's database of quantitative small molecule bioactivity data. Synthetic small molecules form the majority of life saving drugs, and the discovery of safe and efficacious new drugs acting through novel biological mechanisms is a major goal for life science research. The ChEMBL database stores curated two-dimensional chemical structures and abstracted bioactivity data (i.e. binding constants, pharmacology and ADMET data) alongside calculated molecular properties. ChEMBL data are abstracted and curated from the primary scientific literature and cover a significant fraction of the structure–activity relationship data for the discovery of modern drugs. Our research interests focus on applying ChEMBL data to drug-discovery challenges.

SUMMARY OF PROGRESS

- Integrated key EMBL-EBI resources with ChEMBL data;
- Established a robust curate–update cycle for new literature data;
- Extended the deposition and accessibility of neglected disease screening datasets;
- Provided programmatic web service access to ChEMBL data;
- Integrated PubChem bioassay data into ChEMBL, and developed a number of Open Data sharing and standards initiatives;
- Extended ChEMBL data model to address ligand-binding domain, and biological drug data;
- Broad impact across the global research community, with many papers published using ChEMBL data in the development of novel methods for virtual screening, target prediction and network analysis.

MAJOR ACHIEVEMENTS

ChEMBL data content expanded considerably in 2011, and now has over one million distinct compound structures and more than five million experimental bioactivities. This increase is thanks both to a 50% increase in the core, literature-derived data and to the loading of comparable data from the NIH Molecular Libraries screening data. Together, these factors substantially expanded the diversity and density of chemical and target space covered by ChEMBL.

Usage of ChEMBL's web interface grew approximately three-fold in 2011, and interface enhancements were largely driven by feedback from the user community. Downloads of the entire database and integration with other systems continued to be strong. We continued our successful webinar series, which focuses on user interface, web services and the ChEMBL data model. This was complemented by a series of training courses, both on-site in Hinxton and at host institutes throughout Europe.

Extending the chemical content of ChEMBL and integrating it with other chemistry resources brought to light issues such as database synchronisation, complex loading dependencies and differences in representation standards of small molecules. In response, we developed a fast, highly scalable chemical-structure integration system called UniChem. UniChem allows us to integrate and query across other chemistry resources, both within the EBI and across other global chemistry resources, while keeping our curation and focus on the core ChEMBL data. We plan to offer UniChem as an open service for the global community following further testing.

Confidentiality of search structures is often a key concern in drug discovery; accordingly, all ChEMBL services run under the secure, industry standard https: Internet protocol. We extended our technical infrastructure in 2011 to include the upload capability and controlled sharing of private, pre-publication data.

One of the most important applications of the data contained within ChEMBL is in the assessment and scoring of genes and proteins as targets for drug discovery. We have continued to implement an open infrastructure for large-scale scoring of targets for their 'druggability'; our structure-based scoring of binding-site data for complementarity to drug-like small molecules is well used.

Our research interests focus on applying the ChEMBL data to drug-discovery challenges. One major area is the analysis of the properties of successful peptide-derived drugs in which physicochemical properties are very different to those of classic Lipinski-like synthetic molecules. We mapped the complete amino-acid monomer space derived from ChEMBL

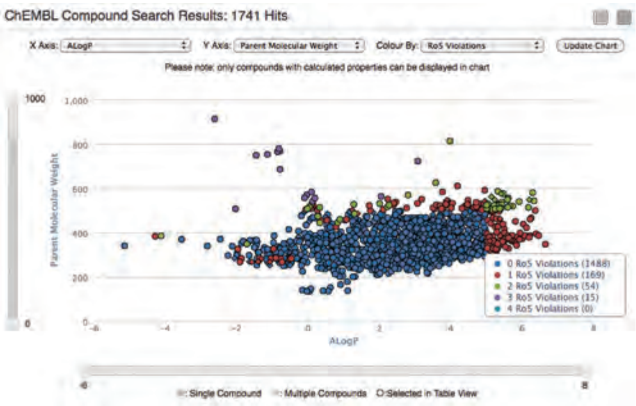


Figure 1. Molecular property space plot of bioactive compounds from ChEMBL.

Selected publications

Gaulton, A., Bellis, L.J., *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40 (Database issue), D1100-7.

Gleeson, M.P., Hersey, A., *et al.* (2011) Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* 10 (3), 197-208.

Krüger, F.A. and Overington, J.P. (2012) Global analysis of small molecule binding to related protein targets. *PLoS Comput. Biol.* 8 (1), e1002333.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

peptides, and then developed QSAR approaches to suggest optimisation strategies.

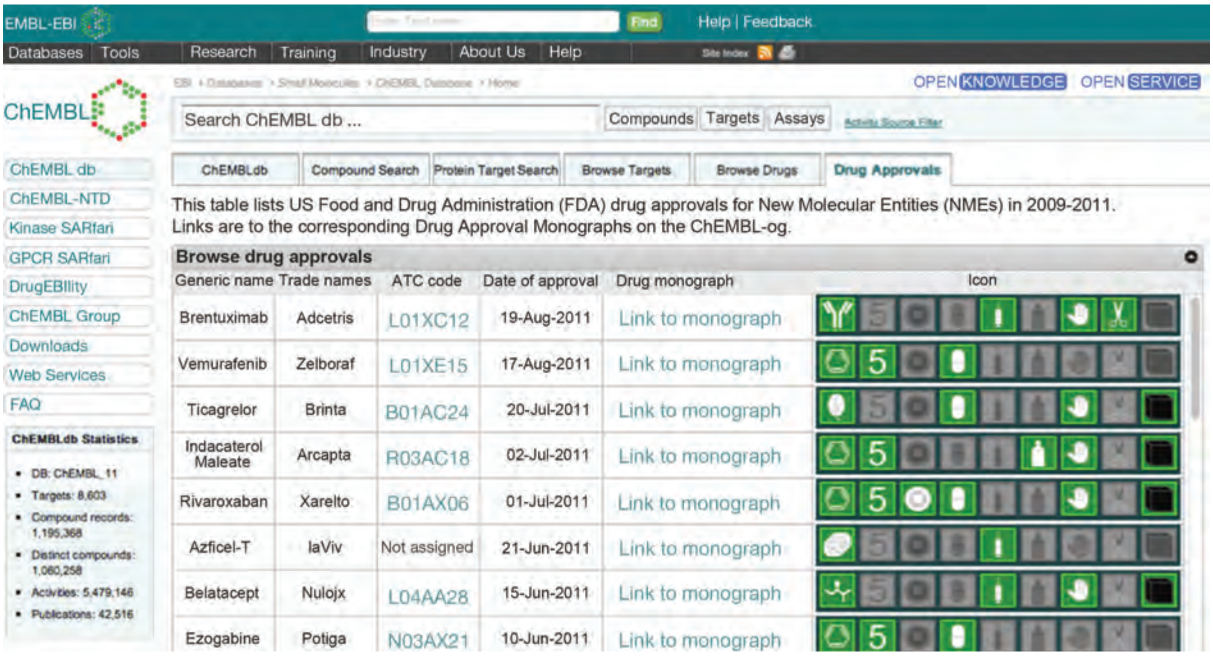
Another research area has been the exploration of the properties of 'tool compounds', the small-molecule probes of cells and pathways that are established as a core part of chemical biology research. We looked specifically at the quantitative conservation and differences of binding of compounds between established rodent models and human systems. We also explored the development of a novel integrated theoretical-, X-ray- and NMR-based approach to binding-site characterisation. Other research has included the mapping of temporal expression-level changes of drug-target and metabolic systems from prenatal to geriatric stages for rodents and humans.

The group's social media outreach has proved particularly popular, with the ChEMBL-og and Twitter feeds gathering a large and diverse user community. The ChEMBL-og features monographs of new drug approvals, tracking trends in drug and target innovation, and provides a unique molecular view on drug discovery, with links to ChEMBL and across other EBI resources.

We participated in three significant, EU-funded, trans-national projects: eTox, DiXa, and EU-OPENSCREEN. eTox is an Innovative Medicines Initiative that aims to build an unprecedented collaborative database of chronic rat-toxicity data and then perform bio-and chemoinformatic analyses and software development to predict toxicity, thereby improving the productivity of pharmaceutical discovery. DiXa is a project to integrate 'omics data across a broad range of public chemical safety datasets. EU-OPENSCREEN helps researchers identify compounds affecting new targets by integrating resources such as high-throughput screening platforms, chemical libraries, bio- and cheminformatics support, screening results, assay protocols, and chemical information.

FUTURE PLANS

In 2012 we will complete the rollout of UniChem across EMBL-EBI as a standard method for cross-resource chemical structure integration, and then add new complementary sources including patent data and commercially available compounds. We will continue to develop and extend approaches to predict the inherent tractability of new target systems for drug discovery, including extension of the data used to cover clinical development stage compounds. We will also develop new informatics approaches to discover and curate key clinical stage molecular and target validation data from the broad literature. Finally, we will explore the integration of ChEMBL data with human variation data for target validation, drug re-use and drug safety purposes.



ChEMBLdb Statistics

- DB: ChEMBL_11
- Targets: 8,603
- Compound records: 1,195,368
- Distinct compounds: 1,060,258
- Activities: 5,479,146
- Publications: 42,516

This table lists US Food and Drug Administration (FDA) drug approvals for New Molecular Entities (NMEs) in 2009-2011. Links are to the corresponding Drug Approval Monographs on the ChEMBL-og.

Browse drug approvals

Generic name	Trade names	ATC code	Date of approval	Drug monograph	Icon
Brentuximab	Adcetris	L01XC12	19-Aug-2011	Link to monograph	
Vemurafenib	Zelboraf	L01XE15	17-Aug-2011	Link to monograph	
Ticagrelor	Brinta	B01AC24	20-Jul-2011	Link to monograph	
Indacaterol Maleate	Arcapta	R03AC18	02-Jul-2011	Link to monograph	
Rivaroxaban	Xarelto	B01AX06	01-Jul-2011	Link to monograph	
Azficel-T	laViv	Not assigned	21-Jun-2011	Link to monograph	
Belatacept	Nulojx	L04AA28	15-Jun-2011	Link to monograph	
Ezogabine	Potiga	N03AX21	10-Jun-2011	Link to monograph	

Figure 2. The ChEMBL database.

Cheminformatics and metabolism

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.

DESCRIPTION OF SERVICES AND RESEARCH

Our team provides the biomedical community with information on small molecules and their interplay with biological systems. Our database portfolio includes ChEBI, EMBL-EBI's database and ontology of chemical entities of biological interest, as well as Rhea and IntEnz, our enzyme-related resources. The group develops methods to decipher, organise and publish the small molecule metabolic content of organisms. We develop algorithms to: predict metabolomes based on genomic and other information; determine quickly 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.

SUMMARY OF PROGRESS

- Developed the Enzyme Portal, which we will release in February 2012;
- Developed the MetaboLights database and archive, which we will release in February 2012;
- Issued several releases (70 to 86) of ChEBI, our ontology and database of chemical entities of biological interest;
- Issued several releases of enzyme resources Rhea (15 to 27) and IntEnz (62 to 74);
- Deployed Rhea web services in beta.

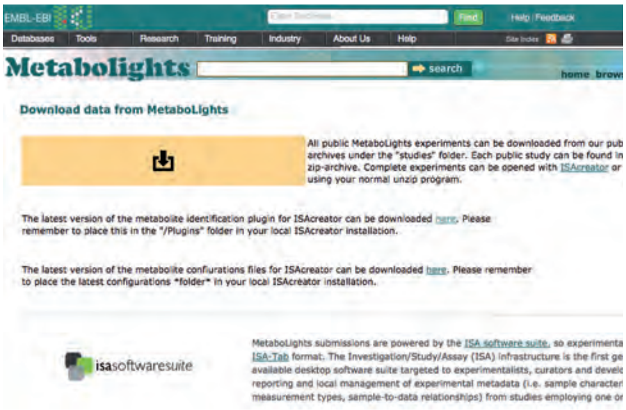
MAJOR ACHIEVEMENTS

In 2011 we focused on developing two new resources: the Enzyme Portal and MetaboLights. The Enzyme Portal unifies access to all enzyme-related information at EMBL-EBI, using information from: UniProtKB, Reactome, ChEMBL, EC2PDB, the Catalytic Site Atlas, MACiE (the cofactor database), IntEnz, Rhea and others. The Enzyme Portal is the first resource at EMBL-EBI to be fully designed and developed based on the demonstrated needs of the users. It removes artificial boundaries for viewing and using enzyme-related information, drawing on reputable resources to characterise different types of data. It shares the look and feel of the new EBI Search, and makes use of the technology developed by the EBI Search team. Users can query the Enzyme Portal with enzyme, gene, compound name, reaction ID or other enzyme-related concepts. Results are organised according to names, classification and function; protein structure; reactions and pathways; small molecules; disease and literature.

MetaboLights is an archive and reference database for metabolomics experiments and derived information. Following the release of the archive layer in early 2012, the development will focus on the reference database. This layer covers metabolite structures; their reference spectra and biological roles, locations and concentrations; and experimental data from metabolic experiments. Our search service development will focus on spectral and chemical similarities.

The ChEBI database team has continued to focus on curating high-quality small molecule data and improving the web interface by displaying content from Wikipedia. Furthermore, they have conducted extensive work on the ontology to improve the classification of natural products and carbohydrates.

In 2011 Rhea web services were publicly deployed as a beta version. Users can use the REST API to search the database programatically and retrieve individual reactions in CML, BioPAX or RXN formats.



Download data from MetaboLights

All public MetaboLights experiments can be downloaded from our pub archives under the "studies" folder. Each public study can be found in zip-archive. Complete experiments can be opened with ISAcreator or using your normal unzip program.

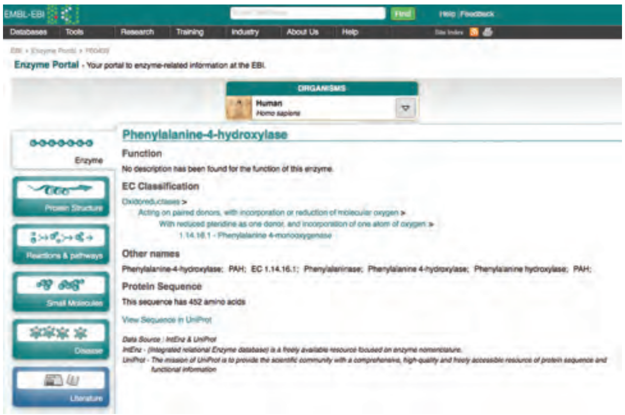
The latest version of the metabolite identification plugin for ISAcreator can be downloaded here. Please remember to place this in the "/Plugins" folder in your local ISAcreator installation.

The latest version of the metabolite confiurations files for ISAcreator can be downloaded here. Please remember to place the latest configurations "folder" in your local ISAcreator installation.

isasoftwaresuite

MetaboLights submissions are powered by the ISA software suite, so experimenta ISA-Tab format. The Investigation/Study/Assay (ISA) infrastructure is the first ge available desktop software suite targeted to experimentalists, curators and develo reporting and local management of experimental metadata (i.e. sample character measurement types, sample-to-data relationships) from studies employing one or

Figure 1. The new Metabolights resource.



Enzyme Portal - Your portal to enzyme-related information at the EBI.

Phenylalanine-4-hydroxylase

Function

No description has been found for the function of this enzyme.

EC Classification

Oxidoreductases >
Acting on paired donors, with incorporation or reduction of molecular oxygen >
With reduced pteridine as one donor, and incorporation of one atom of oxygen >
1.14.16.1 - Phenylalanine 4-monooxygenase

Other names

Phenylalanine-4-hydroxylase; PAH; EC 1.14.16.1; Phenylalaninase; Phenylalanine 4-hydroxylase; Phenylalanine hydroxylase; PAH;

Protein Sequence

This sequence has 452 amino acids

View Sequence in UniProt

Data Source : IntEnz & UniProt
IntEnz - (Integrated relational Enzyme database) is a freely available resource focused on enzyme nomenclature.
UniProt - The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information

Figure 2. Enzyme Portal: sample results page. The new resource will launch in February 2012.

Selected publications

Alcantara, R., Axelsen, K.B., *et al.* (2012) Rhea: a manually curated resource of biochemical reactions. *Nucleic Acids Res.* 40 (D1), D754-60.

De Matos, P., Adams, N., *et al.* (2012) A Database for Chemical Proteomics: ChEBI. *Methods Mol. Biol.* 803, 273-96.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

FUTURE PLANS

In 2012 we will release the Enzyme Portal and the MetaboLights database's archive layer. We will gather user feedback and work on further developing and stabilising these resources. A strong focus will be on developing the MetaboLights reference layer, which will contain information on individual metabolites and their chemical, spectroscopic and biological properties. This layer will strongly interface with EMBL-EBI resources such as ChEBI, Reactome, UniProtKB, the BioSamples Database and the Gene Expression Atlas.

We will work closely with the metabolomics community on data exchange formats and mechanisms. Our internal software engineering efforts will focus on creating curation and data submission components that can be re-used across EMBL-EBI's cheminformatics databases. As part of our efforts to develop metabolomics and metabolism resources, we will continue to work on the first release of a natural-products collection in ChEBI. It will feature information about approximately 5000 natural products, including their structure and detailed biological source.

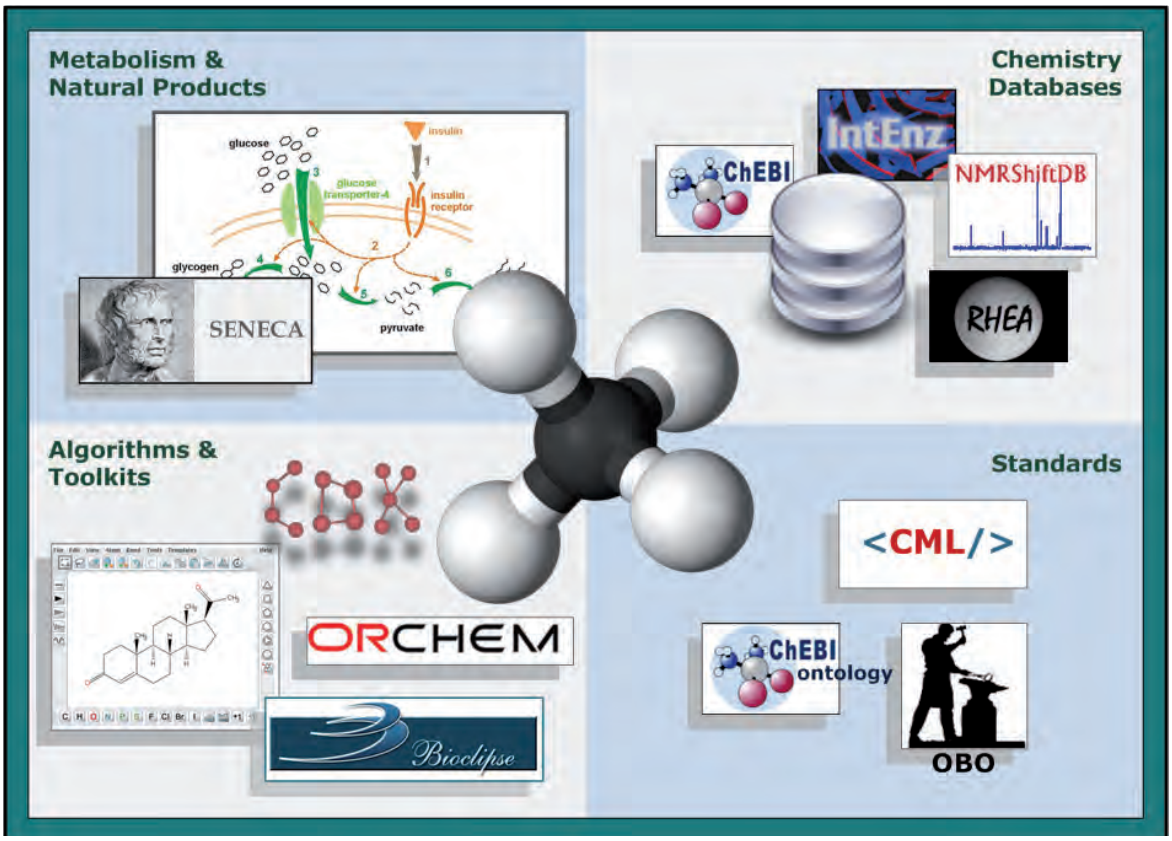


Figure 3. Summary of the activities conducted within the Cheminformatics and Metabolism group: natural products and metabolism, chemistry databases, standards and cheminformatics toolkits.

Functional genomics

Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. Postdoctoral research at New Mexico State University, US. At EMBL-EBI since 1997.

DESCRIPTION OF SERVICES AND RESEARCH

The Functional Genomics Team comprises teams led by Misha Kapushesky, Helen Parkinson, Ugis Sarkans and several staff reporting directly to Alvis Brazma. We focus on functional genomics data services, research in high-throughput sequencing and gene-expression data analysis, and research and development related to biomedical informatics and systems microscopy. We run several of EMBL-EBI's core resources: the ArrayExpress Archive of Functional Genomics Data, the Gene Expression Atlas and the BioSamples Database.

Our PhD students focus on data analysis, building models for systems biology and developing new methods and algorithms. Integration of data across multiple platforms and types of data is another important area of activity. We also contribute substantially to training in transcriptomics and the general use of EMBL-EBI bioinformatics tools.

SUMMARY OF PROGRESS

- Redeveloped the Gene Expression Atlas;
- Completed the migration of ArrayExpress data to a newly developed infrastructure;
- Developed a new interface for ArrayExpress;
- Released the BioSamples Database;
- Developed a prototype for a Systems Microscopy Database;
- In collaboration with Helsinki University of Technology, developed and published new methods for data-driven information retrieval;
- Organised and participated in over 25 training events.

MAJOR ACHIEVEMENTS

Gene Expression Atlas

Misha Kapushesky, pp. 36

The Gene Expression Atlas, part of the ArrayExpress infrastructure, allows users to query gene expression by gene name or property (e.g. gene ontology term) or by tissue type, cell type, disease state or other conditions in which the genes are expressed (Kapushesky *et al*., 2010). In 2011 the Atlas software underwent major redevelopment, making it more robust and setting it on sustainable ground. Most importantly, the Atlas now includes gene expression data from high-throughput sequencing experiments.

Functional Genomics Production

Helen Parkinson, pp. 38

The Functional Genomics Production Team manages data content and user interaction for the ArrayExpress Archive and Gene Expression Atlas as well as the new BioSamples Database. Major developments in 2011 included the completion of migration of ArrayExpress data to a newly developed infrastructure; establishing robust, high-throughput sequencing-based assay data management procedures and pipelines; and the release of the BioSamples Database.

Functional Genomics Development

Ugis Sarkans, pp. 40

The software development team builds and maintains several major components of the ArrayExpress infrastructure and BioSamples Database. Major achievements in 2011 were the completion of the transition to the new ArrayExpress software infrastructure; the development of new, richer user interface for ArrayExpress; and the release of the BioSamples Database.

Systems Microscopy Database and other pilot projects

Gabriella Rustici, Catherine Kirsanova, Johan Rung

We developed a prototype database and interface for the storage and visualisation of data generated in the domain of 'systems microscopy' by combining automated fluorescence microscopy, cell microarray platforms, RNAi assays, quantitative image analysis and data mining. This work was carried out as part of the Systems Microscopy

Network of Excellence, funded under the EU's Seventh Framework Programme (FP7). Jointly with collaborators from the University of Latvia, we also contributed to an FP7-funded pilot project that involves developing a data-sharing infrastructure for an International Cancer Genomics Consortium (ICGC) project on kidney cancer (CAGEKID).

Research

Johan Rung, Angela Goncalves, Mar Gonzales-Porta, Jing Su, Misha Kapushesky, Aurora Torrente, Wanseon Lee, Liliana Greger

Our ongoing research projects are related to regulation of gene expression and analysis of large-scale functional-genomics data. In addition to publishing several papers on RNA-seq analysis, our team has made advances in developing methods for gene expression data meta-analysis, in particular for discovery of diabetes candidate genes and new integrative analysis of RNAseq data. Part of our work involves analysing RNAseq and DNA data from kidney and other cancers, as a part of the International Cancer Genome Consortium. In collaboration with Helsinki University of Technology, we developed and published new methods for data-driven information retrieval.

Training

Gabriella Rustici, Ibrahim Emam, Angela Goncalves, Mar Gonzales-Porta, Emma Hastings, Annalisa Mupo, Tomasz Adamusiak, James Malone

In 2011 we organised and participated in over 25 training events, including the EBI Bioinformatics Roadshows and hands-on courses. These included the EMBO practical course on the analysis of high-throughput sequencing data, the most popular and oversubscribed event in EMBL-EBI's training calendar. We developed six modules for Train online (the EMBL-EBI online training resource, see page 80), ranging from an introductory functional genomics course to step-by-step guides on how to use the ArrayExpress Archive and Expression Atlas.

Reference cited: Kapushesky, M., Emam, I, *et al.* (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucl. Acids Res.* 38 (suppl 1): D690-D698.

Selected publications

Caldas, J., Gehlenborg, N., *et al.* (2012) Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. Bioinformatics 28 (2), 246-53.

Kutter, C., Brown, G.D., *et al.* (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43 (10), 948-55.

Gostev, M., Fernandez-Banet, J., *et al.* (2011) SAIL - a software system for sample and phenotype availability across biobanks and cohorts. *Bioinformatics* 27 (4), 589-91.

Parkinson, H., Sarkans, U., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39 (Database issue), D1002-4.

FUTURE PLANS

In 2012 our services will focus on the deep integration of the ArrayExrpess, Gene Expression Atlas, and BioSamples Databases with the overall EMBL-EBI data resource infrastructure. The Gene Expression Atlas will be conceptually changed, specifically to better accommodate RNA-seq data. Large-scale data integration and systems biology will remain in the focus or our research. We will work to develop methods for RNA-seq data analysis and processing, and apply these to address important biological questions such as ubiquity of gene expression, the role of alternative splicing and splicing mechanisms. With our collaborators from the International Cancer Genome Consortium we will be seeking new insights into cancer genomes and their impacts on functional changes in cancer development. In this area we will focus on discovery and analysis of fusion genes and their role in cancer development.

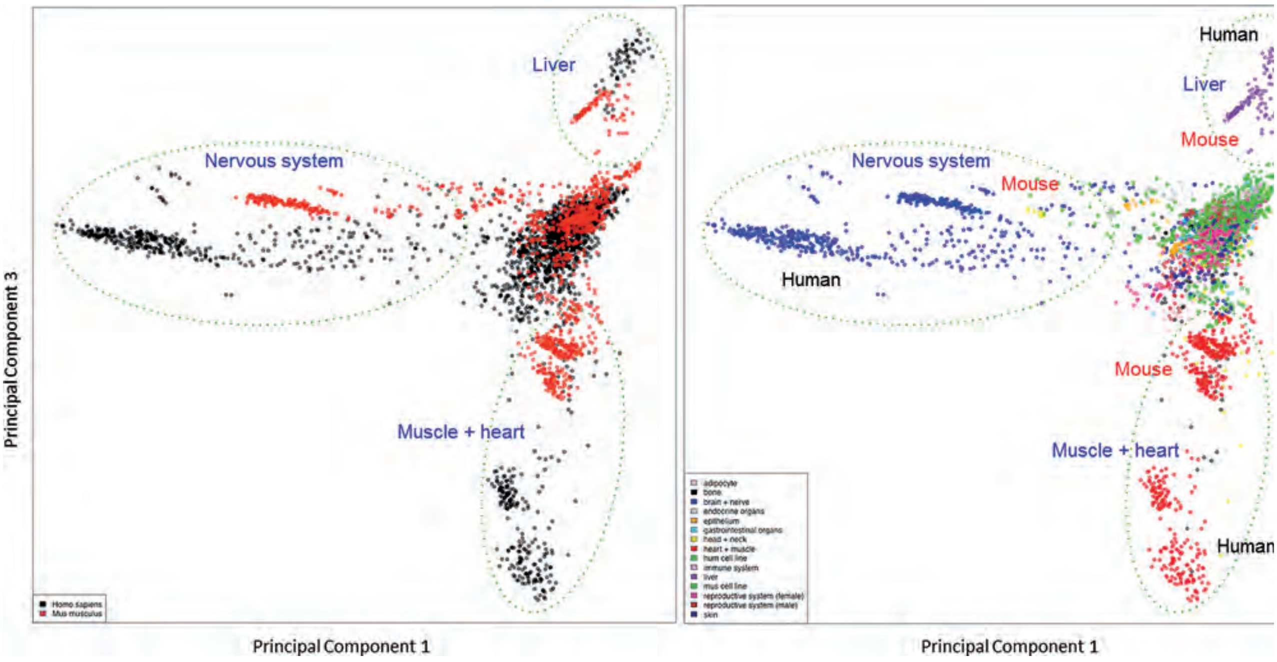
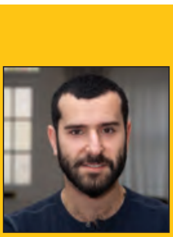


Figure. PCA plots of a combined human and mouse gene-expression data matrix (principal components 1 and 3). Each dot represents a sample, which is labeled by (a) species and (b) tissue type. Reproduced courtesy of Genome Biology.

Functional genomics: Atlas



Misha Kapushesky

BA in Mathematics and Comparative Literature, Cornell University, NY, USA, 2000. PhD in Genetics, University of Cambridge, UK, 2010.
At EMBL-EBI since 2001.

DESCRIPTION OF SERVICES AND RESEARCH

The Functional Genomics Atlas Team develops and runs the Expression Atlas database and the R Cloud service. The Expression Atlas is a value-added database of transcriptomics datasets, providing semantically rich searches and visualisations of gene activity in curated public data from the ArrayExpress Archive. The R Cloud is the cloud-computing infrastructure used by the Atlas and is offered as a remotely accessible R statistical analysis environment service to external users. We provide the Expression Atlas as stand-alone software capable of storing various types of 'omics' data. Since late 2010 we have issued regular monthly releases and, as of August 2011, the database supports 19 species including expression data measured for 19 014 biological conditions in 136 551 assays from 5598 independent studies. The Atlas Team conducts research in the area of functional genomics data analysis and integration with our collaborators in the EU-funded SYBARIS project on biomarkers of antifungal drug resistance and disease susceptibility.

SUMMARY OF PROGRESS

- Continued open-source, stand-alone releases of Expression Atlas software accompanied by regular data releases;
- Developed integrated views of next-generation sequencing data in the Expression Atlas;
- Loaded numerous large-scale microarray and RNA-seq data sets;
- Integrated an RNA-Seq processing pipeline, allowing direct loads of short-read data from the European Nucleotide Archive (ENA);
- Improved automated ontology annotation with EFO and Zooma utilities, and improved genome annotations through BioMart.

MAJOR ACHIEVEMENTS

Expression Atlas and R Cloud software have been in steady development and in 2011 we issued regular software and data releases. The software, which comes with a set of interfaces for data loading and processing, is frequently downloaded from our GitHub site. It is used to run the public EMBL-EBI Expression Atlas and to manage data for the SYBARIS project.

Data and curation improvements

Our focus has been shifting towards next-generation sequencing data sets. Using the ArrayExpressHTS pipeline (Goncalves et al., 2011) we loaded several dozen RNA-Seq experiments into the Atlas. We streamlined the pipeline between the ArrayExpress and European Nucleotide Archive (ENA; see page 14) and data is coming in regularly. We have continued to work on microRNA data curation and re-annotated all relevant microarray platforms to the latest version of miRBase, ensuring maximum dataset compatibility in the Atlas. We improved sample and transcript annotation by automating ontology mapping pipelines with Zooma and developing a completely new internal module for BioMart and other genome-annotation data sources.

User interface improvements

In addition to a thorough redesign of the internal engine, we improved the Atlas user interface received in a number of important ways. It now shows both differentially expressed and non-differentially expressed genes (Figure 3). Advanced filters based on differential expression counts and on compact ontology sub-trees allow users to easily traverse through large result sets to general queries. Together with the Production Team (see page 38), we developed the Vertebrate Bridging Ontology. We also imported cross-species homologous structure annotations, which are used to enhance condition searches in the Atlas; this significantly expanded multi-organism queries along these axes.

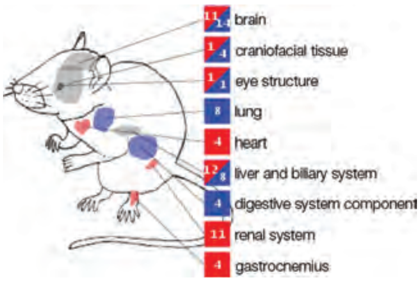


Figure 1. From the Expression Atlas.

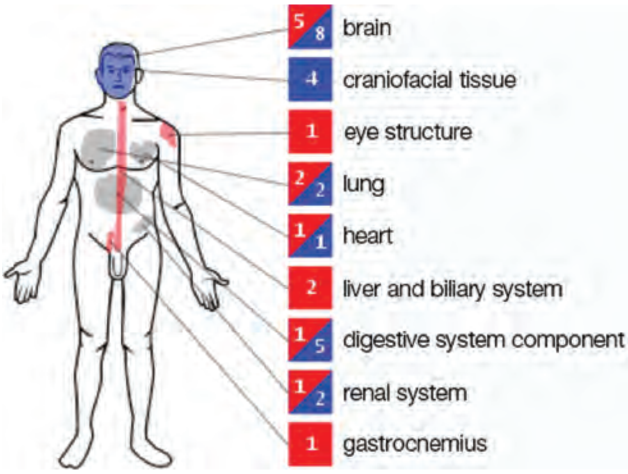


Figure 2. From the Expression Atlas.

SYBARIS

The SYBARIS project studies markers of antifungal drug resistance and disease susceptibility. We transferred all data generated in the project from our partners to EMBL-EBI, managed the project overall and participated in the meta-analysis of various datasets. We contributed to the analysis of genetic determinants of colony morphology in S. cerevisiae, the evolution of drug resistance in C. albicans through mistranslation, and the study of mutations and transcriptional response in the development of azole resistance in A. fumigatus and related species. Several publications have been submitted. In mid-2011 the SYBARIS project reached its mid-term mark and had a positive European Commission project review. The eight European partners reached all milestones and submitted all deliverables on time. As part of the review, we reported an overview of systems biology methods in fungal infectious disease (Santamaría et al., 2011).

Selected publications

Adamusiak, T., Burdett, T., *et al.* (2011) OntoCAT - simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics* 12, 218.

Culhane, A.C., Schroder, M.S., *et al.* (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Res. 40 (Database issue), D1060-6.

Goncalves, A., Tikhonov, A., *et al*. (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 27 (6), 867-9.

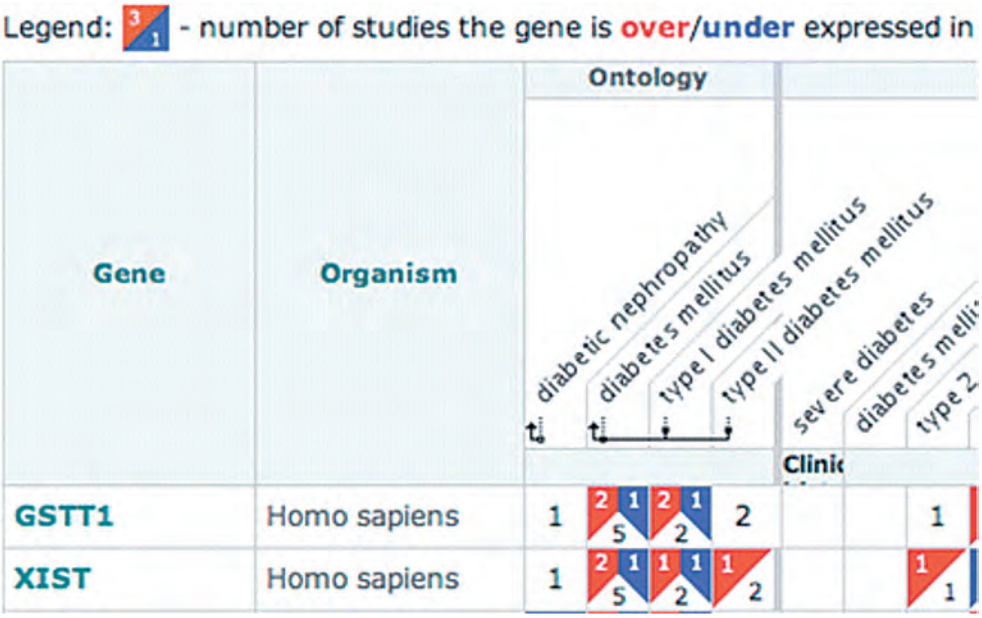
Kapushesky, M., Adamusiak, T., *et al.* (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40 (Database issue), D1077-81.

Kurbatova, N., Adamusiak, T., *et al*. (2011) ontoCAT: an R package for ontology traversal and search. *Bioinformatics* 27 (17), 2468-70.

Santamari¬a, R., Rizzetto, L., *et al.* (2011) Systems biology of infectious diseases: a focus on fungal infections. *Immunobiology* 11 (1212), 1227.

FUTURE PLANS

Following the significant work on making the Atlas and R Cloud more robust completed in 2011, we will concentrate in 2012 on data growth in the Atlas – particularly of high-throughput sequencing datasets. We will continue to improve and extend our RNA-Seq pipeline, which will strengthen our collaboration with the ENA. Another important area for the team will be collaborating with the Marioni group and the Huber group at EMBL Heidelberg to develop novel approaches to statistical analysis of large-scale, next-generation sequencing datasets. Finally, we will deliver an online knowledge base for SYBARIS summarising the project's findings and data submissions to public repositories. The consortium will conclude at the end of 2012. The Expression Atlas team will have a change in leadership in 2012, as Misha Kapushesky will be leaving EMBL.



Legend: 3 1 - number of studies the gene is **over**/**under** expressed in

Gene	Organism	Ontology				Clinic		
		diabetic nephropathy	diabetes mellitus	type I diabetes mellitus	type II diabetes mellitus	severe diabetes	diabetes melli	type 2
GSTT1	Homo sapiens	1	2 1 5	2 1 2	2			1
XIST	Homo sapiens	1	2 1 5	1 1 2	1 2			1 1

Figure 3. An example of the new Atlas display of non-differentially expressed genes. Red background for over-expression, blue for under-expression, white for non-differential expression.

Functional genomics: production

Helen Parkinson

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

DESCRIPTION OF SERVICES AND RESEARCH

Our team manages data content and user interaction for three EMBL-EBI databases: the ArrayExpress Archive (Parkinson, 2010), the Gene Expression Atlas (Kapushesky *et al.*, 2012) and the BioSamples Database (Gostev *et al.*, 2011). All three resources have complex metadata representing experimental types, variables and sample attributes for which we require semantic mark-up in the form of ontologies. We therefore develop both ontologies and software for the annotation of complex biological data including the Experimental Factor Ontology (EFO) for functional genomics annotation (Malone, 2010), the Software Ontology (SWO), Ontology for Biomedical Investigation (OBI), the Coriell Cell Line Ontology and the Vertebrate Anatomy Ontology (VBO; Travillian *et al.*, 2011).

ArrayExpress is a driving biological project for the National Center for BioOntology in the US and we are developing tools for ontology manipulation and for the semantic web to support functional genomics data integration. We have developed new formats that are used to solicit and load submissions to the BioSamples Database. This new resource will be used increasingly as a central resource of biological samples at EMBL-EBI; internal databases such as the European Genotype Phenotype Database (EGA) will link to it. Data exchange with the National Center for Biotechnology Information is underway.

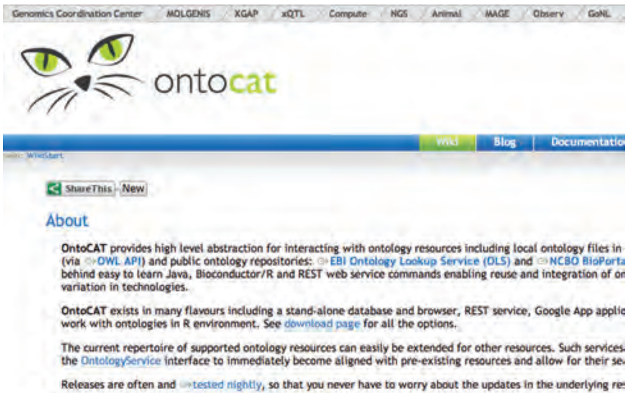
We handle two new types of data. In the context of the KOMP2 project, we manage, analyse, and distribute complex phenotypic data from 20 000 knockout mouse lines that will be generated over the course of five years. We also collaborate to develop tools for the annotation and production of the National Human Genome Research Institute's (NHGRI) genome-wide association study (GWAS) catalogue.

SUMMARY OF PROGRESS

- Re-annotation and data mining of blood array based experiments for meta-analysis;
- Issued 39 monthly releases of the Experimental Factor Ontology;
- Launched the Vertebrate Bridging Ontology;
- Launched the BioSamples Database, supporting formats and tools;
- Contributed to the formation of the KOMP2 Data Coordination Centre;
- Released MAGEComet, a web-based data annotation tool;
- Released ontoCat, an R package supporting ontology traversal and analysis;
- Developed RDF representation of ArrayExpress MAGE-TAB and Gene Expression Atlas content;
- Analysed RNA-Seq data on behalf of the Geuvardis consortium;
- Developed tools for the NHGRI GWAS catalogue.

MAJOR ACHIEVEMENTS

The main tasks of the group are processing, annotating, analysing and curating functional genomics data (e.g. BioSamples, RNA-Seq, and array-based expression) from both direct submissions and import from external resources; we also integrate these data across EMBL-EBI databases. We produce tools and develop infrastructure to support these tasks. For example, in collaboration with colleagues at the National Center for BioOntology we are developing ontology tools, and in collaboration with the Gen2Phen project we are developing data models for cross-species integration of phenotypic data. The EFO (Malone *et al.*, 2010) is released monthly to support the query of data in the Gene Expression Atlas and ArrayExpress. EFO has been extended to approximately 5000 classes (a 40% increase on 2010), is cross referenced to 25 public domain ontologies and has been expanded to support data mining of functional genomics blood data and GWAS data annotation. It can be also used in the EBI Search, which covers EMBL-EBI's core databases.



About

OntoCAT provides high level abstraction for interacting with ontology resources including local ontology files in (via OWL API) and public ontology repositories: EBI Ontology Lookup Service (OLS) and NCBO BioPorta behind easy to learn Java, Bioconductor/R and REST web service commands enabling reuse and integration of on variation in technologies.

OntoCAT exists in many flavours including a stand-alone database and browser, REST service, Google App applic work with ontologies in R environment. See download page for all the options.

The current repertoire of supported ontology resources can easily be extended for other resources. Such services the OntologyService interface to immediately become aligned with pre-existing resources and allow for their se

Releases are often and tested nightly, so that you never have to worry about the updates in the underlying re

Figure 1. ontoCAT, launched in 2011.

We produce open-source software for data management and annotation, ontology building and lexical mapping. In 2011 we released three new applications to complement our existing software: the MAGEComet annotation tool; an Atlas as RDF service for the semantic web community (www.ebi.ac.uk/efo/semanticweb/atlas); and the ontoCat R package supporting ontology traversal and analysis (Kurbatova *et al.*, 2011). Data analysis projects included re-annotation and meta-analysis of blood-derived gene-expression data, extension of R pipelines for the EU-funded Geuvadis project, analysis of RNA-Seq data and microRNA pipeline development.

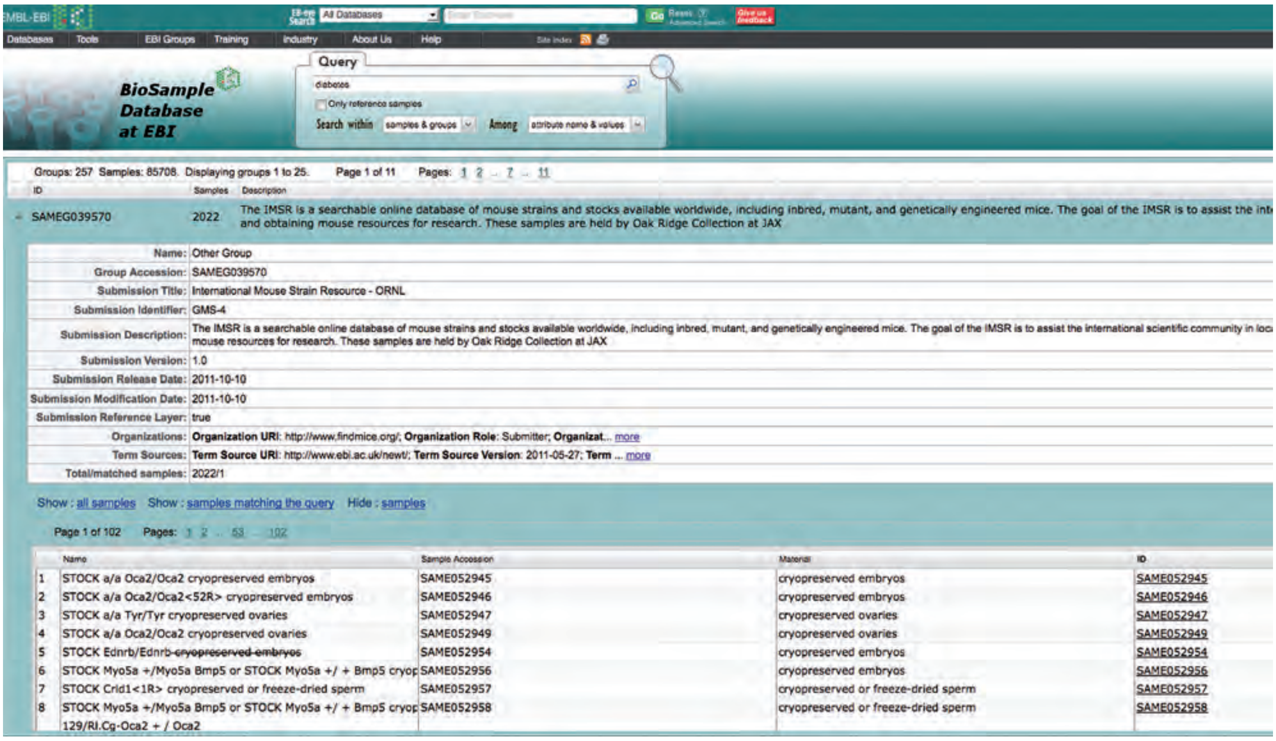
Selected publications

Adamusiak, T., Burdett, T., *et al.* (2011) OntoCAT - simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics* 12, 218.

Gostev, M., Faulconbridge, A., *et al.* (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.* 40 (Database issue), D64-70.

FUTURE PLANS

In 2012 we will start a new National Science Foundation-funded collaboration with the Plant Database Gramene to extend our RNA-Seq data processing pipelines to crop plants and to develop annotation standards for these. Population of the BioSamples Database will continue with internal EMBL-EBI databases as well as external resources; a new submission system is under development for data acquisition. The KOMP2 mouse phenotypic data flow will begin in 2012 and first public versions of the KOMP2 database and data processing pipelines will be made available to support data release. Work on ontologies will include development of EFO views for different user communities, for example GWAS versus Gene Expression and integration of Gene Ontology and pathway-specific terms to support complex gene/disease/pathway queries.



Groups: 257 Samples: 85708. Displaying groups 1 to 25. Page 1 of 11 Pages: 1 2 .. 7 .. 11

ID	Samples	Description
SAMEG039570	2022	The IMSR is a searchable online database of mouse strains and stocks available worldwide, including inbred, mutant, and genetically engineered mice. The goal of the IMSR is to assist the int and obtaining mouse resources for research. These samples are held by Oak Ridge Collection at JAX

Name:	Other Group
Group Accession:	SAMEG039570
Submission Title:	International Mouse Strain Resource - ORNL
Submission Identifier:	GMS-4
Submission Description:	The IMSR is a searchable online database of mouse strains and stocks available worldwide, including inbred, mutant, and genetically engineered mice. The goal of the IMSR is to assist the international scientific community in loca mouse resources for research. These samples are held by Oak Ridge Collection at JAX
Submission Version:	1.0
Submission Release Date:	2011-10-10
Submission Modification Date:	2011-10-10
Submission Reference Layer:	true
Organizations:	**Organization URI**: http://www.findmice.org/; **Organization Role**: Submitter; **Organizat**... more
Term Sources:	**Term Source URI**: http://www.ebi.ac.uk/newt/; **Term Source Version**: 2011-05-27; **Term** ... more
Total/matched samples:	2022/1

Show : all samples Show : samples matching the query Hide : samples

Page 1 of 102 Pages: 1 2 .. 53 .. 102

	Name	Sample Accession	Material	ID
1	STOCK a/a Oca2/Oca2 cryopreserved embryos	SAME052945	cryopreserved embryos	SAME052945
2	STOCK a/a Oca2/Oca2<52R> cryopreserved embryos	SAME052946	cryopreserved embryos	SAME052946
3	STOCK a/a Tyr/Tyr cryopreserved ovaries	SAME052947	cryopreserved ovaries	SAME052947
4	STOCK a/a Oca2/Oca2 cryopreserved ovaries	SAME052949	cryopreserved ovaries	SAME052949
5	STOCK Ednrb/Ednrb ~~cryopreserved embryos~~	SAME052954	cryopreserved embryos	SAME052954
6	STOCK Myo5a +/Myo5a Bmp5 or STOCK Myo5a +/ + Bmp5 cryop	SAME052956	cryopreserved embryos	SAME052956
7	STOCK Crid1<1R> cryopreserved or freeze-dried sperm	SAME052957	cryopreserved or freeze-dried sperm	SAME052957
8	STOCK Myo5a +/Myo5a Bmp5 or STOCK Myo5a +/ + Bmp5 cryop 129/RI.Cg-Oca2 + / Oca2	SAME052958	cryopreserved or freeze-dried sperm	SAME052958

Figure 2. A query result from the BioSamples Database showing samples, their annotation and sample groups.

Functional genomics: software development

Ugis Sarkans

PhD in Computer Science, University of Latvia, 1998. Postdoctoral research at the University of Wales, Aberystwyth, 2000. At EMBL-EBI since 2000.

DESCRIPTION OF SERVICES AND RESEARCH

Our team has been developing software for ArrayExpress since 2001. As of December 2011, ArrayExpress holds data from more than 770 000 microarray hybridisations and is one of the major data resources of EMBL-EBI. The software development team builds and maintains several components of the ArrayExpress infrastructure, including data management tools for the ArrayExpress Archive (the MIAME-compliant database for the data that support publications); the ArrayExpress Archive user interface (UI); and MIAMExpress (a data annotation and submission system). Since 2010 our team has participated in building the BioSamples Database, a new EBI resource. We also collaborate on a number of ‘multi-omics’ projects in a data-management capacity.

SUMMARY OF PROGRESS

- Brought into service a new ArrayExpress data management infrastructure;
- Maintained data submission flow and access during migration to the new data infrastructure;
- Evolved the Archive user interface to provide a richer, more consistent service;
- Launched the BioSamples Database, which contains more than 1 million samples.

MAJOR ACHIEVEMENTS

In 2011 we finished migration to the new ArrayExpress infrastructure, based on the MAGE-TAB data exchange format. For a significant part of the year the ArrayExpress user interface was working on the old and new databases in parallel, while the team (together with the Functional Genomics Production team) was working on various aspects of the new software and data migration. This major infrastructure change took place in the background, and was not noticeable to users.

We improved several tools that are used internally for data management, and also revamped tools that were affected by the change in our primary data format (e.g. data loaders). Specifically, we redeveloped tools for access-control management (essential for pre-publication data) as well as the automated MIAME scorer, which is now able to score relevant experiments according to MINSEQE (Minimum Information about a high-throughput SeQuencing Experiment).

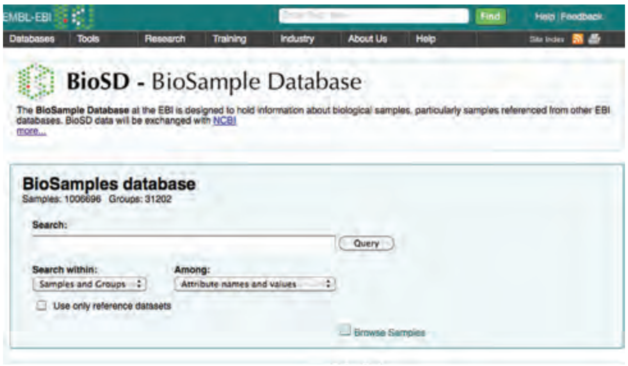
The ArrayExpress Archive user-interface work progressed in parallel with back-end infrastructure development. In preparation for retiring the old back-end, we redeveloped some older parts of the user interface; now, different types of objects are presented to users via the same UI paradigm. We re-implemented microarray design search and display, as well as protocol search and visualisation. In one novel view, users can view detailed experiment annotations as a user-friendly table in a browser, instead of relying on Excel spreadsheets.

One goal in developing EMBL-EBI’s new BioSamples Database is to clean up and aggregate aspects of information about biological samples that are served by different data resources. Our team participates in the design and development process of this database, lending our experience in handling various aspects of biological sample information management and reusing or and adapting relevant parts of the ArrayExpress software for these purposes. In 2011 period the BioSamples Database was released and grew to hold more than 1 million samples.

The team was involved in the SIROCCO project, which focuses on small RNA molecules that regulate gene expression during growth and development. Together with Anton Enright’s group, we built the SIROCCO Data Centre for storage and handling of datasets generated within the consortium. The system utilises the SIMBioMS data management system, a multi-module solution for data management in biomedical studies.

FUTURE PLANS

A priority area for ArrayExpress in 2012 is data quality. We will ensure that all our datasets are available through data-analysis packages, such as the ArrayExpress package in Bioconductor, in a usable form. We will also link ArrayExpress to GenomeSpace, a new data integration initiative lead by the Broad Institute. We will give more attention to helping users with data submissions. In particular, we will adjust the Annotare microarray data annotation tool for ArrayExpress use. Better support for sequencing-based functional genomics data submission will be developed. We will continue to improve our UI, making it more consistent across various sections and enhancing usability aspects such as query support. We will also



EMBL-EBI

Databases | Tools | Research | Training | Industry | About Us | Help

BioSD - BioSample Database

The **BioSample Database** at the EBI is designed to hold information about biological samples, particularly samples referenced from other EBI databases. BioSD data will be exchanged with NCBI
more...

BioSamples database
Samples: 1006696 Groups: 31202

Search:

Search within: Samples and Groups

Among: Attribute names and values

□ Use only reference datasets

Figure 1. The BioSamples Database, launched in 2011.

Selected publications

Parkinson, H., Sarkans, U., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39 (Database issue), D1002-4.

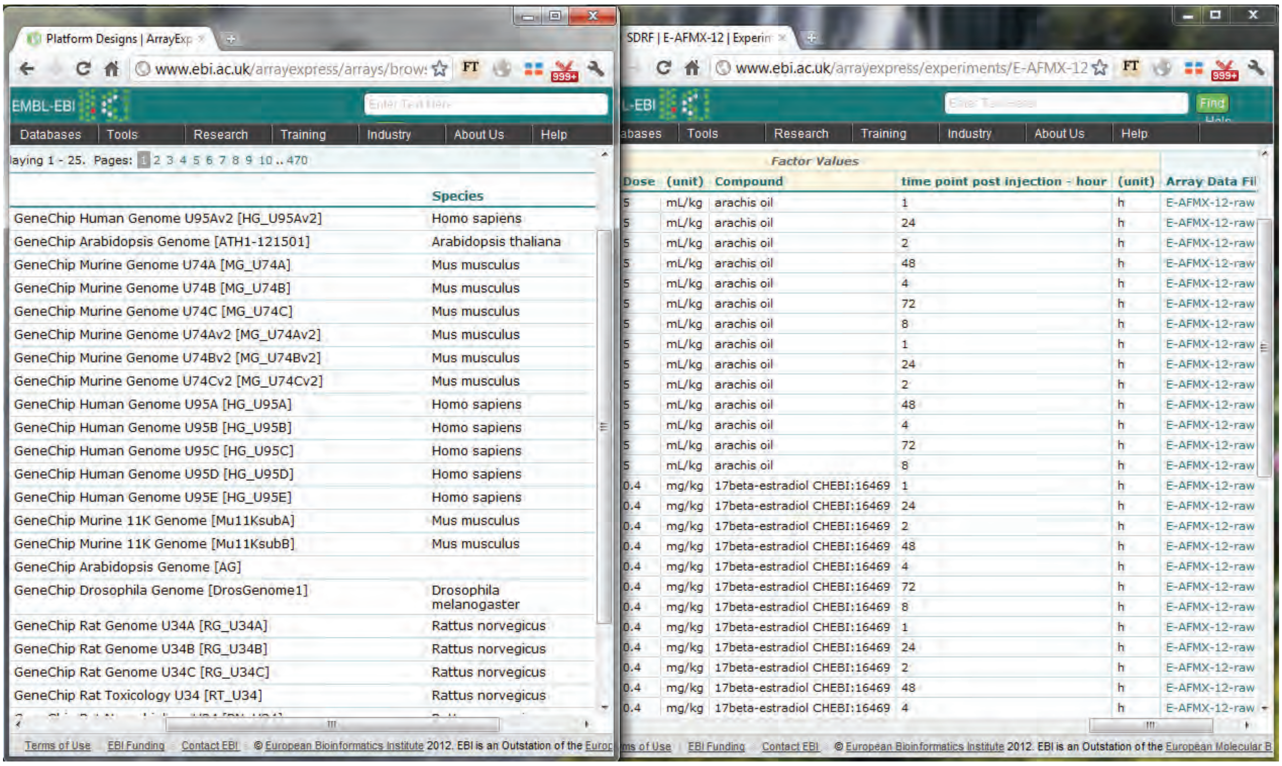
Nicholson, G., Rantalainen, M., *et al.* (2011) A Genome-Wide Metabolic QTL Analysis in Europeans Implicates Two Loci Shaped by Recent Positive Selection. *PLoS Genet.* 7 (9), e1002270.

Nicholson, G., Rantalainen, M., *et al.* (2011) Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.* 7, 525.

work towards adjusting the ArrayExpress Archive interface to the new EBI website presentation and functionality paradigm.

As we continue working on both ArrayExpress and BioSamples databases, we will ensure that both projects cross-feed on various levels: design experience, software and data exchange. ArrayExpress will serve as a test bed for demonstrating how the BioSamples Database can be useful for other resources at the EBI. We will ensure that biological samples that have been submitted to the BioSamples Database can be reused in data submissions processed by ArrayExpress, and that appropriate links are maintained in both directions. We will actively work with other databases at EMBL-EBI, ensuring that the BioSamples Database can satisfy their needs with respect to biological-sample data submissions, management, queries and presentation. We will also reuse the ArrayExpress UI software for the BioSamples project.

Our team participates in several pilot projects in a data management role. We believe that being close to large consortia generating different kinds of high-throughput data places us in a better position to fulfil our main objective of developing ArrayExpress and BioSamples infrastructures. Our participation in diXa, a toxicogenomics data management project, will enable us to build better links between BioSamples, ArrayExpress and other EMBL-EBI assay-data resources. We are also beginning to work with a project devoted to autism research, which will contribute to a deeper understanding of ways to manage complex endophenotype data, such as imaging data.



Platform Designs | ArrayExp

www.ebi.ac.uk/arrayexpress/arrays/brow:

EMBL-EBI

Databases | Tools | Research | Training | Industry | About Us | Help

laying 1 - 25. Pages: 1 2 3 4 5 6 7 8 9 10 .. 470

	Species
GeneChip Human Genome U95Av2 [HG_U95Av2]	Homo sapiens
GeneChip Arabidopsis Genome [ATH1-121501]	Arabidopsis thaliana
GeneChip Murine Genome U74A [MG_U74A]	Mus musculus
GeneChip Murine Genome U74B [MG_U74B]	Mus musculus
GeneChip Murine Genome U74C [MG_U74C]	Mus musculus
GeneChip Murine Genome U74Av2 [MG_U74Av2]	Mus musculus
GeneChip Murine Genome U74Bv2 [MG_U74Bv2]	Mus musculus
GeneChip Murine Genome U74Cv2 [MG_U74Cv2]	Mus musculus
GeneChip Human Genome U95A [HG_U95A]	Homo sapiens
GeneChip Human Genome U95B [HG_U95B]	Homo sapiens
GeneChip Human Genome U95C [HG_U95C]	Homo sapiens
GeneChip Human Genome U95D [HG_U95D]	Homo sapiens
GeneChip Human Genome U95E [HG_U95E]	Homo sapiens
GeneChip Murine 11K Genome [Mu11KsubA]	Mus musculus
GeneChip Murine 11K Genome [Mu11KsubB]	Mus musculus
GeneChip Arabidopsis Genome [AG]	
GeneChip Drosophila Genome [DrosGenome1]	Drosophila melanogaster
GeneChip Rat Genome U34A [RG_U34A]	Rattus norvegicus
GeneChip Rat Genome U34B [RG_U34B]	Rattus norvegicus
GeneChip Rat Genome U34C [RG_U34C]	Rattus norvegicus
GeneChip Rat Toxicology U34 [RT_U34]	Rattus norvegicus

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the Europ

SDRF | E-AFMX-12 | Experin

www.ebi.ac.uk/arrayexpress/experiments/E-AFMX-12

L-EBI

abases | Tools | Research | Training | Industry | About Us | Help

Factor Values					
Dose	(unit)	Compound	time point post injection - hour	(unit)	Array Data Fil
5	mL/kg	arachis oil	1	h	E-AFMX-12-raw
5	mL/kg	arachis oil	24	h	E-AFMX-12-raw
5	mL/kg	arachis oil	2	h	E-AFMX-12-raw
5	mL/kg	arachis oil	48	h	E-AFMX-12-raw
5	mL/kg	arachis oil	4	h	E-AFMX-12-raw
5	mL/kg	arachis oil	72	h	E-AFMX-12-raw
5	mL/kg	arachis oil	8	h	E-AFMX-12-raw
5	mL/kg	arachis oil	1	h	E-AFMX-12-raw
5	mL/kg	arachis oil	24	h	E-AFMX-12-raw
5	mL/kg	arachis oil	2	h	E-AFMX-12-raw
5	mL/kg	arachis oil	48	h	E-AFMX-12-raw
5	mL/kg	arachis oil	4	h	E-AFMX-12-raw
5	mL/kg	arachis oil	72	h	E-AFMX-12-raw
5	mL/kg	arachis oil	8	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	1	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	24	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	2	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	48	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	4	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	72	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	8	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	1	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	24	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	2	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	48	h	E-AFMX-12-raw
0.4	mg/kg	17beta-estradiol CHEBI:16469	4	h	E-AFMX-12-raw

ms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular B

Figure 2. New views in the ArrayExpress user interface: array design list view (left) and single experiment detail view (right).

Protein Data Bank in Europe (PDBe): Bringing structure to biology

Gerard Kleywegt

PhD University of Utrecht, 1991. Postdoctoral researcher, then independent investigator, University of Uppsala, 1992-2009. Coordinator, 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.

DESCRIPTION OF SERVICES AND RESEARCH

The 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 the Biological Magnetic Resonance Bank (BMRB) in the US and the Protein Data Bank Japan (PDBj). PDBe is a deposition and annotation site for the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB; see Scientific Advisory Boards, page 100).

The major goal of PDBe is to provide integrated structural data resources that evolve with the needs of biologists. To that end, our team endeavours to: handle deposition and annotation of structural data expertly; provide an integrated resource of high-quality macromolecular structures and related data; and maintain in-house expertise in all the major structure-determination techniques (i.e. X-ray crystallography, Nuclear Magnetic Resonance spectroscopy and 3D Electron Microscopy). Our specific focus areas are: advanced services, ligands, integration, validation and experimental data.

SUMMARY OF PROGRESS

- Processed 1496 depositions to the PDB and 163 depositions to EMDB. One of the PDB depositions was the 10 000th structure annotated at the EBI;
- Launched several new tools on the PDBe website:
  - PDBportfolio: pictorial summaries of PDB entries;
  - UniPDB: schematic views of structural coverage of UniProt entries;
  - Vivaldi: validation and analysis of NMR structures;
  - new modules for the PDBeXplore structure browser (by chemistry, taxonomy and GO classification);
  - PDBeXpress: sophisticated analyses of PDB data using a simple interface;
  - various tools to search, mine, analyse and visualise EMDB data;
- In collaboration with our wwPDB and EMDataBank partners, released remediated versions of the PDB and EMDB archives;
- Carried out a substantial number of user-training roadshows and workshops, and expanded our outreach and communications activities.

MAJOR ACHIEVEMENTS

Following a major turnover in 2009-2010, the team's many new members have successfully transitioned into their new roles and have made a positive impact on our services. Leadership of the team was strengthened in 2011 with the appointment of Dr Sameer Velankar as Team Leader in charge of content and integration (see page 46). He joins Dr Tom Oldfield, who was appointed Team Leader in 2010 to lead on databases and services (see page 44).

In 2011 PDBe annotation staff curated a record number of entries, both in absolute numbers (1496 PDB entries, up 9% from 2010) and in relative terms (16.1% of all deposited PDB entries worldwide, up from 15.4% in 2010). The Wellcome Trust blog ran an article about one of these entries: the 10 000th PDB structure annotated at EMBL-EBI.

The team is collaborating with partners in the US and Japan to create a common tool for handling the deposition and annotation of structural data on biomacromolecules – obtained using any technique or combination of techniques – by all wwPDB and EMDataBank partners. The PDBe team is a major contributor to software development in this project and is responsible for the workflow manager, the validation module and the deposition interface. The new tool will go into production at all deposition sites in late 2012 (see Oldfield, page 44).

A new version of the PDB archive featuring substantial improvements to the description of peptide-like inhibitors was released in 2011. The next round of remediation is underway. A PDBe-hosted workshop in 2011, comprising the major developers of X-ray structure refinement programs, resulted in a decision to migrate to the powerful and flexible mmCIF format so that the archive can accommodate depositions of large entries and complex chemistry. The team also started to implement the recommendations of the wwPDB X-ray Validation Task Force, which were published in October 2011.

PDBe's Electron Microscopy staff improved the quality, accuracy and integrity of the EMDB archive and improved software to ensure higher quality of future depositions. They developed functionality to improve and extend the EMDataBank atlas pages, which now feature an interactive viewer for EM maps and models as well as static images and plots to help users assess the quality of map–model fits. They also developed a powerful search tool and a data-mining service for EMDB data. PDBe's NMR staff reorganised and redesigned important areas of the website and developed Vivaldi, which allows users to display and analyse NMR structures in the PDB and to assess their quality. Other developments related to PDBe services are described by Sameer Velankar and Tom Oldfield.

The PDBe team organises its own outreach and training activities and participates in EMBL-EBI Bioinformatics Roadshows (see page 80). In 2011 we were involved in 11 user-training roadshows (in Australia, Czech Republic, Greece, Portugal, South Africa, Sweden and Turkey). We co-organised the EMBO course on Computational Structural Biology and organised an EM data-management workshop, both on-site in Hinxton. We exhibited with our wwPDB partners at the IUCr Congress, a major crystallography conference held in Madrid, and delivered presentations at many scientific institutes and professional meetings (18 lectures, 15 poster presentations). We also expanded our outreach and communications programme by: co-organising "PDB40", a symposium to celebrate the 40th anniversary of the PDB archive; launching Quips, interactive articles (pdbe.org/quips) about biologically important structures; and increasing our presence in social media and other fora.

Selected publications

Lawson, C.L., Baker, M.L., *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39 (Database issue), D456-64.

Velankar, S., Alhroub, Y., *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 39 (Database issue), D402-10.

Velankar, S. and Kleywegt, G.J. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. Acta Crystallogr. *D Biol. Crystallogr.* 67 (Pt 4), 324-30.

FUTURE PLANS

In the coming years we plan to make PDBe the logical first stop on any quest for structural information. To transform the structural archive into a truly useful resource for biomedical and related disciplines, we will focus on developing five key areas: advanced services (e.g. PDBePISA, PDBeFold, PDBeMotif and the new PDB browsers); annotation, validation and visualisation of ligand data; integration with other biological and chemical 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 they support the structural models and inferences.

In 2012 we expect to complete user testing and public release of the new joint wwPDB Deposition and Annotation system. Redesign of the PDB and EMDB web pages and search system will also be undertaken, and public release is expected in 2013.

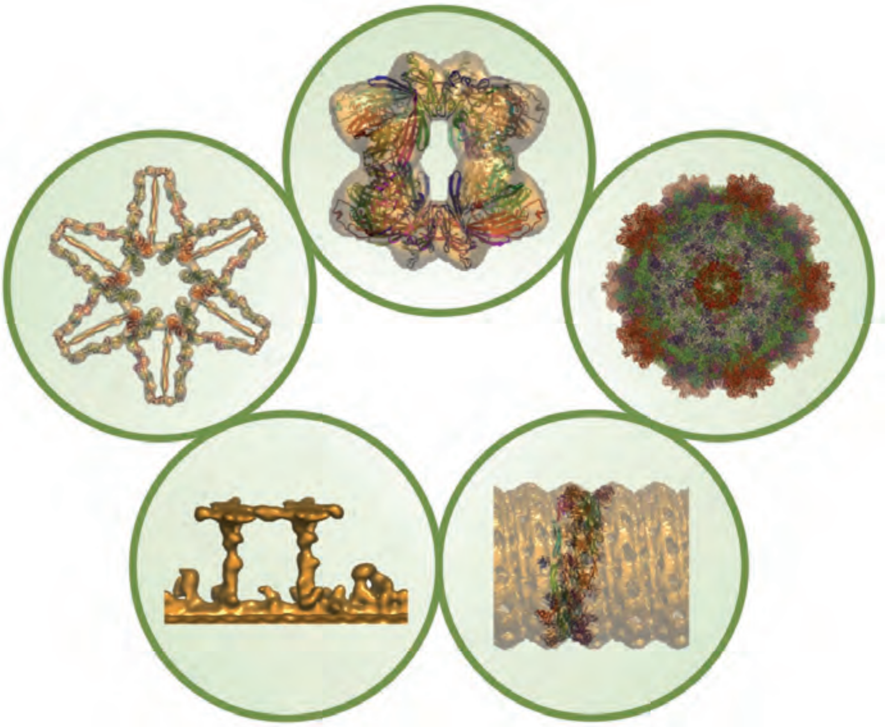


Figure. Sample entries from the Electron Microscopy Data Bank (EMDB) and selected fitted models from the PDB. EMDB is the global public repository for EM density maps, founded at the EBI in 2002. EMDB entries range from near-atomic resolution maps of viruses to tomographic reconstructions of huge macromolecular machines and complexes inside the cell. Biologists are increasingly using hybrid techniques involving combinations of different electron microscopy methods (e.g. single-particle processing, tomography) as well as other structure-determination techniques to answer important biological questions. In many cases, they can find an atomic model from the PDB for parts or all of a structure held in EMDB. In 2011, PDBe introduced new tools to search, mine, analyse and visualise EMDB data and related atomic models. Pictured (clockwise from top): human alpha B-crystallin (EMD-1894, PDB: 2ygd), cytoplasmic polyhedrosis virus (EMD-5256, PDB: 3izx), GMPPCP-stabilised human dynamin 1 delta PRD polymer (EMD-1949, PDB: 3zys), radial spokes from *Chlamydomonas Reinhardti* flagella (EMD-1941), myosin V inhibited state (EMD-1201, PDB: 2dfs).

Protein Data Bank in Europe (PDBe): databases and services

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.

DESCRIPTION OF SERVICES

The Protein Data Bank in Europe (PDBe) is one of six core databases located at EMBL-EBI and is also a partner in the Worldwide Protein Data Bank organisation (wwPDB) along with RCSB and BMRB in the US and PDBj in Japan. The PDBe team manages two production systems: the weekly update of deposited data and the weekly increment of new released 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.

SUMMARY OF PROGRESS

- Improved the up-time of PDBe resources based on core databases;
- Updated the main data-process flow to an automated system for weekly increments;
- Significantly enhanced and extended three existing services and added two new services;
- Continued development of a new deposition and annotation system for the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB) in collaboration with wwPDB partners. Undertook project coordination, development of workflow system components, database implementation and development of the deposition system.

MAJOR ACHIEVEMENTS

During 2011 PDBe put in place a much more structured production environment and processes to manage the deposition of data, the release of new data and all back-end support for services. These changes resulted in an improvement in the delivery of the weekly release cycle in line with the wwPDB universal release time (Wednesday midnight in the UK). This happens to the nearest minute without loss of service, with correct delivery of the new data and validation and reporting of any problems. Given the update time of midnight these processes must be automated and self-validating. Additional benefits included a reduction in the amount of time staff spend maintaining the production systems, which allows the team time to concentrate on new systems.

The PDBe team makes major contributions to the coordinated development of a new wwPDB deposition and annotation system. PDBe staff have been directly involved with the design and planning of this work along with the other wwPDB partners, and have developed software for the project, including a workflow system and the deposition user interface for the scientists depositing data.

There have been major updates of the Protein Interfaces, Surfaces and Assembly server (PDBePISA) and structure-similarity server (PDBeFold), which are popular resources provided by the PDBe team. During 2011 these services were adapted so they are more robust and run properly in a production environment. We made a large number

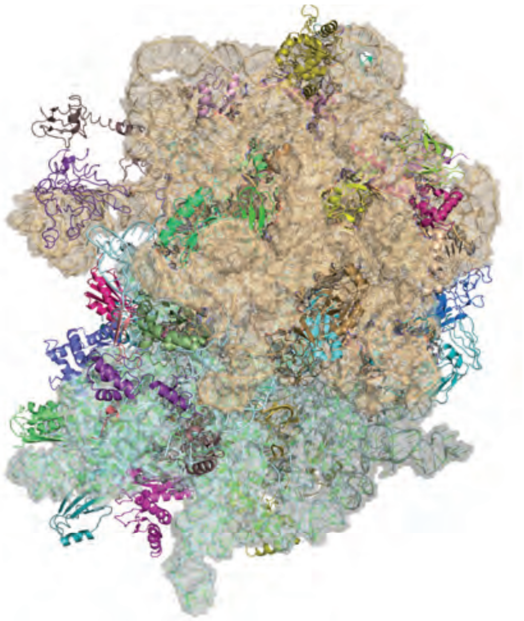


Figure 1. The ever-larger and increasingly complex structures deposited in the PDB require robust data handling and archiving systems. The PDBe Databases and Services team addresses these challenges.

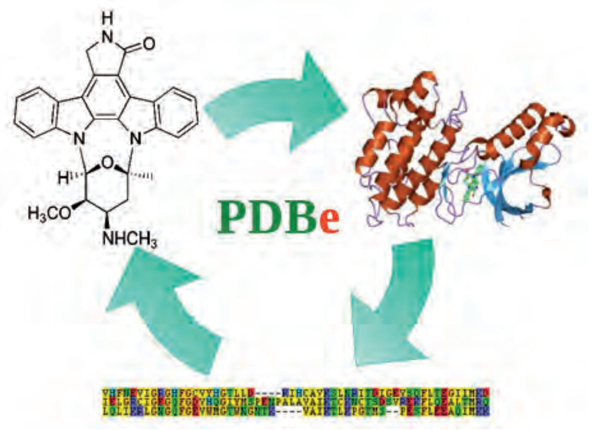


Figure 2. In order to make the structures more useful, we need to iink them to sequence and chemistry data. Our databases and services endeavour to capture the relevant information linking these different types of information.

of improvements, providing a significant extension in functionality for users.

We performed a major update of the PDBeChem chemistry service, completely rewriting all code to enable the system to support future development.

PDBeXpress makes the very powerful and extensive service PDBeMotif more accessible to a wider range of users. We have essentially provided shortcuts to PDBeMotif functionality with simple user interfaces. PDB Highlights is a new service that provides short-cuts for users to find macromolecular structures of a certain type quickly, for example: the structure with the highest resolution or the most RNA chains. This is part of our effort to make structural information more accessible to biologists and to allow future unification of services across PDBe and EMBL-EBI.

Selected publications

Lawson, C.L., Baker, M.L., *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39 (Database issue), D456-64.

Velankar, S., Alhroub, Y., *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 39 (Database issue), D402-10.

FUTURE PLANS

The PDBe team is currently moving its services to the London data centres and the aim is to complete this process in spring 2012.

With the future provision of new services based on structure validation data there will be an emphasis in 2012 on extending the core databases and infrastructure. This will require an optimisation of the loading tools to manage both the increased number of depositions and the breadth of data required to support new services.

wwPDB is aiming to bring the new deposition and annotation system online at the end of 2012 at all PDB and EMDB deposition sites.

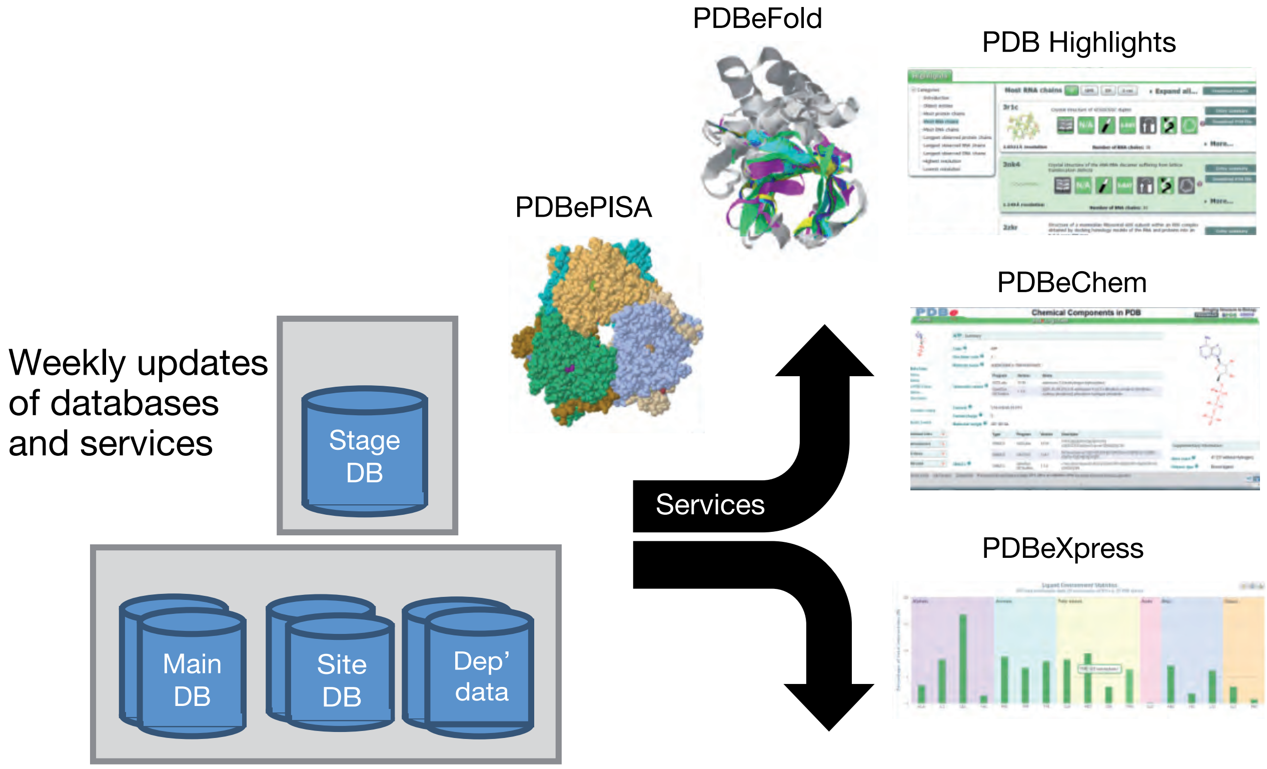


Figure 3. PDBe maintains multiple Oracle servers to enable good performance and minimise down-time for all the PDBe web services. This enables the PDBe team to manage two overlapping weekly flows of data for deposition and dissemination of information. There have been major updates to the popular PDBePISA, PDBeFold and PDBeChem services. PDB Highlights and PDBeXpress are two new services that went into use in 2011.

Protein Data Bank in Europe (PDBe): content and integration

Sameer Velankar

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

DESCRIPTION OF SERVICES

As a founding member of the Worldwide Protein Data Bank (wwPDB), the Protein Data Bank in Europe (PDBe) manages the global biomacromolecular structure archive, the Protein Data Bank (PDB). The wwPDB partners accept and annotate worldwide depositions of biomacromolecular structures determined using X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, 3D Electron Microscopy (EM) and other structure-determination methods. PDBe is a founding member of EMDataBank, the organisation that manages EMDB.

Our goal is to ensure that PDBe truly serves the needs of the biomedical community. As part of that effort, we are constantly improving the web interface for existing tools and services and designing new tools to make structural data available. In the SIFTS project, we integrate structural data with other biological data to facilitate discovery. These integrated data form the basis for many query interfaces that allow macromolecular structure data to be presented in its biological context. Our specific focus areas are: data integrity, data quality, integration and data dissemination to the non-expert biomedical community.

SUMMARY OF PROGRESS

- Improved data integrity for the PDB and EMDB archives through wwPDB and EMDataBank remediation efforts that addressed multiple issues;
- Improved data integration (SIFTS) infrastructure to produce more accurate assignments of GO and InterPro cross-references to PDB data. The new infrastructure facilitates tracking changes and additions of new cross-references from Pfam, CATH, SCOP, IntEnz, GO and InterPro;
- Launched a number of new tools, including PDBportfolio (pictorial summaries of PDB entries), UniPDB (schematic views of structural coverage of UniProt entries) and new modules for the PDBeXplore structure browser (by chemistry, taxonomy and GO classification);
- Launched an article series called Quips (Quite Interesting Pdb Structures), which features short stories about selected structures coupled with an interactive structure viewer;
- Developed several new tutorials and on-line training modules;
- Conducted 11 roadshows to train PDBe users and reach new user communities. For the first time, PDBe presented roadshows outside of Europe: in Australia and South Africa;
- Expanded PDBe's presence in social media to inform its users of new developments.

MAJOR ACHIEVEMENTS

In 2011 PDBe annotation staff curated a record number of PDB entries, the majority of which were annotated within one working day of being deposited. The year saw 1496 entries deposited to PDB (up 9% from 2010). One of these entries was the 10 000th entry deposited and annotated at PDBe since PDB deposition and annotation services were established at EMBL-EBI in 1998. PDBe staff also annotated 163 EMDB entries, representing 58% of all EMDB entries deposited in 2011.

PDBe staff were also involved in a new release of the PDB archive (v. 4.0) that addressed data issues related to residual B-factors, annotation and representation of peptide inhibitors and antibiotics, representation of X-ray entries in correct crystallographic conventions, issues related to partial occupancy of atoms, incomplete coordinate sets, polymers containing non-standard linkages and taxonomy information related to strain names.

EMDB data was remediated to improve the quality and consistency of the archive. This involved ensuring a consistent representation of data in the archive, correcting data and adding missing information such as journal information, defocus units, author information and contour levels.

Our efforts to train PDBe users were bolstered by the introduction of our Quips article series. These articles, often authored by experts from the community, explain one or more biologically important structures to a non-expert audience and are accompanied by animated 3D molecular graphics views to help with exploration of macromolecular structure data. A Quips article on autotaxin was developed specifically to accompany a review in the journal *Nature Reviews Molecular Cell Biology.*

Since 2000 PDBe has provided the valuable data resource Structure Integration with Function, Taxonomy and Sequences (SIFTS) in close collaboration with UniProt. SIFTS is used by major bioinformatics resources and underpins many PDBe tools and services. Within SIFTS, the mapping of GO and InterPro assignments was improved in 2011 and the production process was redesigned to make it more robust. This led to the development of new tools (PDBeXplore modules) that allow browsing of structural data based on GO data and taxonomy data. An additional PDBeXplore interface was developed that allows browsing of PDB data based on small-molecule information in the PDB archive.

In addition to the new PDBeXplore modules, we developed a tool that shows the structural coverage in the PDB for a specified UniProt entry (UniPDB). Furthermore, PDBportfolio shows pictorial representations of various functional and structural properties of a PDB entry, including annotation of Pfam and CATH domains on the 3D structure. It also shows the quaternary structure, ligand-binding environment and some experimental information.

The PDBe team worked with a number of users to understand how our the usability of our website could be improved. The study was carried out during roadshows in Australia, Sweden and South Africa. User feedback from these roadshows informed changes to the PDBe website, such as improvements to the atlas pages for PDB entries and the layout of the front page.

PDBe expanded its presence in various social media (i.e. Facebook and Twitter), reaching a large audience of users and potential users (several hundred followers and 'likes'). We use these platforms to inform users of new developments and to get feedback on existing services. Based on feedback, we now post weekly release statistics to both sites.

Selected publications

Velankar, S., Alhroub, Y., *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 39 (Database issue), D402-10.

Velankar, S. and Kleywegt, G.J. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. *Acta Crystallogr. D Biol. Crystallogr.* 67 (Pt 4), 324-30.

FUTURE PLANS

Our goal is to make biomacromolecular structure data available to non-experts in the wider biomedical community so they can exploit the wealth of information available in the PDB and EMDB. To that end, we will continue to concentrate on improving data quality and the integrity of the archives. We will advance our efforts to integrate structural data with other relevant biomedical information. To make it easy for non-expert users to assess the quality of any macromolecular structure, we will work on validation mechanisms and develop an intuitive interface to display their results. We will continue to make our data available to other biomedical resources and to improve our services and web site. The upgrades will be released in six-month cycles.

In 2012 the new joint wwPDB Deposition and Annotation system will be released (see Tom Oldfield). This will mean a major change to the annotation practices at PDBe and we will work closely with our wwPDB partners to ensure an efficient transition. We will also start a complete redesign process for the PDB and EMDB web pages and search systems, to be released in 2013.

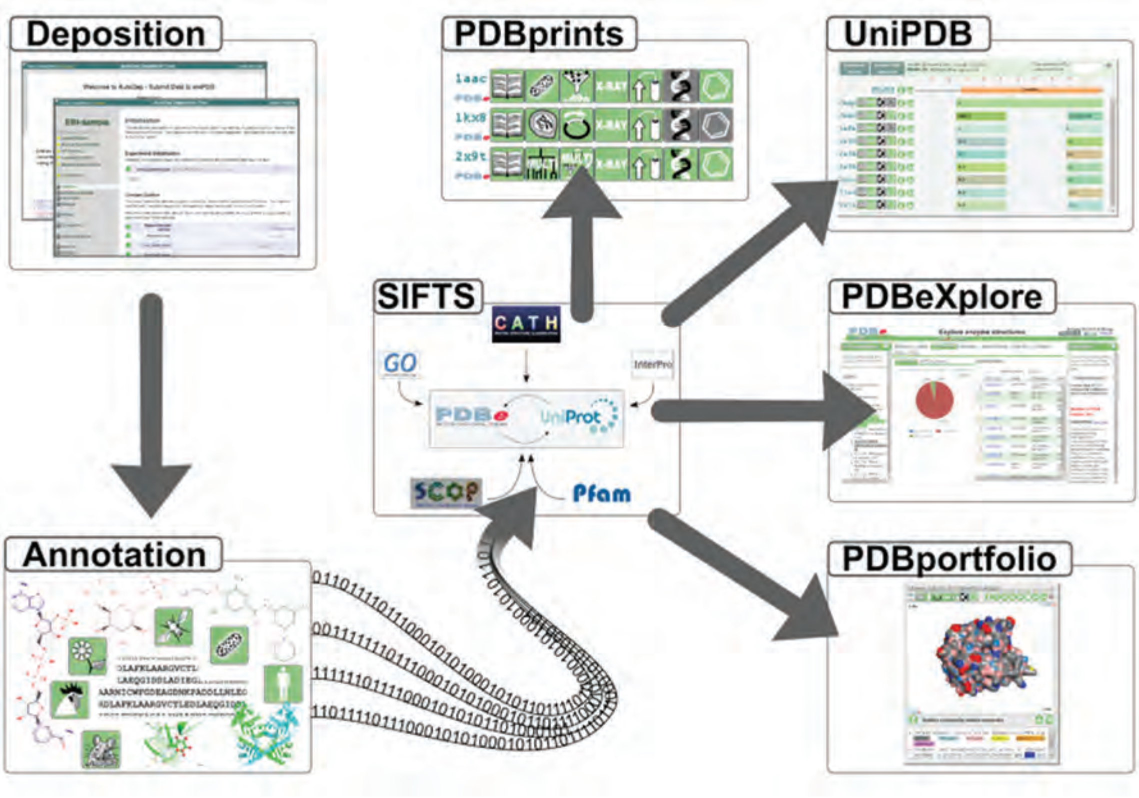


Figure. EMDB and PDB deposition and annotation provides the basis for data integration efforts such as Structure Integration with Function, Taxonomy and Sequences (SIFTS). The SIFTS resource underpins the development of various tools and services that facilitate dissemination of biomacromolecular structure data to the biomedical community.

Literature services

Johanna McEntyre

PhD in plant biology, Manchester Metropolitan University, 1990. Editor, Trends in Biochemical Sciences, Elsevier, Cambridge, UK, 1997. Staff Scientist, NCBI, National Library of Medicine, NIH, USA, 2009. Team Leader at EMBL-EBI since 2009.

DESCRIPTION OF SERVICES AND RESEARCH

The scientific literature often represents both the start and end point of a scientific project. However, as with other biomedical data resources, the volume of articles published is overwhelming, too much for any one person to read. Placing the literature within the context of related public data resources will equip researchers better for data analysis, navigation and discovery.

With several thousand new research articles published every day, linking articles to each other – and to the broader scientific literature such as textbooks, theses and patents – will become a necessity if we are to leverage the investment in scientific research to greater potential. Text-mining represents a high-throughput approach to the identification of biological concepts in articles, which can then form the basis for the development of new applications and stimulate precise, deep linking to related data resources in the future.

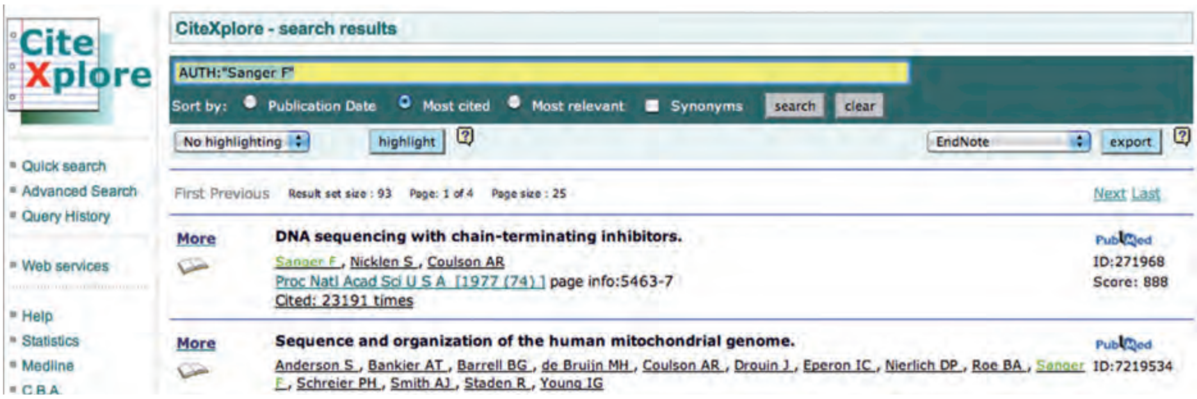
The goal of Literature Services at EMBL-EBI is to build text-based resources for the life sciences, integrated with other public-domain data resources hosted at EMBL-EBI. To this end, we run two literature databases: CiteXplore and UK PubMed Central. CiteXplore contains over 26 million abstracts and includes PubMed as well as data from Agricola and patents from the European Patent Office. UK PubMed Central comprises over 2 million full-text articles, of which about 400 000 are open access.

The databases are linked to a number of EMBL-EBI data resources by (1) using the references appended to database records by curators and submitters, and (2) through text mining to identify terms of interest, such as gene symbols, and using these to link to appropriate databases.

We also calculate citation network information for the records we hold: over 10 million articles have been cited at least once, representing the largest public-domain citation network in the world. We plan to use this infrastructure to develop novel and useful search and browse features for publications mapped to data, and to share the article content and annotation as widely as possible, both programmatically and for individual users.

SUMMARY OF PROGRESS

- Assumed leadership of UK PubMed Central (April 2011 – April 2016).
- Improved technical coordination of UK PubMed Central by serving all components (database, indexing, application) from EMBL-EBI.
- Handled over 8 million web service requests per month.
- Introduced citation-count sort order for our citation network of over 10 million cited articles.
- Partnered with OpenAIRE Plus to explore dataset management options.



AUTH:"Sanger F"

Sort by: ○ Publication Date ● Most cited ○ Most relevant □ Synonyms

Result set size : 93 Page: 1 of 4 Page size : 25

DNA sequencing with chain-terminating inhibitors.
Sanger F , Nicklen S , Coulson AR
Proc Natl Acad Sci U S A [1977 (74)] page info:5463-7
Cited: 23191 times
PubMed ID:271968 Score: 888

Sequence and organization of the human mitochondrial genome.
Anderson S , Bankier AT , Barrell BG , de Bruijn MH , Coulson AR , Drouin J , Eperon IC , Nierlich DP , Roe BA , Sanger F , Schreier PH , Smith AJ , Staden R , Young IG
PubMed ID:7219534

Figure 1. CiteXplore search result.

MAJOR ACHIEVEMENTS

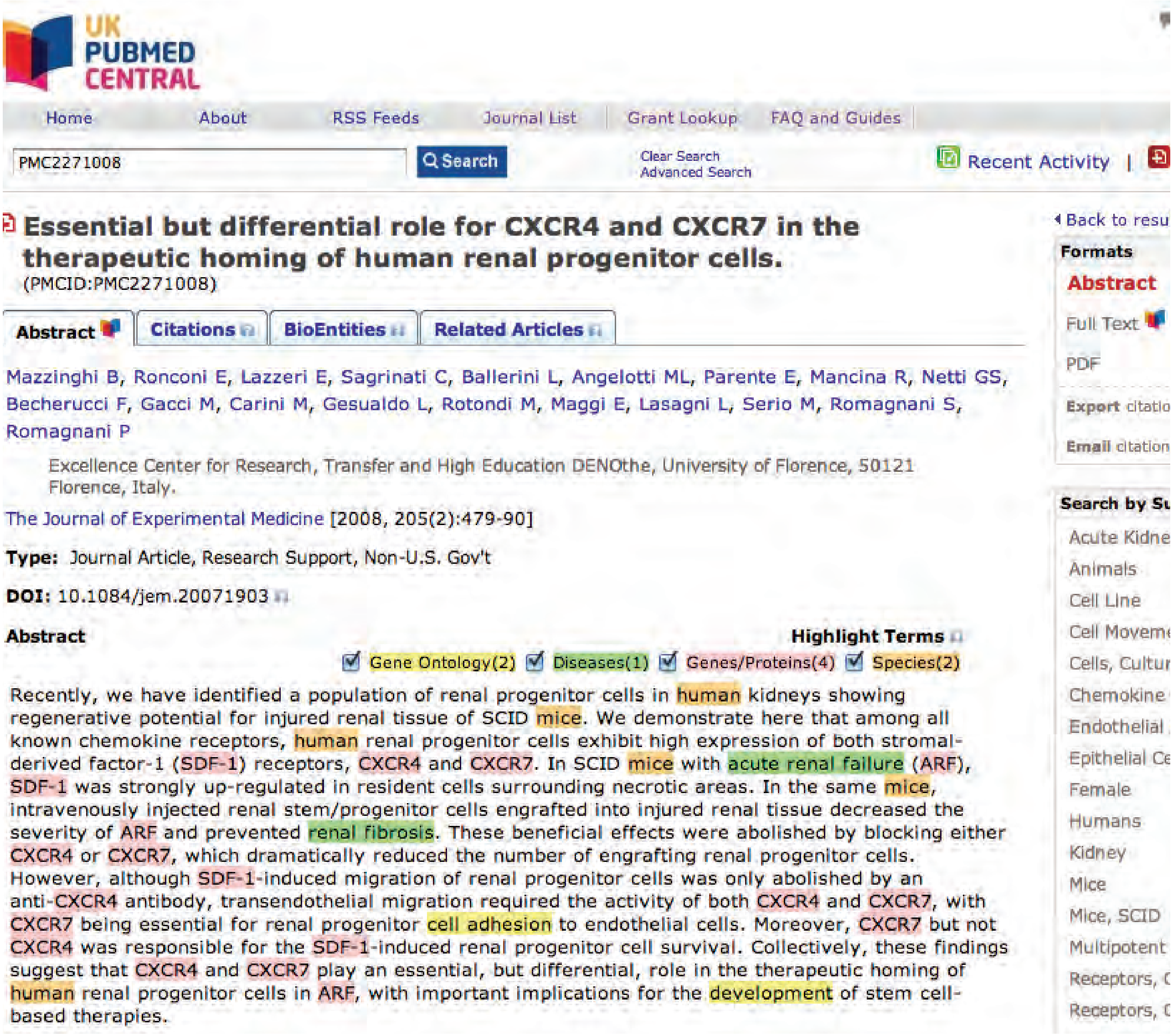
The major achievements of the Literature Services centred on winning the contract and grant ward to run UKPMC from April 2011, with the service running fully from EMBL-EBI by July 2011. This new and high-profile responsibility for EMBL-EBI will give rise to future integration and development opportunities impacting all data resources at EMBL-EBI. Furthermore, it is a step towards realizing the vision of UKPMC – to evolve into a European resource of international importance.

FUTURE PLANS

Building on the article collections and websites we host at EMBL-EBI, we plan to extend UK PubMed Central and CiteXplore into a cohesive literature search-and-retrieval system that represents European science on a global scale. In the near future, we will add the full text of a number of books and reports alongside the research article collection. We will also further the concept of 'Literature Labs', which will be an opportunity for us to engage actively with the text-mining community in road-testing applications based on UK PubMed Central and CiteXplore content that will enable us to integrate the literature with public data resources more deeply.

Selected publication

McEntyre, J.R. Ananiadou, S., *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.* 39 (Database issue), D58-65.



Essential but differential role for CXCR4 and CXCR7 in the therapeutic homing of human renal progenitor cells.
(PMCID:PMC2271008)

Mazzinghi B, Ronconi E, Lazzeri E, Sagrinati C, Ballerini L, Angelotti ML, Parente E, Mancina R, Netti GS, Becherucci F, Gacci M, Carini M, Gesualdo L, Rotondi M, Maggi E, Lasagni L, Serio M, Romagnani S, Romagnani P

Excellence Center for Research, Transfer and High Education DENOthe, University of Florence, 50121 Florence, Italy.

The Journal of Experimental Medicine [2008, 205(2):479-90]

Type: Journal Article, Research Support, Non-U.S. Gov't

DOI: 10.1084/jem.20071903

Abstract

Highlight Terms

☑ Gene Ontology(2) ☑ Diseases(1) ☑ Genes/Proteins(4) ☑ Species(2)

Recently, we have identified a population of renal progenitor cells in human kidneys showing regenerative potential for injured renal tissue of SCID mice. We demonstrate here that among all known chemokine receptors, human renal progenitor cells exhibit high expression of both stromal-derived factor-1 (SDF-1) receptors, CXCR4 and CXCR7. In SCID mice with acute renal failure (ARF), SDF-1 was strongly up-regulated in resident cells surrounding necrotic areas. In the same mice, intravenously injected renal stem/progenitor cells engrafted into injured renal tissue decreased the severity of ARF and prevented renal fibrosis. These beneficial effects were abolished by blocking either CXCR4 or CXCR7, which dramatically reduced the number of engrafting renal progenitor cells. However, although SDF-1-induced migration of renal progenitor cells was only abolished by an anti-CXCR4 antibody, transendothelial migration required the activity of both CXCR4 and CXCR7, with CXCR7 being essential for renal progenitor cell adhesion to endothelial cells. Moreover, CXCR7 but not CXCR4 was responsible for the SDF-1-induced renal progenitor cell survival. Collectively, these findings suggest that CXCR4 and CXCR7 play an essential, but differential, role in the therapeutic homing of human renal progenitor cells in ARF, with important implications for the development of stem cell-based therapies.

Figure 2. UK PubMed Central in 2011. EMBL-EBI will be leading the project until 2016.

Developing and integrating tools for biologists

Peter Rice

BSc University of Liverpool, 1976. EMBL Heidelberg, 1987-1994. Sanger Centre 1994-2000. LION Bioscience 2000-2002. At EMBL-EBI since 2003.

DESCRIPTION OF SERVICES AND RESEARCH

The team focuses on the integration of bioinformatics tools and data resources. We also investigate and advise on the e-Science and Grid technology requirements of EMBL-EBI through application development, training exercises and participation in international projects and standards development. We are responsible for the development of the EMBOSS open-source sequence-analysis package and for the EMBRACE project, which integrates access to bioinformatics tools and data content through standard-compliant web services.

SUMMARY OF PROGRESS

- Issued two releases of EMBOSS;
- Developed the EDAM ontology of bioinformatics data types and methods.

MAJOR ACHIEVEMENTS

The EMBOSS development work, fully funded by the Biotechnology and Biological Sciences Research Council (BBSRC), was carried out by a team of three developers. We succeeded in catching up on a backlog of maintenance and feature requests, and released two versions of EMBOSS (6.4.0 on 15 July and 6.5.0 ready for release in early 2012 when the BBSRC grant ends). These releases support the standardisation of EMBOSS internals for three books we have published with Cambridge University Press. In 2011 we concentrated on new developments and added many extensions to the current developers' version of the code.

The EMBOSS website was completely remodelled and released in early 2012. The content is modelled on the three books, with automated updates from the most recent developers' code.

We re-wrote the query language used to access data; now, EMBOSS supports multiple queries across multiple fields for databases and for file-based inputs. Server definitions allow several thousand public data resources to be automatically defined for all EMBOSS installations.

We extended EMBOSS input formats to include standard next-generation sequence formats, in particular the SAM and BAM formats used by SAMtools and other packages and the various versions of FASTQ format first used at the Wellcome

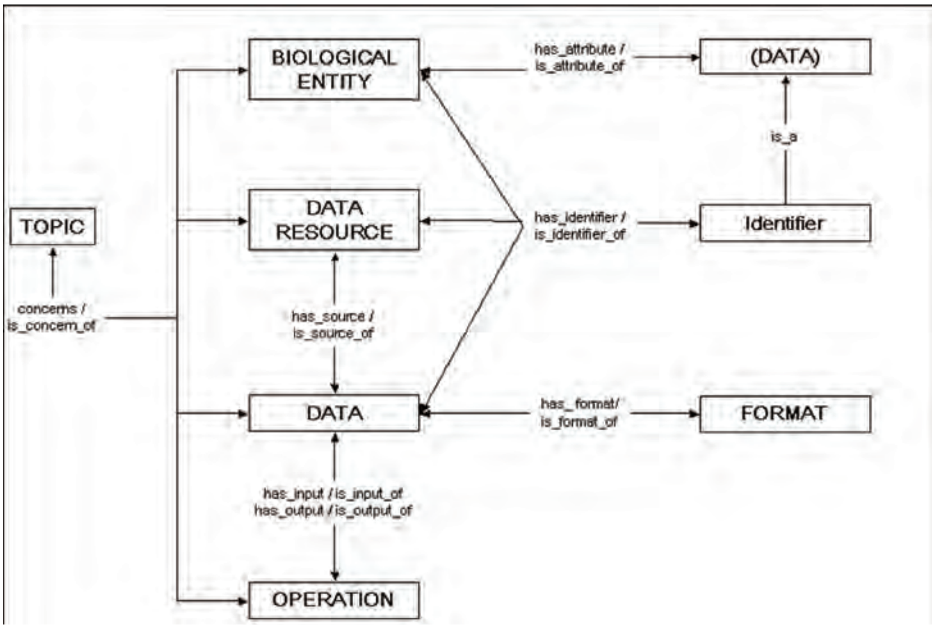
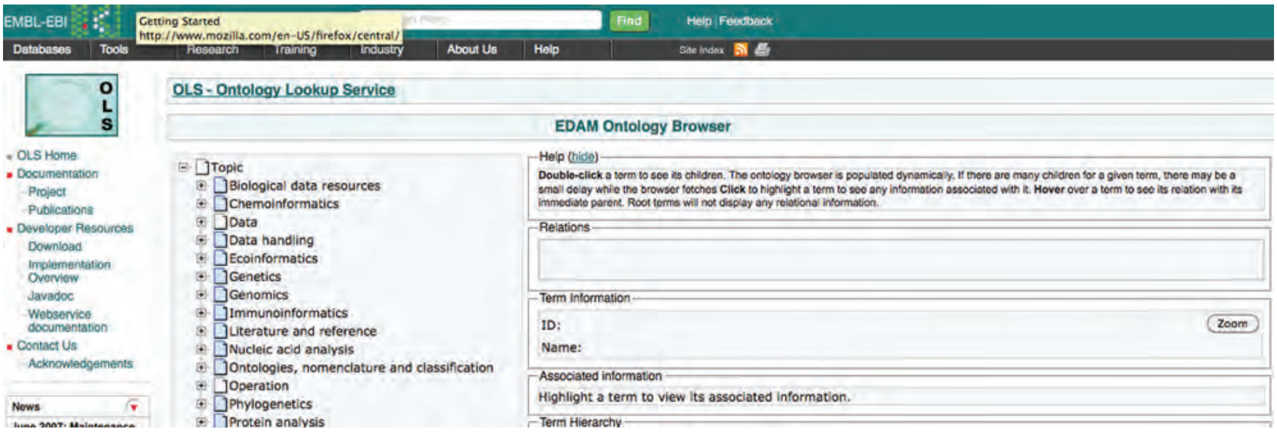


Figure 1. EDAM terms and relations.



- Topic
  - Biological data resources
  - Chemoinformatics
  - Data
  - Data handling
  - Ecoinformatics
  - Genetics
  - Genomics
  - Immunoinformatics
  - Literature and reference
  - Nucleic acid analysis
  - Ontologies, nomenclature and classification
  - Operation
  - Phylogenetics
  - Protein analysis

Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches **Click** to highlight a term to see any information associated with it. **Hover** over a term to see its relation with its immediate parent. Root terms will not display any relational information.

ID:

Name:

Highlight a term to view its associated information.

Figure 2. EDAM ontology browser in the Ontology Lookup Service.

Trust Sanger Institute. These formats are now used to populate sequence-assembly objects containing mapped reads from regions of an assembly, and can be used to stream assembly data when converting to an alternative format.

As a result of our efforts in 2011, EMBOSS supports genome-scale reference sequences, including support for building EMBL database contig entries from multiple entries in the whole genome shotgun division. Sequence annotation now fully supports GFF3 standards, with detailed discussion of special cases among developers in other Open-Bio Foundation projects.

Our team supports a C version of the Ensembl API, contributed by Michael Schuster in the Ensembl team at EMBL-EBI (see page X). This allows developers to use EMBOSS to create efficient applications to process large amounts of Ensembl-derived data in C rather than Perl.

Where EMBOSS applications depend on other packages (especially for the third-party EMBASSY packages) these applications are now checked when the EMBOSS application is first started, giving an immediate run-time error with a consistent message explaining how to provide the location of the external program.

For EMBRACE we recognised the need to provide semantic-level annotation of the many data and tools services. Such annotation enables the EMBRACE registry to provide descriptive searches for service discovery and for the semantic joining of the output of one service to the input of another. Our solution was to develop a new ontology: EDAM (EMBRACE Datatypes And Methods), which has over 2000 terms covering operations, topics (e.g. sequence analysis), data types, data formats and data resources. Using these terms we were able to annotate over 200 EMBOSS applications comprehensively by including EDAM references in their ACD command definition files. These annotations are then automatically transcribed into the WSDL web service definition files for SoapLab web services that launch these same EMBOSS applications. The first full release of EDAM will appear early in 2012, fully supported by EMBOSS applications.

To cover the full set of public data resources, EMBOSS has created a data resource catalogue (DRCAT) which was included in the most recent release and defines data resources and their public interfaces in ways that allow automated access to each resource from within EMBOSS, either directly or through cross-references from other input sources.

FUTURE PLANS

In 2012 the EMBOSS package will no longer be developed at EMBL-EBI. Development will continue with contributions from the recent development team, with continued support through the EMBOSS mailing lists, Wiki and website. The EMBOSS user and developer communities will be surveyed to determine future directions for the project. The EMBRACE consortium ends in July 2012. Our work on EMBRACE-compliant services will continue with the further development of the EDAM ontology as part of EMBOSS and the maintenance of fully annotated SOAP services within SoapLab, also through EMBOSS. DRCAT will be developed further, in collaboration with other catalogues of public resources.

Services Teams

PROTEIN AND NUCLEOTIDE TEAMS

Joint Team Leader (Proteins)

Rolf Apweiler

Joint Team Leader (Nucleotides)

Ewan Birney

Project Coordinators

Elspeth Bruford

Pascal Kahlem

Group Cooperation Officer

Chuck Cook

Curators and Bioinformaticians

Louise Daugherty

Ruth Seal

Matt Wright

Group Secretaries

Shelley Goddard

Tracy Mumford

PhD Students

Joe Foster

Markus Fritz

Dace Ruklisa

EIPOD Postdoctoral Fellow

Mikhail Spivakov

EUROPEAN NUCLEOTIDE ARCHIVE

Team Leader

Guy Cochrane

Bioinformaticians

Blaise Alako

Xin Liu

Marc Rossello

Senior Scientific Database Curator

Clara Amid

Scientific Database Curators

Ana Cerdeño-Tárraga

Richard Gibson

Petra Ten Hoopen

Software Engineers

Lawrence Bower

Iain Cleland

Neil Goodgame

Mikyung Jang

Arnaud Oisel

Nima Pakseresht

Swapna Pallreddy

Rajesh Radhakrishnan

Dmitriy Smirnov

Daniel Vaughan

Vadim Zalunin

EMBLBank Coordinator (Development)

Rasko Leinonen

Submissions Processing

Sheila Plaister

Visitor

Alexander Senf

VERTEBRATE GENOMICS

Senior Team Leader

Paul Flicek

Group Coordinators

Fiona Cunningham (Ensembl)

Javier Herrero (Ensembl)

Glenn Proctor* (Ensembl Software)

Project Leaders

Laura Clarke (Resequencing Informatics)

Ian Dunham (Ensembl Regulation)

Xose Fernandez (Ensembl Outreach

Rhoda Kinsella (Ensembl Production)

Gautier Koscielny* (Mouse Informatics)

Ilkka Lappalinen (Variation Archive)

Terry Meehan* (Mouse Informatics)

Damian Smedley (Mouse Informatics)

Giulietta Spudich (Ensembl Outreach)

Andy Yates* (Ensembl Core)

Scientific Programmers

Nathan Johnson

Damian Keefe*

Stephen Keenan

Vasudev Kumanduri

Michael Maguire

Pablo Marin-Garcia

David Richardson*

Richard Smith

Dylan Spalding*

Ian Streeter*

Albert Vilella

Steven Wilder

Holly Zheng Bradley

Ensembl Developers

Ikhlak Ahmed*

Kathryn Beal

Stephen Fitzgerald

Laurent Gil

Leo Gordon

Andreas K. Kähäri*

Monika Komorowska*

Pontus Larsson*

Ian Longden

Thomas Maurel*

Will McLaren

Matthieu Muffato*

Miguel Pignatelli*

Graham Ritchie

Daniel Sobral

Kieron Taylor*

Bioinformaticians

Chao-Kung Chen

Edoardo Marcora*

Phil Wilkinson

Curators

Jacqueline MacArthur*

Lisa Skipper

User Support Officers

Jeff Almeida-King

Denise Caravahlo-Silva*

Bert Overduin

Michael Schuster

Jana Vandrovcova*

Postdoctoral Fellows

Benoit Ballester*

David Thybert

PhD Students

Andre Faure

Thomas Rensch*

Petra Schwalie

Visitors

Jia-ming Chang*

Carsten Kemena*

Mateus Patricio*

Team Secretary

Kerry Smith

NON-VERTEBRATE GENOMICS

Team Leader

Paul Kersey

Coordinators

Daniel Bolser *

Kevin Howe *

Eugene Kulesha

Daniel Lawson
Daniel Staines

Bioinformaticians
Paul Davies *
Paul Derwent
Avazeh Ghanbarian *
Daniel Hughes
Uma Maheswari
Karyn Megy
Michael Nuhn
Michael Paulini *
Alessandro Vullo
Gary Williams *
Derek Wilson

Software Engineers
Matthias Haimel
Arnaud Kerhornou
Gautier Koscielny *
Andy Yates *

Web Developers
Jay Humphrey *
Nick Langridge
Iliana Toneva

Visitors
Mary Ann Tuli *

PROTEOMICS SERVICES

Team Leader
Henning Hermjakob

Bioinformaticians
David Croft
Attila Csordas
Marine Dumousseau
Pierre Grenon
David Ovelleiro
Rui Wang

Coordinators
Sandra Orchard
Juan Antonio Vizcaino

Curators (including senior curators)
Bernard de Bono
Margaret Duesbury
Phani Garapati
Bijay Jassal
Steven Jupe
Jyoti Khadake
Pablo Porras Millan
Mark Williams

Software Engineers (including senior software engineers)
Bruno Aranda
Richard Cote
Antonio Fabregat Mundo
Johannes Griss
Rafael Jimenez
Samuel Kerrien
Daniel Rios
Florian Reisinger
Chris Taylor
Sarala Wimalaratne

Visitors (including visiting students)
Phil Charles
Noemi Del Toro Ayllon
Jhon Gomez Carvajal
Gavin Koh
Nelson Ndegwa Gichora
Yasset Perez Riverol
Gustavo Salazar
Jose Villaveces
Matthieu Visser

INTERPRO

Team leader
Sarah Hunter

Database Software Engineering Group
Matt Corbett
Conor McMenamin *

Annotation Coordinator
David Lonsdale *

Content Coordinator
Alex Mitchell *

Scientific Database Curators
Sarah Burge *
Prudence Mutowo *
Amaia Sangrador

Bioinformaticians
Chris Hunter
Craig McAnulla
Siew-Yit Yong

Software Development Coordinator
Phil Jones

Senior Software Engineers
John Maslen *
Antony Quinn

Database Production Manager
Ujjwal Das *

Web Developer
Sebastien Pesseat

Software Developers
Matthew Fraser *
Maxim Scheremetjew *

UNIPROT: DEVELOPMENT

Team Leader
Maria Martin

Project leaders
Alexander Fedotov/Jie Luo
Samuel Patient

Software Engineers
Ricardo Antunes
Elisabet Barrera Casanova
Benoit Bely
Borisas Bursteinas
Francesco Fazzini
Leyla Jael Garcia Castro
Wudong Liu
Nikolas Pontikos
Steven Rosanoff
Tony Sawford
Edward Turner
Xavier Watkins
Tony Wardell
Alan Wilter Sousa da Silva
Hermann Zellner

Web Developers
Mark Bingley

Bioinformaticians
Diego Poggioli

User Experience Analyst
Sangya Pundir

Trainees
Carlos Bonilla
Michiel Schneider

UNIPROT: CONTENT

Team Leader
Claire O'Donovan

Project Leaders
Emily Dimmer
Michele Magrane

Senior Scientific Database Curators
Wei Mun Chan
Rachael Huntley

Scientific Database Curators
Yasmin Alam-Faruque
Gayatri Chavali
Elena Cibrian-Uhalte
Reija Hieta
Rachel Jones
Duncan Legge
Prudence Mutowo
Klemens Pichler
Harminder Sehra

Bioinformatician
Julius Jacobsen

Visitor
Lorna Richardson
Eleanor Stanley

CHEMBL

Team Leader
John Overington

Group Coordinator
Anne Hersey

Data Mining and Analysis
Francis Atkinson

Data Manager
Ruth Akhtar

Chemical Content Curator
Louisa Bellis

Data Integration
Jon Chambers
Anna Gaulton

Web Application Developers
Mark Davies
Shaun McGlinchey

Scientific Application Developer
Kazuyoshi Ikeda
Furkan Yegin

Biologial Content Curator
Yvonne Light

Postdoctoral Fellow
Patricia Bento

PhD Students
Felix Krueger
Rita Santos
Ben Stauch

Visiting Scientists
Bissan Al-Lazikani
Lee Harland
Roger Sayle

CHEMINFORMATICS AND METABOLISM

Team Leader
Christoph Steinbeck

Bioinformaticians
Kalai Vanii Jayaseelan
Janna Hastings

Coordinators
Paula de Matos

Curators (including senior curators)

Marcus Ennis
Zara Josephs
Gareth Owen
Steven Turner

Software Engineers (including senior software engineers)

Rafael Alcántara Martin
Hong Cao
Pablo Conesa Mingo
Adriano Dekker
Kenneth Haugh
Joseph Onwubiko
Mark Rijnbeek
Amit Walinjkar

PhD Students

Stephan Beisken
John May
Pablo Moreno

Postdoctoral Fellows

Luis de Figueiredo

Visitors (including visiting students)

Merche Castillo
Samy Deghou
Duan Lian
Peter Murray-Rust
Venkatesh Muthukrishnan
Reza Salek
Andreas Truszkowski

FUNCTIONAL GENOMICS

Senior Team Leader

Alvis Brazma

Biomedical Statistician

Johan Rung

Bioinformaticians

Liliana Greger
Jing Su

Software/Web Developer

Catherine Kirsanova

Research/Training Coordinator

Gabriella Rustici

PhD students

Angela Goncalves
Mar Gonzalez-Porta

Group Secretary

Lynn French

Administrative Assistant

Kathryn Holmes

Visitor

Aurora Torrente

FUNCTIONAL GENOMICS ATLAS

Team Leader

Misha Kapushesky

Software Engineers

Alexey Filippov
Olga Melnichuk
Robert Petryszak
Nataliya Sklyar
Andrew Tikhonov

Research Fellow

Wanseon Lee

Bioinformatician

Nikolay Pultsin

FUNCTIONAL GENOMICS PRODUCTION

Team Leader

Helen Parkinson

Bioinformaticians

Tomasz Adamusiak
Anna Farne
Adam Faulconbridge
Emma Hastings
Ele Holloway
Simon Jupp
Maria Keays
Natalja Kurbatova
James Malone
Amy Tang
Julie Taylor
Tobias Ternent
Ravensara Travillian
Danielle Welter
Eleanor Williams

Software Engineers

Tony Burdett
Jon Ison

Visitors

Morris Swertz
Drashtti Vasant
Vincent Xue

FUNCTIONAL GENOMICS DEVELOPMENT

Technical Team Leader

Ugis Sarkans

Coordinator

Nikolay Kolesnikov

Software Engineers

Marco Brandizi
Miroslaw Dylag
Ibrahim Emam
Ekaterina Pilicheva
Rui Pereira*
Anjan Sharma*

Data Manager

Stathis Kanterakis*

Database Administrator

Roby Mani

Visiting Student

Pauls Vasilis*

PROTEIN DATA BANK IN EUROPE

Senior Team Leader

Gerard Kleywegt

Team Leaders

Tom Oldfield (Databases and Services)
Sameer Velankar (Content and Integration)

Project Leaders

Aleksandras Gutmanas (NMR)
Miriam Hirshberg (QA)
Ardan Patwardhan (EMDB)
Sanchayita Sen (Curation)
Jawahar Swaminathan (Curation)*

Administrator

Pauline Haslam

Curators

Matthew Conroy
Gaurav Sahni
Martyn Symmons

Software Engineers

Younes Alhroub
Jose Dana
Manuel Fernandez Montecelo*
Glen van Ginkel
Swanand Gore
Pieter Hendrickx
Ingvar Lagerstedt
Saqib Mir
Luana Rinaldi
Eduardo Sanz-Garcia*
Michael Wainwright

Database Administrator

Robert Slowley

Visitors

Egon Heuson*
Laurence Newman*

LITERATURE SERVICES

Team Leader

Johanna McEntyre

Software Engineers

Paula Buttery
Norman Cobley
Alan Horne
Stephen Spencer

Literature Analyst

Andrew Caines

Developers

Yuci Gou*
Jyothi Katuri
Oliver Kilian*
Nikos Marinos*
Xingjun Pi*
Xiaofei Wang*

Project Manager

Philip Rossiter*

DEVELOPING AND INTEGRATING TOOLS FOR BIOLOGISTS

Team Leader

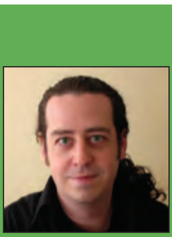
Peter Rice

Software Engineers

Alan Bleasby
Jon Ison
Mahmut Uludag

Research

Bertone Group: Pluripotency, reprogramming and differentiation



Paul Bertone

PhD Yale University, 2005. At EMBL–EBI since 2005. Investigator, Stem Cell Institute, University of Cambridge. Joint appointments in Genome Biology and Developmental Biology Units.

DESCRIPTION OF RESEARCH

We investigate the cellular and molecular processes underlying mammalian stem cell biology using a combination of experimental and computational approaches. 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.

Pluripotent ES cells can produce lineage-specific precursors and tissue-specific stem cells, with an accompanying restriction in commitment potential. These exist in vivo as self-renewing multipotent progenitors localised in reservoirs within developed organs and tissues. The properties of proliferation, differentiation and lineage specialisation are fundamental to cellular diversification and growth patterning during organismal development, as well as the initiation of cellular repair processes throughout life.

Our research group applies the latest high-throughput technologies to investigate the functions of key regulatory proteins and their influence on the changing transcriptome. We focus on early lineage commitment of ES cells, neural differentiation and nuclear reprogramming. The generation of large-scale data from functional genomic and proteomic experiments will help to identify and characterise the regulatory influence of key transcription factors, signalling genes and non-coding RNAs involved in early developmental pathways, leading to a more detailed understanding of the molecular mechanisms of vertebrate embryogenesis.

SUMMARY OF PROGRESS

- Mapped genome-wide binding sites of key pluripotency factors and chromatin modifications in mouse embryonic stem cells;
- Determined major molecular characteristics of human brain cancer stem cells;
- Developed and optimised protocols for comprehensive RNA sequencing;
- Resolved the complete transcriptomes of stem cells at various developmental stages.

MAJOR ACHIEVEMENTS

Cellular differentiation is normally a one-way process. Remarkably, induced pluripotent stem (iPS) cells can now be generated from various somatic cell types via transduction of reprogramming factors. Early attempts to revert differentiated cells into an ES cell-like state suffered from one or more notable deficiencies that indicated iPS cells were not truly pluripotent, where reprogramming is stalled at an incompletely pre-pluripotent (pre-iPS) stage. We are undertaking several related projects to characterise the reversion of differentiated cells to a pluripotent state, in collaboration with Austin Smith and Jose Silva at the Wellcome Trust Centre for Stem Cell Research. To investigate this process we are applying the ChIP-seq approach to map direct targets of regulatory factors in mouse ES cells, excluding potential non-specific interactions with control samples derived from ES cells genetically devoid of the factors of interest. These are propagated in culture through chemical inhibition of Mek/Erk pathways and glycogen synthase kinase-3 (Gsk-3). The combination of leukemia inhibitory factor (LIF) exposure and inhibition of Mek/Erk signalling is sufficient to convert pre-iPS cells rapidly and at high efficiency into authentic iPS cells.

The fundamental processes that regulate cell differentiation are not well understood and are likely to be misregulated in cancer. A second focus in the lab is the study of neural cancer stem cells derived from human glioma multiforme tumours. These glioma neural stem (GNS) cells have been isolated and expanded using the same culture conditions previously used for the establishment of normal neural stem cells. The normal and diseased counterparts are morphologically and immunohistologically indistinguishable, yet the differentiation behaviour of the cancer stem cells is clearly aberrant and they are able to give rise to authentic tumors upon xenotransplantation.

Together with Steven Pollard at University College London, we applied a combination of sequencing and microarray approaches to determine the genetic architecture of individual GNS cell lines and the variations unique to each. Together with comprehensive transcriptome sequencing, these data provide a unified view of genomic aberrations and global expression patterns that contribute to the cancer state. We are further characterising the differentiation of GNS cells to neurons and oligodendrocytes, an event that is positively correlated with patient survival rates in cases of glioblastoma multiforme.

FUTURE PLANS

We will continue working to understand the molecular mechanisms that support pluripotency in ground-state embryonic stem cells, and to map the transition between the pluripotent state and early lineage commitment. We also plan to use the ChIP-seq approach to capture the epigenetic status of cells undergoing reversion to pluripotency. It is believed that a stabilising process in lineage selection involves the progressive restriction of transcriptional potential of cells as they transition through the lineage hierarchy, mediated through chromatin modifications. This hypothesis suggests that subsequent induction of somatic cells to a pluripotent state would then invoke widespread epigenetic erasure, in order to restore the cell to a state where global lineage commitment options are available. We will also further characterise the molecular properties of neural cancer stem cells, and assess the role of genetic aberrations and variation across individuals in the multipotent capacity of cell lines of different origins. This will involve genome and transcriptome sequencing, time-course expression profiling and functional experiments to identify alterations in disease versus normal cell types.

Selected publications

Reynolds, N., Salmon-Divon, M., *et al*. (2011) NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression. *EMBO J.* 31, 593-605.

Albers, C.A., Cvejic, A., *et al.* (2011) Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat. Genet.* 43 (8), 735-7.

Git, A., *et al.* (2010) Systematic comparison of microarray profiling, real-time PCR and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* 16, 991-1006.

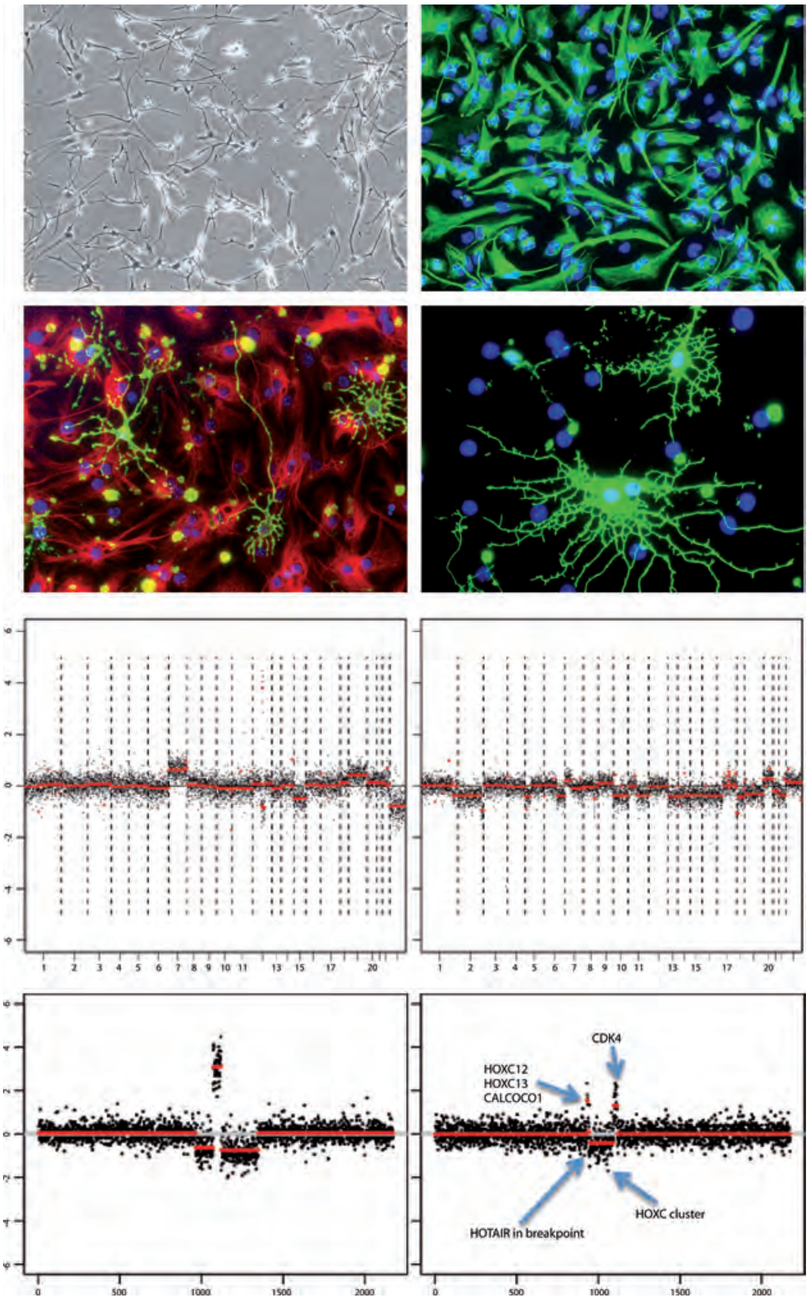
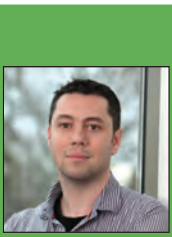


Figure. Neural cancer stem cells propagate indefinitely in culture (top) and can differentiate into the major cell types of the central nervous system, such as astrocytes and oligodendrocytes (second row). Array CGH and genome resequencing identify chromosomal abnormalities (third row) and the disruption of genes affected by them (bottom row).

Enright Group: Functional genomics and analysis of small RNA function



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.

DESCRIPTION OF RESEARCH

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 unannotated 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

- Published a new and exhaustive study on piwi-bound RNAs in the mouse germline with the O'Carroll lab in EMBL Monterotondo;
- Developed the Kraken pipeline including REAPER, new tools for extremely rapid and efficient analysis and processing of next-generation sequencing data;
- Joined the SIROCCO consortium as an analysis partner on the characterisation of non-coding RNAs in plants and animals.

MAJOR ACHIEVEMENTS

A new model for piwiRNAs in the Mouse germline. We collaborated with Donál O'Carroll at EMBL-Monterotondo on the analysis of piRNAs in the mouse germline. These molecules act as guardians of the germline and silence transposable elements, which become active following the demethylation required in germline development. Transposons can become a significant threat to the genome when their loci are demethylated, as they can become active and reintegrate elsewhere in the genome. Our work challenged the previous model of piwiRNA function, which held that piwiRNA transcripts have evolved in tandem with transposable elements (i.e. LINE1 and IAP elements) and act as complementary 'guide molecules', allowing a piwi protein to bind to a target transposon transcript. In this model, transposon silencing is carried out in a 'ping-pong' system in which two enzymes (Mili and Miwi2) use piwiRNAs to slice target transposons, generating more piwiRNAs that could further slice other transposon transcripts. The passing of piwiRNAs between these two molecules ostensibly caused an amplification effect that rapidly and effectively destroyed active transposons. We tested this model in the male mouse germline using a strategy that combined genetic and genomic approaches. The O'Carroll group generated mutant animals in which the slicing activity of either Mili or Miwi2 was mutated, allowing them to bind piwiRNAs but not to perform slicing. These proteins were immuno-precipitated and our group sequenced the population of bound piwiRNAs using next-generation sequencing technologies. Nenad Bartonicek was responsible for the analysis of the bound RNA populations found in both mutant and wild-type piwi proteins. We showed that the slicing activity of the Mili enzyme was required for effective transposon binding and silencing, while mutations to the Miwi2 enzyme had little or no effect on silencing activity. This work allowed us to generate a new model for piwiRNA function in the mouse in which Mili is responsible for binding, slicing and amplification and destroys transposons as they are produced, while Miwi2 is more likely to re-enter the nucleus and cause the transposon loci to be re-methylated.

New tools and pipelines for sequence analysis. As part of the SIROCCO consortium for the analysis of small RNAs in animals and plants, we generated a new set of tools and techniques for dealing with extremely large datasets. The group's Mat Davis and Stijn van Dongen developed the Kraken pipeline, which includes a novel algorithm called REAPER. The system is an extremely fast and lightweight approach written in C that scales to the kind of extremely large datasets that are being more commonly encountered in modern genomics. In 2011 we began the process of releasing this pipeline to the community and applying it to very large-scale datasets such as the piwiRNA project described above.

Long non-coding RNAs. Harpreet Saini and Mat Davis continued to develop strategies for analysing RNA-sequencing data for transcriptome reanalysis. We are working with the Furlong Lab in EMBL Heidelberg and the O'Carroll lab in EMBL Monterotondo to sequence and assemble long non-coding RNAs involved in the development of the Mouse germline and in *Drosophila* development.

Functional genomics and analysis of small RNA function. Evolutionary analysis of miRNAs was led by José Afonso Guerra-Assunção. Small, non-coding molecules do not lend themselves well to standard sequence-based phylogenetic approaches to understanding their evolution. Nevertheless, a great deal can be learned about the evolution of small RNA regulation in vertebrates. Our first approach was to identify likely orthologues and paralogues of known miRNAs across large numbers of species. The miRBase database of miRNA sequences comprises mainly species from which a miRNA was first isolated; it does not always search for its counterparts in diverse organisms. We developed a novel mapping strategy for identifying likely miRNA loci in multiple organisms based on a query miRNA, and mapped all miRBase miRNAs across the animal genomes available in Ensembl. We provided an online tool (MapMi; Guerra-Assuncao *et al.*, 2010) for performing this mapping or querying the results of the miRBase mapping. We also developed a system for large-scale exploration of the syntenic arrangement of miRNAs. We collaborated on a number of projects to assess the impact of single nucleotide polymorphisms (SNPs) between individuals at the level of miRNAs or their targets.

Selected publication

De Fazio, S., Bartonicek, N., *et al.* (2011) The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* 480, 259–263.

FUTURE PLANS

Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. Leonor Quintais will develop strategies for dealing with large-scale CLIP assays for microRNA target analysis. 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 will continue to build strong links with experimental laboratories that work on miRNAs in different systems. 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.

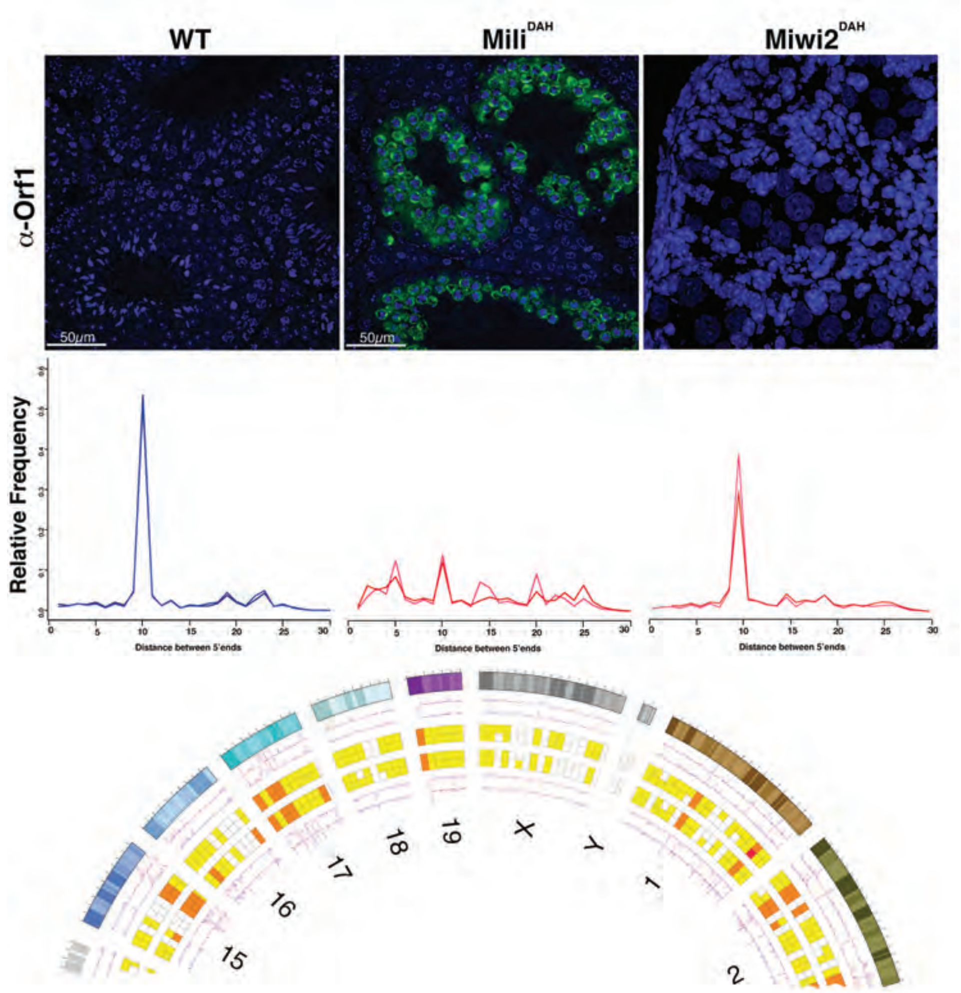


Figure. Top row: Immunoflourescent staining (green) for the protein (ORF1) expressed from LINE1 transposons when active and blue DAPI staining for nuclei in mouse testes sections. The wild-type sample shows no expression while the Mili DAH sample shows rampant transposon activity when the slicing function of Mili is mutated. The final column shows that no LINE1 activity is present when the slicing activity is mutated in Miwi2. The second row shows the frequency of 5'/3' overlaps of specific lengths from piwiRNAs sequenced from each sample. Normal 'ping-pong' activity is observed in the wild type, while this is defective in MiliDAH. Again the Miwi2DAH sample behaves similarly to wild type indicating that the slicing activity of this enzyme is not required for successful transposon silencing. The semi-circular plot shows how sequenced piwiRNAs map across a section of the mouse genome, with piwiRNA hotspots shown as orange and red areas on the heatmap.

Goldman Group: Evolutionary tools for genomic analysis

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.

DESCRIPTION OF RESEARCH

Our research concentrates on the mathematics and statistics of data analyses that use evolutionary information in sequence data and phylogenies to infer the history of living organisms, describe and understand processes of evolution and make predictions about the function of genomic sequence. We aim to increase our understanding of the process of evolution and provide new tools to elucidate the function of biological molecules as they change over evolutionary timescales. 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. In recent years, collaborations with sequencing consortia have provided essential state-of-the-art data and challenges to inspire, develop, and apply novel methods. Collaborations between group members who are involved in theoretical development and those who carry out comparative analysis of genomic data remain a stimulating source of inspiration in all of our research areas. Traditionally, the group has been strong in examining the theoretical foundations of phylogenetic reconstruction and analysis. In 2011 we continued to gain expertise in 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.

SUMMARY OF PROGRESS

- Expanded our phylogeny-aware alignment method PAGAN to perform extension of existing alignments with new data;
- Completed an investigation of the effects of alignment error and alignment filtering on detecting positive selection in proteins;
- Analysed protein-coding evolution in the African great apes, identifying genes undergoing accelerated evolution and estimating genome-wide levels of evolutionary constraint;
- Developed NG-SAM, a protocol for next generation sequencing of repetitive genome regions without the use of molecular cloning.

MAJOR ACHIEVEMENTS

Accurate multiple alignment of large data sets and sequences of very different length is demanding. We have expanded our phylogeny-aware alignment method PAGAN to perform extension of existing alignments with new data. Unlike alternative methods, PAGAN performs well under different evolutionary scenarios and provides superior accuracy for both DNA and protein data, the improvement being especially large for short sequences. The great accuracy of PAGAN-generated alignments of noisy reads extends the possible uses of NGS methods in evolutionary analyses.

We have also brought to completion an investigation of the effects of alignment error and alignment filtering on detecting positive selection in proteins. This study showed that different aligners have strikingly different tendencies to generate false positive results. The PRANK software, developed in the group (Löytynoja and Goldman, 2008), provided the best performance among the aligners tested. In contrast, an investigation of the effect of alignment filtering on phylogenetic inference and led us to conclude that in that context, filtering does more harm than good.

In collaboration with Nick Mundy and Stephen Montgomery (Department of Zoology, University of Cambridge), we analysed protein-coding evolution in the African great apes, identifying genes undergoing accelerated evolution in this clade and estimating genome-wide levels of evolutionary constraint. This analysis was contributed to the gorilla genome manuscript, which is currently in press.

AYB, our base-caller for the popular Illumina NGS platform, continues to improve and consistently produces more accurate reads than all available alternatives. Recent work shows that the amount of information about the sequence, as measured by the calibrated quality scores for each base, also exceeds alternatives and suggests that the reads produced are much more effective when used in downstream applications such as single nucleotide polymorphism (SNP) detection.

Life Technologies' latest Exact Call Chemistry is a new development of their SOLiD® next generation sequencing platform that allows the correction of miscalls in some circumstances. The new chemistry has a structure known as a 'convolutional code'; we examined its properties and showed both the circumstances in which errors can be corrected and what happens to the read when they cannot. Based on our insights into how this technology works, we produced several recommendations about how the nucleotide sequence archives (such as EMBL-EBI's Sequence Read Archive) should represent the data to make the best trade-off between faithful representation of the error structure of the reads and the space required to store them.

Despite the rise of next generation sequencing, the study of variability in repetitive regions remains a challenge due to the limited read length. One solution proposed for resolving such regions is Sequence Assembly aided by Mutagenesis (SAM), which relies on the fact that introducing enough random mutations makes the assembly possible. In collaboration with Jan Korbel's group at EMBL-Heidelberg we have developed NG-SAM, a version of the SAM protocol coupled to the NGS workflow without the use of molecular cloning. Using a realistic simulation pipeline in order to study the feasibility of the approach, we conclude that it may be successfully put into practice.

FUTURE PLANS

In 2012 we plan to: further characterise common sources of alignment error in genome-wide studies; develop new methods for detecting and mitigating such errors; and design more advanced methods for simulating protein-coding sequences. At a theoretical level, we will complete and draw conclusions from an ongoing study in phylogenetic inference, of 'long-branch attraction', which refers to the apparent propensity of long branches to 'congregate' in estimated phylogenies. Specifically, we hypothesise that long-branch attraction might result from the accumulation of independent deviations attributable to each long branch but not arising from any interactions among them. We will also attempt to assess the irreversibility of substitution processes observed in real biological sequences by comparing similarly parametrised reversible and irreversible Markov models.

Selected publications

Albers, C.A., Cvejic, A., *et al.* (2011) Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat. Genet.* 43 (8), 735-7.

Jordan, G. (2011) Analysis of alignment error and sitewise constraint in mammalian comparative genomics. *PhD thesis, EMBL and University of Cambridge*.

Jordan, G. and Goldman, N. (2011) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* Article in press.

Kosiol, C. and Goldman, N. (2011) Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. *J. Mol. Biol.* 411 (4), 910-23.

Lindblad-Toh, K., Garber, M., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

Washietl, S., Findeiss, S., *et al.* (2011) RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17 (4), 578-94.

We will continue to analyse raw data from emerging sequencing technologies, with an eye to improving the reads produced and understanding the correct way to represent the inherent uncertainties of the data in downstream analyses. The remaining time on our Wellcome Trust sequencing technologies grant will be spent looking at whether techniques for spectral correction of short reads can also be applied to recalibration.

As high-throughput RNA sequencing (RNA-seq) has become the most important method in transcriptomics, data analysis methods are proliferating – each accounting for more biological factors. To enable fair and insightful benchmarking of novel methods, we started development of a simulation tool for RNA-seq library construction and sequencing; release is expected in early 2012.

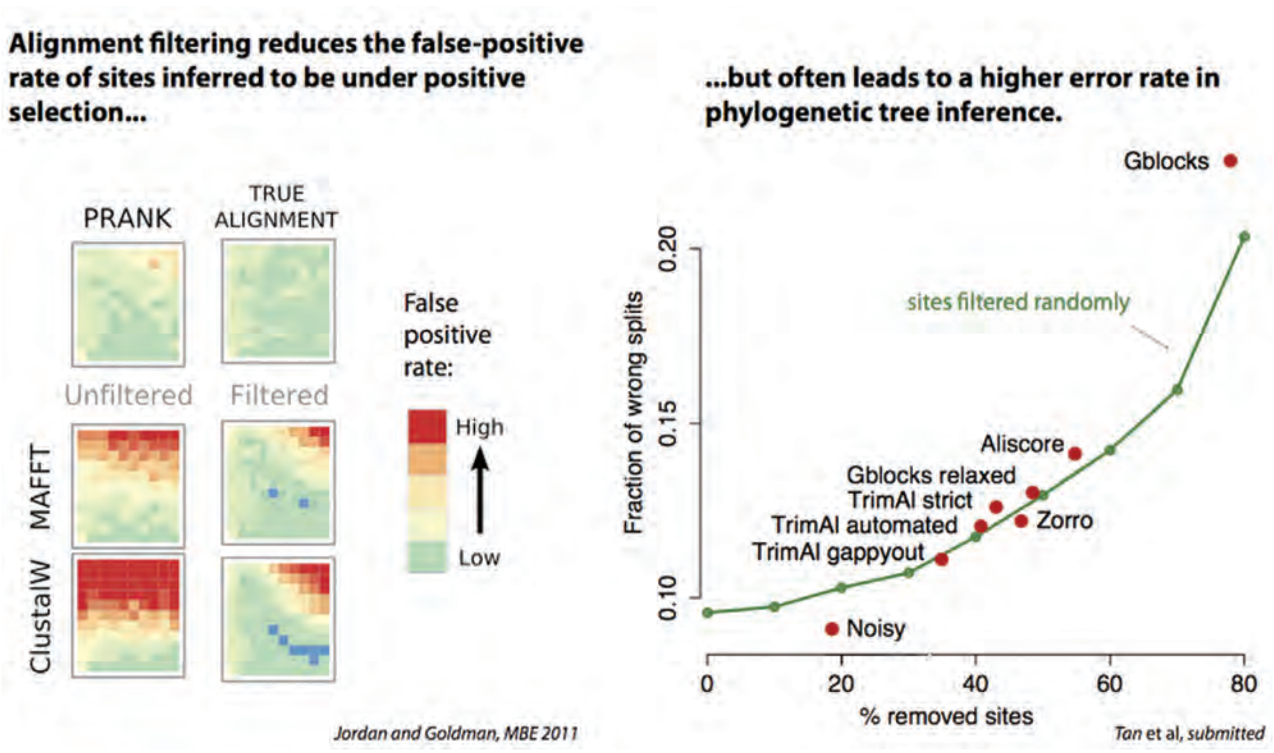
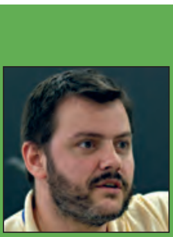


Figure. Alignment filtering reduces the false-positive rate of sites inferred to be under positive selection (left) but often leads to a higher error rate in phylogenetic tree inference (right).

Le Novère Group: Computational systems neurobiology



Nicolas Le Novère

PhD Pasteur Institute, Paris, 1998. Postdoctoral research at the University of Cambridge, UK, 1999-2001. Research fellow, CNRS, Paris, 2001-2003. At EMBL-EBI since 2003.

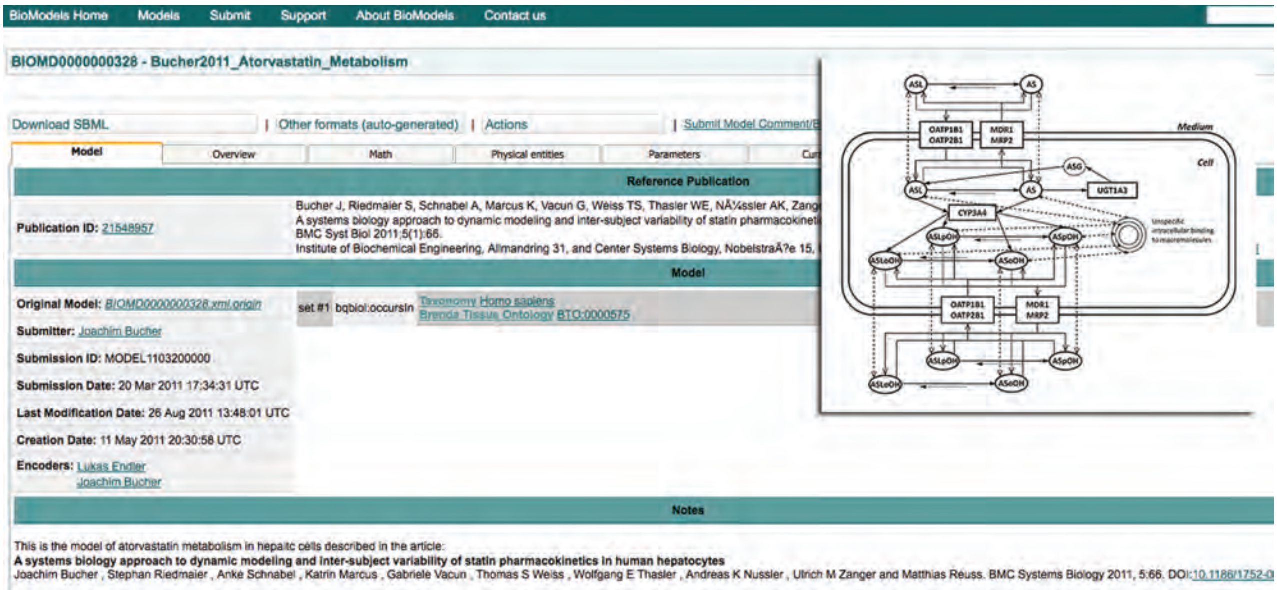
DESCRIPTION OF RESEARCH

Our research interests revolve around signal transduction in neurons, ranging from the molecular structure of proteins involved in neurotransmission to signalling pathways and electrophysiology. In particular, we focus on the molecular and cellular basis of neuroadaptation. By building detailed and realistic computational models, we try to understand how neurotransmitter-receptor movement, clustering and activity influence synaptic signalling. Downstream from the transduction machinery, we build quantitative models of the integration of signalling pathways known to mediate the effects of neurotransmitters, neuromodulators and drugs of abuse. We are particularly interested in understanding the processes of cooperativity, pathway switch and bi-stability.

The group provides community services that facilitate research in computational systems biology. For example, we lead the development of standard representations, encoding and annotating schemes, tools and resources for kinetic models in chemistry and cellular biology. The Systems Biology Markup Language (SBML) is designed to facilitate the exchange of biological models between different software. The Systems Biology Graphical Notation (SBGN) is an effort to develop a common visual notation for biochemists and modellers. We also develop standards for model curation (e.g. MIRIAM) and controlled vocabularies (e.g. SBO, the Systems Biology Ontology) to improve model semantics. In order to manage perennial cross-references, we are manage MIRIAM Registry and its associated Identifiers.org Uniform Resource Identifier (URI) scheme. BioModels Database is the reference resource where scientists can store, search and retrieve published mathematical models of biological interest, launch online simulations or generate sub-models.

SUMMARY OF PROGRESS

- Through progress on the modelling of signalling pathways involved in synaptic plasticity, gained a deeper understanding of the complex equilibria and kinetic events involved in calcium signalling;
- Increased the number of models provided by BioModels Database by >20%, with more than 760 models publicly distributed in 2011;
- Developed the Identifiers.org resolver, which allows MIRIAM URIs to be used in the semantic web.



BIOMD0000000328 - Bucher2011_Atorvastatin_Metabolism

Reference Publication

Publication ID: 21548957

Bucher J, Riedmaier S, Schnabel A, Marcus K, Vacun G, Weiss TS, Thasler WE, NÃ¼ssler AK, Zang
A systems biology approach to dynamic modeling and inter-subject variability of statin pharmacokineti
BMC Syst Biol 2011;5(1):66.
Institute of Biochemical Engineering, Allmandring 31, and Center Systems Biology, NobelstraÃ?e 15,

Model

Original Model: *BIOMD0000000328.xml.origin*

set #1 bqbiol:occursIn Taxonomy Homo sapiens
Brenda Tissue Ontology BTO:0000575

Submitter: Joachim Bucher

Submission ID: MODEL1103200000

Submission Date: 20 Mar 2011 17:34:31 UTC

Last Modification Date: 26 Aug 2011 13:48:01 UTC

Creation Date: 11 May 2011 20:30:58 UTC

Encoders: Lukas Endler
Joachim Bucher

Notes

This is the model of atorvastatin metabolism in hepaitc cells described in the article:
A systems biology approach to dynamic modeling and inter-subject variability of statin pharmacokinetics in human hepatocytes
Joachim Bucher , Stephan Riedmaier , Anke Schnabel , Katrin Marcus , Gabriele Vacun , Thomas S Weiss , Wolfgang E Thasler , Andreas K Nussler , Ulrich M Zanger and Matthias Reuss. BMC Systems Biology 2011, 5:66. DOI:10.1186/1752-0

Figure 1. The BioModels Database: main page of a model of Atorvastatin metabolism, together with a schema featured in the ‘Model of the Month’ page.

MAJOR ACHIEVEMENTS

In 2011 we focused on modelling signalling pathways that are involved in synaptic plasticity. Progress in this area led to a deeper understanding of the complex equilibria and kinetic events involved in calcium signalling. In particular, we shed light on the mechanisms by witch calcium, calmodulin and calcium/calmodulin kinase II interact.

We continued to develop BioModels Database, which grew by 20% during the reporting period. This resource now offers to the community more than 800 models representing 150 000 mathematical relations. Those computational models of biological processes can be used in a variety of formats, as such or to build other models.

A major achievement in 2011 was the development of the Identifiers.org resolver, which provides shared stable identifiers (e.g. for database entries, terms in ontologies) that can be resolved to multiple resources. The MIRIAM system of global identifiers has allowed people to give unambiguous, perennial identifiers to individual data records since 2005. It is a free system that ensures links to data and related information are up to date, which in turn makes it possible to access and reuse the results of experiments in novel ways. By adopting MIRIAM, data providers essentially agree to use one unique reference identifier to describe the same data record. Identifiers.org improves on the MIRIAM system. Annotating data sets with these resolvable URIs eliminates the need to write code to generate links from annotations.

Selected publications

Courtot, M., Juty, N., *et al.* (2011) Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 7, 543.

Dräger, A., Rodriguez, N., *et al.* (2011) JSBML: a flexible Java library for working with SBML. *Bioinformatics* 27 (15), 2167-8.

Kettner C., *et al.* (2010) Meeting Report from the Second “Minimum Information for Biological and Biomedical Investigations” (MIBBI) workshop. *SIGS*, 3: 259-266

Schulz, M., Krause, F., *et al.* (2011) Retrieval, alignment, and clustering of computational models based on semantic annotations. *Mol. Sys. Biol.* 7, 512.

Waltemath, D., Adams, R., et al. (2011) Minimum information about a simulation experiment (MIASE). 7 (4), Art. No.: e1001122.

FUTURE PLANS

The activity of the neurobiology side of the group will expand to cover the signalling pathways involved in synaptic plasticity more comprehensively. While the emphasis will remain on biochemistry, whole-neuron behaviours will be incorporated, in particular electrophysiology.

On the technology side, the software infrastructure running the BioModels Database will be rewritten to cope with new challenges (e.g. size and type of models, authentication and security, easy deployment). Concerning the content, we will extend the support to other types of models (e.g. PK/PD models) and new formats.

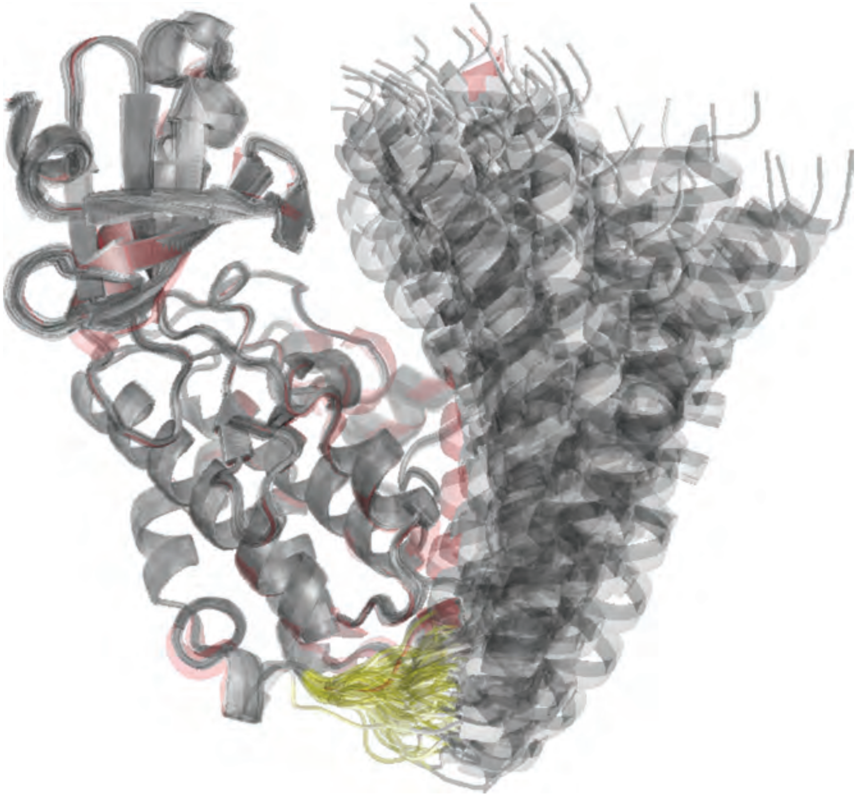
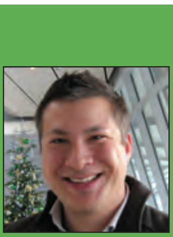


Figure 2. Overlay of 100 model structures created with MODELLER, where structural information was omitted for the four residues linking the kinase domain of CaMKII with the inhibitory helix. These four residues are shown in yellow. The structure corresponding to the published structure of the kinase domain (with the linker region intact, PDB ID: 2BDW, chain A) is shown in red.

Luscombe Group: Genomics and regulatory systems



Nicholas Luscombe

*BA University of Cambridge 1996. PhD University College London 2000. Anna Fuller Postdoctoral Fellow, Yale University 2000-2004.
At EMBL-EBI since 2005.*

DESCRIPTION OF RESEARCH

Cellular life must recognise and respond appropriately to diverse internal and external stimuli. By ensuring the correct expression of specific genes at the appropriate times, the transcriptional regulatory system plays a central role in controlling many biological processes: these range from cell cycle progression and maintenance of intracellular metabolic and physiological balance to cellular differentiation and developmental time-courses. Numerous diseases result from a breakdown in the regulatory system, and one third of human developmental disorders have been attributed to dysfunctional transcription factors. Furthermore, alterations in the activity and regulatory specificity of transcription factors are now established as major sources for species diversity and evolutionary adaptation.

Much of our basic knowledge of transcriptional regulation is derived from molecular biological and genetic investigations. In the past decade, the availability of genome sequences and development of new laboratory techniques have led to the generation of an unprecedented volume of information describing the function and organisation of regulatory systems. Genomic studies now allow us to examine the regulatory system from a whole-organism perspective. However, observations made with these data are often unexpected and appear to complicate our view of gene-expression control.

The flood of biological data raises many interesting questions that require the application of computational methods. Together, bioinformatics and genomics make it possible to uncover general principles and to provide global descriptions of entire systems. Armed with these data, we are in a strong position to achieve answers to interesting biological questions?

The Luscombe Group's research is dedicated to understanding how transcription is regulated, and how this regulatory system is used to control biologically interesting phenomena. We work on two major groups of organisms in parallel: higher eukaryotes and bacteria.

SUMMARY OF PROGRESS

- Genome-wide analysis of the repertoire, usage and cross-species conservation of transcription factors in the human genome;
- Identification and characterisation of nucleoporins as major, genome-wide regulators of transcription in higher eukaryotes;
- Examination of how bacterial cellular systems are controlled through the combination of transcription factors, histidine kinases and small molecules;
- Development and application of iCLIP techniques to determine protein–RNA interactions on a transcriptome-wide scale.

MAJOR ACHIEVEMENTS

Higher eukaryotes

One cannot understand transcriptional control without knowing the identity of the regulators. Therefore, we performed a high-quality analysis of the transcription factor repertoire in the human genome (Vaquerizas *et al.*, 2009). This work is now the main reference for mammalian transcription factor repertoires and we have attracted collaborators who wish to identify their binding specificities on a large scale (Jolma *et al.*, 2010).

New mechanisms for transcriptional control

Transcriptional control operates on many levels in eukaryotes. In collaboration with Dr Asifa Akhtar (Max-Planck-Institute for Immunobiology, Freiburg), we use the dosage-compensation system in flies as a model for chromosome-wide regulation. We characterised the first histone-modification enzyme displaying context-dependent substrate specificities (Kind *et al.*, 2008; Raja *et al.*, 2010). Most recently, we also discovered that components of the nuclear pore complex control the expression for approximately 25% of fly genes by shaping the 3D organisation of the genome (Vaquerizas *et al.*, 2010). In a theoretical project, we also used this knowledge to build a model explaining RNA-polymerase II behaviour, given the locations of transcription factors and nucleosomes in promoters (Zaugg and Luscombe, 2011).

Organism-wide regulation in bacteria

Bacteria are attractive systems for organism-wide analysis, as they are extremely sensitive to changes in external conditions. Most genomic studies of bacteria so far have focused only on transcription but this provides only a partial view of an organism's regulatory apparatus. We described additional levels of control in *Escherichia coli* including kinases and metabolites (Seshasayee *et al.*, 2010; Seshasayee and Luscombe, submitted); using these data, we then assessed how the regulatory network intersects with the metabolic system (Seshasayee *et al.*, 2008; 2011).

Experimental work

We have established a wet-lab component to our research, kindly hosted by Professor Gordon Dougan at the Wellcome Trust Sanger Institute. We have just published the first ChIP-Seq studies in a prokaryote (Kharamanoglou *et al.*, 2011; Prieto *et al.*, 2011) and are in the process of expanding this to a comparative study of transcriptional regulation in multiple Salmonella strains.

Beyond transcriptional regulation

In collaboration with Dr Jernej Ule (MRC Laboratory of Molecular Biology, Cambridge, UK), we developed a technique to assess protein–RNA interactions at single-nucleotide resolution called iCLIP (Konig *et al.*, 2010; Wang *et al.*, 2010). The technique was applied to study the role of heterogeneous nuclear ribonucleoprotein subunit C (hnRNP C) in regulating splicing. We showed that hnRNP C binds to both introns and exons in order to promote or repress splicing at specific sites on transcripts.

Selected publications

Kahramanoglou, C., Seshasayee, A.S., *et al.* (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res.* 39 (6), 2073-91.

Prieto, A.I., Kahramanoglou, C., *et al.* (2011) Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res.* Article in press.

Zaugg, J.B. and Luscombe, N.M. (2012) A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast. *Genome Res.* 22 (1), 87-94.

FUTURE PLANS

We will continue our analysis of genome-scale data to understand how transcription is regulated, and how it is used to control interesting systems. A major focus continues to be our close interactions with research groups performing functional genomic experiments.

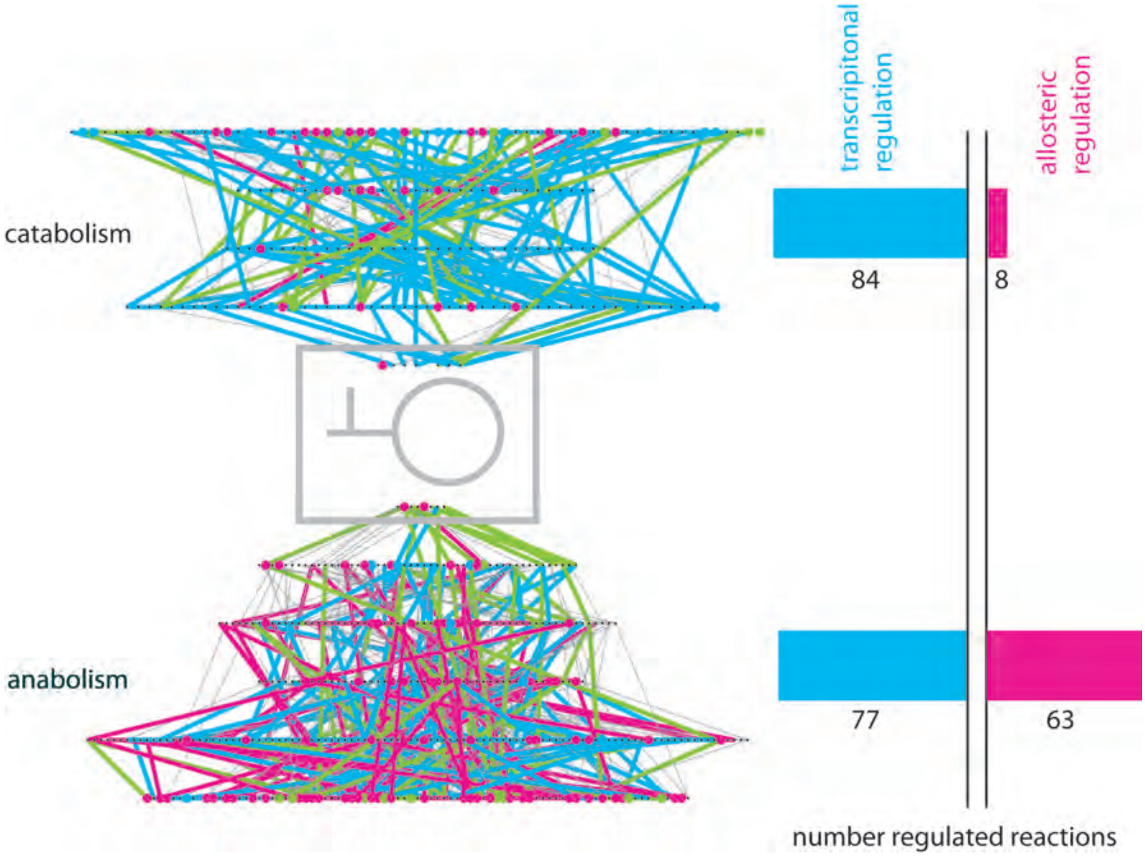


Figure. A network representation displays the *E. coli* metabolic system. Nodes represent small molecules and edges depict enzymatic reactions. The reactions are coloured according to whether they are controlled transcriptionally (blue), allosterically (cyan) or by both methods (green). Allosteric feedback predominantly regulates anabolic pathways, whereas transcriptional feedback controls both anabolic and catabolic pathways.

Marioni Group: Computational and evolutionary genomics

John Marioni

PhD in Applied Mathematics, University of Cambridge, 2008. Postdoctoral research in the Department of Human Genetics, University of Chicago. At EMBL since September 2010.

DESCRIPTION OF RESEARCH

Over the past 50 years, numerous studies have emphasised the critical role of gene expression levels in evolution, developmental processes and disease progression. Variability in the transcriptional landscape can help explain phenotypic differences both between and within species; for example, differential expression of the *Tan* gene between North American Drosophila species underlies divergence of pigmentation, while variation in expression levels of the *MMP3* gene within the human population alters both vascular tissue remodelling and risk of developing atherosclerosis. As a result, identifying and characterising the regulatory mechanisms responsible for changes in gene expression is critically important. Recently, the advent of next-generation sequencing technology has revolutionised our ability to do this. By facilitating the generation of unbiased, high-resolution maps of genomes, transcriptomes and regulatory features such as transcription factor binding sites, these new experimental techniques have given rise to a detailed view of gene expression regulation in both model and, importantly, non-model organisms.

To make the most of these technological developments, it is essential to develop effective statistical and computational methods for analysing the vast amounts of data generated. Only by harnessing experimental and computational biology will we be able to truly understand complex biological processes such as gene regulation. With this in mind, my group focuses on the development of computational methods for interrogating high-throughput genomics data. Our work focuses primarily on modelling variation in gene expression levels in different contexts: between individual cells from the same tissue; across different samples taken from the same tumour; and at the population level where a single, large sample of cells is taken from the organism and tissue of interest. We apply these methods to a range of biological questions from studying the regulation of gene expression levels in a mammalian system to the development of the brain in a marine annelid. In all of these projects we collaborate with outstanding experimental groups, both within and outside EMBL. Together, we frame biological questions of interest, design studies and analyse and interpret the data generated.

SUMMARY OF PROGRESS

- Combined RNA-sequencing with DNA sequence information to interrogate gene and isoform regulatory mechanisms in a mammalian system;
- Started to develop approaches (both experimentally and computationally) for assessing the quality of single-cell RNA-sequencing data;
- Began to derive models for accurately capturing the variability present in single-cell RNA-sequencing data;
- Developed tools for combining spatial and quantitative gene expression data in the context of tumour progression and in early embryonic development.

MAJOR ACHIEVEMENTS

Gene regulation in mammals

Understanding the regulatory mechanisms that underlie gene expression levels is essential to understanding how mammals have evolved. To this end, we collaborated with the laboratories of Duncan Odom at the Cancer Research UK-Cambridge Research Institute and those of Alvis Brazma and Paul Flicek at EMBL-EBI to collect and develop models for analysing RNA-sequencing data obtained from liver samples derived from both the parents and F1 crosses of two different strains of mice. This system allows us to categorise genes, in an unbiased fashion, into different sets depending upon the mechanism by which they are regulated. We can extend this analysis to study variation in the regulation of isoform usage over a short evolutionary distance.

Modelling single-cell RNA-sequencing data

Studying gene expression levels at the single-cell level is crucial in many biological contexts. Some examples include: early developmental stages in which there are few cells, each with often very distinctive functions; in the brain, which is known to contain a highly heterogeneous population of cells; and in tumour samples, which might harbour multiple different sub-types. Although developments in next-generation sequencing technology have made the generation of such data practical, a systematic assessment of the quality and limitations of these approaches has not yet been performed. To address this issue, we collaborated with groups in the EMBL Developmental Biology Unit to design dilution experiments to study both the technical and biological variability present in single-cell RNA-sequencing data. In parallel, we developed computational methods for quantifying the variability present in gene-expression profiles obtained from multiple cells taken from the same tissue.

Evolution of model systems

We are working with Detlev Arendt's group at EMBL Heidelberg to generate the first comprehensive catalogue of expression levels within the brain of a Bilaterian (the marine annelid *Platynereis dumerilii*). A postdoc jointly shared between the two collaborating groups is performing the experimental work. Moreover, we are deriving methods for combining the single-cell RNA-seq data with a previously generated low-throughput spatial map of expression within the Platynereis brain – this will allow each sequenced cell to be associated with a spatial location. To take advantage of this unique system we are developing probabilistic models that can segment the brain into biologically distinct regions, taking into account both the spatial and the quantitative aspects of the data being collected.

Selected publications

Perry, G.H., *et al.* (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* Published online 29 December; doi: 10.1101/gr.130468.111.

Perry G.H., Marioni, J.C, *et al.* (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Mol. Ecol.* 19, 5332-44.

Marioni *et al.*, (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays *Genome Res.* 18:1509-17.

FUTURE PLANS

In 2012 we will continue to develop the methods outlined above, and will work with our experimental collaborators to apply them to relevant and important biological questions. From a computational perspective, modelling single-cell transcriptomics data will increase in importance. Methods for storing, visualising, interpreting and analysing the data generated will be critical if we are to exploit these data to the fullest extent. We will also work on methods for analysing conventional next-generation sequencing data, building on work that we have performed previously.

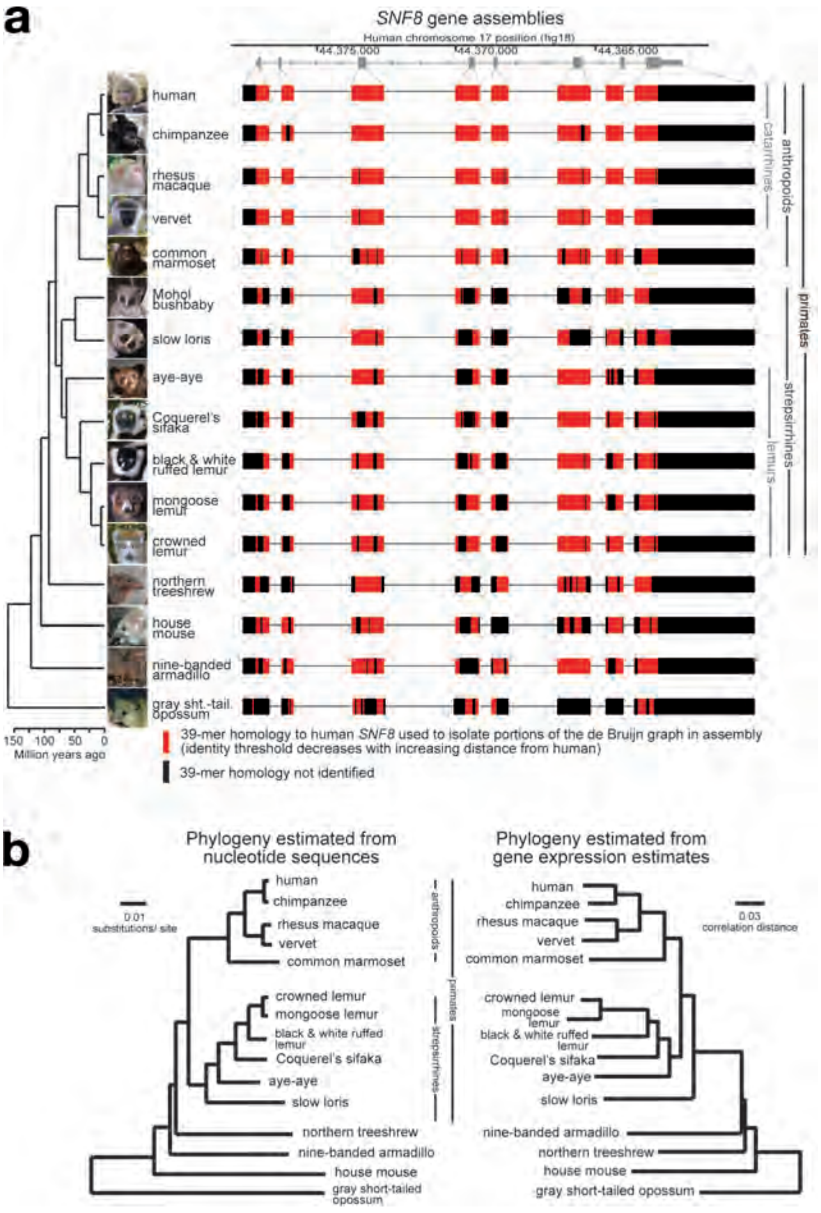


Figure. Transcript assembly and phylogenetic reconstruction from RNA-seq data generated from liver samples taken from 12 primates and 4 non-primate outgroups. (a) A typical example of an assembled gene, SNF8, with complete cross-species exon conservation. Red bars depict identified homologies to the human SNF8 RefSeq coding sequence that were used to isolate the appropriate region of the de Bruijn graph during the assembly process. Divergence times are approximate and based on consensus estimates from previous studies. (b) Neighbour-joining trees estimated from nucleotide sequence and gene expression data. Nucleotide sequence distance matrix was computed from concatenated multi-species alignments of coding sequences of 515 genes that were assembled for all 16 species. Gene expression pairwise correlation distance matrix was computed for species mean expression estimates using all genes assembled in at least six species (6,494 genes). As expected, the known primate phylogeny was recapitulated perfectly from the nucleotide sequence data with the only discrepancy among non-primate mammals being the juxtaposition of the mouse and armadillo branches, likely explained by long branch attraction that is a common issue in phylogenetic analyses that include rodents. Variation in the expression data also follows a phylogenetic pattern, but with slow loris erroneously placed outside all other primates, and the misplacement of armadillo. [Taken from Perry *et al.*, Genome Research 2012.]

Rebholz-Schuhmann Group: Phenotype ontologies and disease

Dietrich Rebholz-Schuhmann

PhD in immunology, University of Düsseldorf, 1989. Senior Scientist at GSF, Munich, 1995. Director of Healthcare IT, LION Bioscience AG, Heidelberg, 1998. At EMBL-EBI since 2003.

DESCRIPTION OF RESEARCH

Text mining comprises the fast retrieval of relevant documents from the whole body of the scientific literature and the extraction of facts from these texts. Text-mining solutions are becoming mature enough to be automatically integrated into workflows for research and into services for the general public, for example delivery of annotated full text documents as part of UK PubMed Central (UKPMC).

Research in the Rebholz Group focuses on extracting facts from the literature. Our goal is to connect literature content automatically to other biomedical data resources and to evaluate the results. Our research targets the recognition of biomedical terms (genes, proteins, gene ontology labels) and the identification of relationships between them. Our work is split into three tightly coupled parts: named entity recognition and its quality control (e.g. UKPMC project); knowledge discovery (e.g. identification of gene–disease associations); and further development of the IT infrastructure for information extraction and fact delivery.

SUMMARY OF PROGRESS

- Further developed different solutions to normalise the representation of concepts in the literature: LexEBI, IeXML, Whatizit and CALBC (collaborative annotation of a large-scale corpus). All solutions contribute to the annotation and indexing of the full-text scientific literature as part of UK PubMed Central (see page X) and the SESL (Semantic Enrichment of the Scientific Literature) project;
- Developed and evaluated new solutions based on public phenotype resources to achieve cross-validation of candidate genes between Human and Mouse;
- Developed and evaluated a new machine-learning based solution for the characterisation of scientific statements to qualify information extraction results according to the quality of the author's statements;
- As part of the SESL project, processed full-text documents from major publishers to integrate the extracted evidences with bioinformatics data resources (e.g. UniProt, ArrayExpress) and to deliver the assertions from the SPARQL endpoint to the project partners in the pharmaceutical industry.

MAJOR ACHIEVEMENTS

Named Entity Recognition. Standardisation of the scientific literature: UKPMC, LexEBI, and CALBC. *Adam Bernard, Senay Kafkas, Jee-Hyub Kim, Vivian Lee, Ian Lewin, Chen Li, Maria Liakata, Shyamasri Saha, Ying Yan*

Our research focuses on identifying named entities (e.g. genes, proteins, diseases) in the literature and linking them to entries in a reference database. We have provided several solutions: LexEBI (a terminological resource), IeXML (an annotation framework for documents), Whatizit (an information-extraction infrastructure) and CALBC (an evaluation infrastructure). We generate LexEBI in order to provide full coverage of domain knowledge in molecular biology for gene and protein names, chemical entities, diseases, species and ontological terms. In 2011 several bioinformatics resources (e.g. the BioThesaurus) were incorporated into LexEBI, which interlinks terms across all resources according to their similarity. LexEBI is integrated with IeXML and the EBI's information-extraction infrastructure for indexing the full body of scientific literature (UK PubMed Central).

By harmonising annotations from automatic text mining solutions, CALBC produced SSC-II: a large-scale, annotated biomedical corpus with four semantic groups (chemical entities and drugs; genes and proteins; diseases and disorders; species). SSC-II was used for the Second CALBC Challenge, wherein participants were asked to annotate the corpus with their annotation solutions (Rebholz-Schuhmann *et al.*, 2010). The final version SSC-III includes a smaller version (174 999 Medline abstracts with 2 548 900 annotations) and a larger one (714 283 Medline abstracts with 10 304 172 annotations). In 2011 we used our text-mining solution (see Liakata and Saha, below) to annotate the open-access subset of UK PubMed Central; we invite the community to collaborate in improving these solutions in the interests of expediting knowledge discovery.

Knowledge discovery and novel text-mining solutions. *Maria Liakata, Shyamasree Saha*

Text mining is largely about retrieving relevant documents quickly from the whole body of the scientific literature and extracting useful facts. Scientists use a well-established discourse structure to: relate their work to the state of the art; express their motivation and hypotheses; and report on their methods, results and conclusions. We defined 11 categories, or 'Core Scientific Concepts' (CoreSCs), at the sentence level: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. The automatic recognition of these CoreSCs can greatly facilitate biomedical information extraction by characterising the different types of facts and evidences available in a scientific publication. In 2011 we trained machine-learning classifiers (SVM and CRF) on a corpus of 265 full-text articles to recognise CoreSCs automatically; we then compared the two methods and evaluated our automatic classifications against a manually annotated gold standard. We achieved promising accuracies with 'Experiment', 'Background' and 'Model' categories (F1 scores of 76%, 62% and 53%, respectively). The most discriminative features were local sentence features (e.g. unigrams, bigrams, grammatical dependencies) while features encoding the document structure (e.g. section headings) also played an important role.

Development and use of phenotype ontological resources.

Robert Hoehndorf, Anika Oellrich

The use of bioinformatics data in clinical environments requires the consistent and complete representation of phenotypes. Research into the molecular origins of hereditary human diseases benefits from computational methods for prioritising gene candidates, especially for orphan diseases where little evidence is available. High-throughput phenotype studies systematically assess the phenotypic outcome of targeted mutations in model organisms, and the comparisons of the experimentally determined phenotypes from model organisms should contribute to disease gene prioritization in humans. We developed a special method for gene prioritisation based on comparing phenotypes of mouse models with those of human diseases. The method was evaluated on known gene–disease associations for human and for mouse; the results demonstrate better prediction performance in comparison to previous phenotype-based approaches

IT infrastructure development for information extraction. The SESL Triple Store: retrieval over large literature content.

Samuel Croset, Christoph Grabmüller, Silvestras Kavaliauskas, Chen Li

Diabetes mellitus Type II (DmT2) is a multifactorial disease and the genetic causes are the subject of ongoing research. The integration of all data relevant to this multifactorial disease (e.g. metabolic pathways and signalling results) is complex due to the number of data resources as well as the volume of information and facts from the scientific literature. Applying state-of-the-art text processing techniques and Semantic Web technology, content from scientific publications can be integrated automatically into the biomedical research data infrastructure, making use of 'Linked Data' principles (see Berners-Lee, 2001) and structuring facts and data according to established standards. The automatic integration of biomedical reference data repositories (e.g. OMIM, UniProtKB, Gene Expression Atlas) with literature content from scientific publishers forms the core to the SESL prototype. The combined evidence facilitates seamless exploration of gene–disease associations for DmT2 across all resources.

The SESL Triple Store gives access to 2232 unique gene–disease associations (GDAs) from different resources and allows cross-validation of functional annotations between data repositories and the scientific literature. The literature-based evidence forms approximately 40% (~14.7 million triples) of the Triple Store for all literature content and 15.4% (~4 million triples) for the public SESL prototype. UniProtKb and UMLS form the other large portions of the repository. Querying across the repositories gives a selection of genes relevant to DmT2 (e.g. TCF7L2, HNF-4-alph) as well as other associated genes (e.g. PTP1B, PLANH1) and related diseases (e.g. metabolic syndrome).

Reference cited: Berners-Lee, T., Hendler, J. and Lassila, O. "The Semantic Web." *Sci. Am.*, 17 May 2001.

Selected publications

Hoehndorf, R., Dumontier, M., *et al.* (2011) Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS One* 6 (7), e22006.

Rebholz-Schuhmann, D., Yepes, A.J., *et al.* (2011) Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J. Biomed. Semantics* 2 (Suppl 5), S11.

Thompson, P., McNaught, J., *et al.* (2011) The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 12, 397.

The standardisation of document representation and the full integration of the document repository with standardised extraction methods is helping us exploit literature content in the context of a shared biomedical research infrastructure. The literature can be integrated through public and private repositories of derived content, or through automatic processing of the literature content with public text-processing services. In 2011 we presented solutions that serve as a standard for such an integrated infrastructure; this approach will greatly facilitate the brokering of facts and content.

FUTURE PLANS

Our ongoing research is concerned with the extraction of information that describes the bioactivities of chemical entities, that cross-validates molecular pathways against the protein interaction networks from the scientific literature and that gathers gene-disease associations from different literature-based repositories. The research continues at a reduced scale, mainly supported by the PhD students and the research fellows.

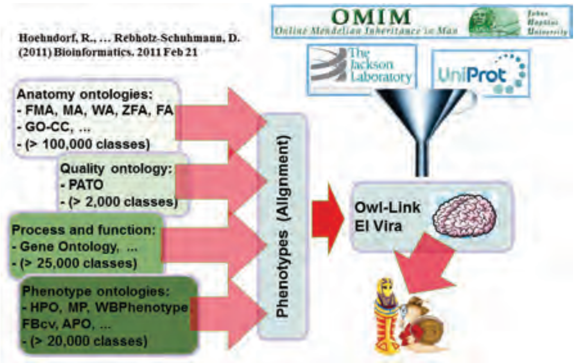


Figure. Ontological resources can be exploited to represent phenotypes at different levels of specification and generalisation. In addition, they can be combined or aligned through upper level ontologies to enable inference across the different feature representations (see Owl-Link, El Vira). Once this has been achieved, the integrated ontologies can be used for knowledge discovery, in particular for the identification of disease candidate genes. Alternative methods use phenotypic profiling of genes and diseases after lexical matching of mouse phenotypes to human phenotypes.

Saez-Rodriguez Group: Systems biomedicine

Julio Saez-Rodriguez

PhD University of Magdeburg, 2007. Postdoctoral work at Harvard Medical School and MIT. At EMBL-EBI since 2010. Joint appointment, Genome Biology Unit.

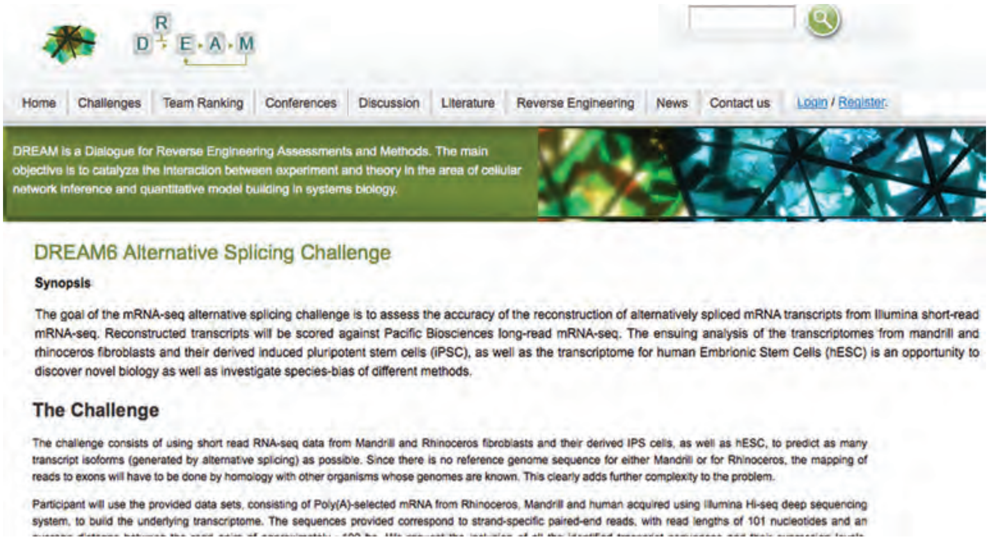
DESCRIPTION OF RESEARCH

Our group aims to achieve a functional understanding of signalling networks and their deregulation in disease, and to apply this knowledge to novel therapeutics. Human cells are equipped with complex signalling networks that allow them to receive and process the information encoded in myriad extracellular stimuli. Understanding how these networks function is a rich scientific challenge but also has practical applications, as alterations in the functioning of these 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 signaling 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 CellNOpt, an R and MatLab platform to model signaling networks using logic formalisms of different quantitative and time resolution;
- Co-organised (including the set up and hosting of the website) the sixth edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods);
- Established methods to analyse large drug screenings in cancer cell lines.



DREAM is a Dialogue for Reverse Engineering Assessments and Methods. The main objective is to catalyze the interaction between experiment and theory in the area of cellular network inference and quantitative model building in systems biology.

DREAM6 Alternative Splicing Challenge

Synopsis

The goal of the mRNA-seq alternative splicing challenge is to assess the accuracy of the reconstruction of alternatively spliced mRNA transcripts from Illumina short-read mRNA-seq. Reconstructed transcripts will be scored against Pacific Biosciences long-read mRNA-seq. The ensuing analysis of the transcriptomes from mandrill and rhinoceros fibroblasts and their derived induced pluripotent stem cells (iPSC), as well as the transcriptome for human Embrionic Stem Cells (hESC) is an opportunity to discover novel biology as well as investigate species-bias of different methods.

The Challenge

The challenge consists of using short read RNA-seq data from Mandrill and Rhinoceros fibroblasts and their derived IPS cells, as well as hESC, to predict as many transcript isoforms (generated by alternative splicing) as possible. Since there is no reference genome sequence for either Mandrill or for Rhinoceros, the mapping of reads to exons will have to be done by homology with other organisms whose genomes are known. This clearly adds further complexity to the problem.

Participant will use the provided data sets, consisting of Poly(A)-selected mRNA from Rhinoceros, Mandrill and human acquired using Illumina Hi-seq deep sequencing system, to build the underlying transcriptome. The sequences provided correspond to strand-specific paired-end reads, with read lengths of 101 nucleotides and an

Figure 1. The DREAM Challenge website. The challenges address fundamental questions in systems biology, bridging theory and experimentation. The questions aren't simple: : for example, researchers test a variety of algorithms to deduce the structure of a biological network based on experimental data. DREAM allows researchers to compare the strengths and weaknesses of these methods and provides a sense of how reliable a given model may be.

MAJOR ACHIEVEMENTS

In 2011 we developed CellNOpt, an R and MatLab platform for modelling signalling networks using logic formalisms of different quantitative and time resolution. CellNOpt uses high-throughput biochemical data to generate models spanning simple Boolean logic models that coarsely describe signalling networks and continuous fuzzy-logic models, as well as differential equation systems describing details in the dynamics of the underlying biochemical processes.

Our group co-organised the sixth edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods), a community effort organised around challenges to advance the inference of mathematical models of cellular networks. This year we run four challenges: (i) Inference of the kinetic parameters of three gene regulatory networks; (ii) reconstruction of the alternatively spliced mRNA transcripts from short-read mRNA-seq data; (iii) prediction of gene expression levels from promoter sequences in eukaryotes; and (iv) diagnosis of Acute Myeloid Leukaemia from patient samples using flow cytometry data.

In the context of the Sanger Institute/EMBL-EBI (ESPOD) postdoctoral programme, we established various methods to analyse large drug screenings in cancer cell lines, particularly for the prediction of drug response from genetic features.

We also refined methods to infer new effects of already approved compounds from gene expression data, leading to new drug repurposing opportunities.

FUTURE PLANS

We will continue to develop methods and tools to understand signal transduction in human cells, as well as their potential to yield insights of medical relevance. Our main focus will be on modelling signalling networks using phospho-proteomics data with our tool CellNOpt, and finding ways to employ different proteomics technologies and sources of information about pathways. We will also continue to develop methods to infer 'drug mode of action' and 'drug repurposing' by integrating genomic and transcriptomic data with drug screenings. Using these methods we hope to address questions such as:

- What are the origins of the profound differences in signal transduction between healthy and diseased cells and in particular, in the context of cancer, between normal and transformed cells?
- What are the differences in signal transduction among cancer types? Can we use these differences to predict disease progression?
- Do these differences reveal valuable targets for drug development? Can we study the side effects of drugs using these models?

Selected publication

C. Terfve, J. and Saez-Rodriguez (2012) Modeling signaling networks using high-throughput phospho- proteomics. Advances in Systems Biology, *Adv. Exp. Med. Biol.* 736:19-57, 2012.

Prill, R.J., Saez-Rodriguez, J., *et al.* (2011) Crowdsourcing network inference: The DREAM predictive signaling network challenge. *Sci. Signal.* 4 (189), mr7.

Saez-Rodriguez, J., Alexopoulos, L.G. and Stolovitzky, G. (2011) Setting the standards for signal transduction research. *Sci. Signal.* 4 (160), pe10.

Saez-Rodriguez, J., Alexopoulos, L.G., et al. (2011) Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.* 71 (16), 5400-11.

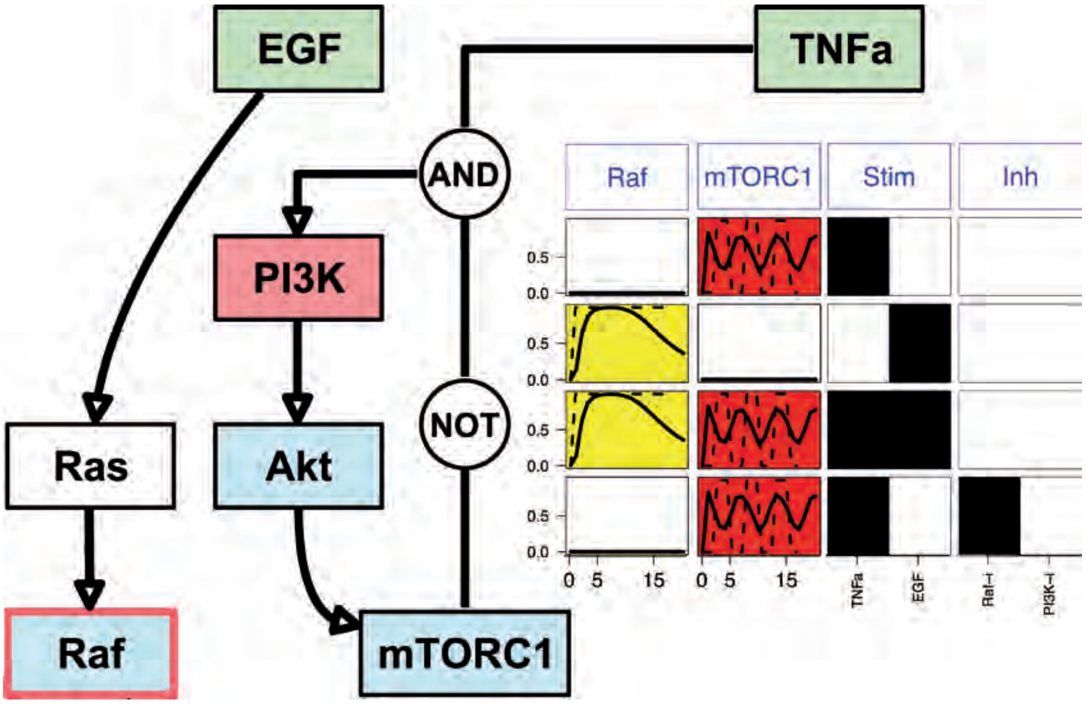


Figure 2. An illustration of how we use our logic modeling method CellNOpt to better understand deregulation of signal transduction in disease. Left: simple pathway model; right: experimental data and match between model simulations and data.

Thornton Group: Computational biology of proteins: structure, function and evolution

Janet Thornton

PhD King's College and National Institute for Medical Research, London, UK, 1973. Postdoc at the University of Oxford, NIMR and Birkbeck College. Lecturer, Birkbeck College, 1983-1989. Prof. of Biomolecular Structure, University College London (UCL) since 1990. Bernal Prof. at Birkbeck College, 1996-2002. Director of the Centre for Structural Biology, Birkbeck College and UCL, 1998-2001. Director of EMBL-EBI since 2001.

DESCRIPTION OF RESEARCH

The goal of our research is to understand more about how biology works at the molecular level, with a particular focus on proteins and their 3D structure and evolution. We are exploring how enzymes perform catalysis, which involves gathering relevant data from the literature and developing novel software tools to characterise enzyme mechanisms and navigate through catalytic and substrate space. In parallel we are investigating the evolution of these enzymes to discover how one enzyme can evolve new mechanisms and new specificities. This involves the integration of 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. The group is also interested in understanding the molecular basis of ageing in different organisms, through a strong collaboration with experimental biologists at UCL. Our role is to analyse functional genomics data from flies, worms or mice and relate these observations to effects on life span by combining information on function, context (pathways and interactions) and evolutionary relationships.

SUMMARY OF PROGRESS

- Continued to analyse enzyme mechanisms, with a recent focus on lyases;
- Developed sophisticated, novel algorithms to compare enzyme reactions and to navigate through reaction (EC) space. These tools have been used to re-investigate enzyme classification;
- Completed an analysis of the evolution of 276 enzyme superfamilies to explore the range of reactions they catalyse and how they evolve new functions, based on a phylogenetic analysis involving sequences and structures;
- Analysed human mutation data from the 1000 Genomes Project, developing a pipeline to map new mutations on protein sequences and using their structures to try to understand or predict their effects on function;
- In our analysis of the functional genomics ageing, focused on the analysis of survival curves and pathway analysis using expression data. We have made great progress in developing an approach to capture the information about the Insulin signaling pathway and to work out mechanisms about how this pathway is activated or suppressed by individual mutations, which affect longevity.

MAJOR ACHIEVEMENTS

In 2011 we completed a review of enzymes and the catalytic reactions they perform. The result was a deeper understanding of these complex proteins, from the chemical versatility of the catalytic toolkit – including the use of cofactors (both metal ions and organic molecules) – to the complex mapping of reactions to proteins (which is rarely one-to-one). This work also shed light on the structural complexity of enzymes and their active sites, often involving multidomain or multisubunit assemblies. Our findings underscore how the enzymes we see today reflect millions of years of evolution, involving de novo design followed by exquisite regulation and modulation to create optimal fitness for life (Holliday *et al.*, 2011). These studies were based on the MACiE database of Enzyme mechanisms (Holliday *et al.*, 2011), which has grown to include over 350 enzyme mechanisms and many additional features and search tools.

We developed FunTree, a new resource that brings together sequence, structure, phylogenetic, chemical and mechanistic information for structurally defined enzyme superfamilies. Gathering this range of data into a single resource allows the investigation of how novel enzyme functions have evolved within a structurally defined superfamily as well as providing a means to analyse trends across many superfamilies. This is done using trees generated from structurally informed multiple sequence alignments, which draw on both domain structural alignments supplemented with domain sequences and whole sequence alignments based on commonality of multi-domain architectures. These trees are decorated with functional annotations (e.g. metabolite similarity) as well as annotations from manually curated resources such the catalytic site atlas and MACiE for enzyme mechanisms. FunTree is freely available through a web interface (Furnham *et al.*, 2011).

In 2011 we completed a review of the use of bioinformatics in ageing. In this review we provide a description of existing databases and computational tools that are available for ageing researchers. We also describe approaches to data interpretation in the field of ageing including gene expression, comparative and pathway analysis. We review recent biological insights gained from applying bioinformatics methods to analyse and interpret ageing data in different organisms, tissues, conditions or cell types, and suggest future directions to address current challenges in the field (Weiser D. *et al.*, (2011).

FUTURE PLANS

We will continue our work on understanding more about enzymes and their mechanisms using structural and chemical information. This will include a study of how the enzymes, their families and their pathways have evolved and how genetic variations in individuals impact on structure, function and disease. We will seek to gain a better understanding of reaction space and its impact on pathways, which will allow improved chemistry queries across our databases. We will continue to use evolutionary approaches to improve our prediction of protein function from sequence and structure. In the ageing project, we are interested in tissue specificity, analysis of survival curves and combining transcriptome data sets with network analysis for flies, worms and mice, to compare the different pathways and ultimately explore effects related to human variation and age.

Selected publications

Furnham, N., Sillitoe, I., *et al.* (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.* 40 (Database issue), D776-82.

Holliday, G.L., Fischer, J.D., *et al.* (2011) Characterising the complexity of enzymes based on their mechanisms and structures using a bio-computational analysis. *FEBS J.* 278 (20), 3835-45.

Holliday, G.L., Andreini, C., *et al.* (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40 (Database issue), D783-9.

Wieser, D., Papatheodorou, I., *et al.* (2011) Computational biology for ageing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 366 (1561), 51-63.

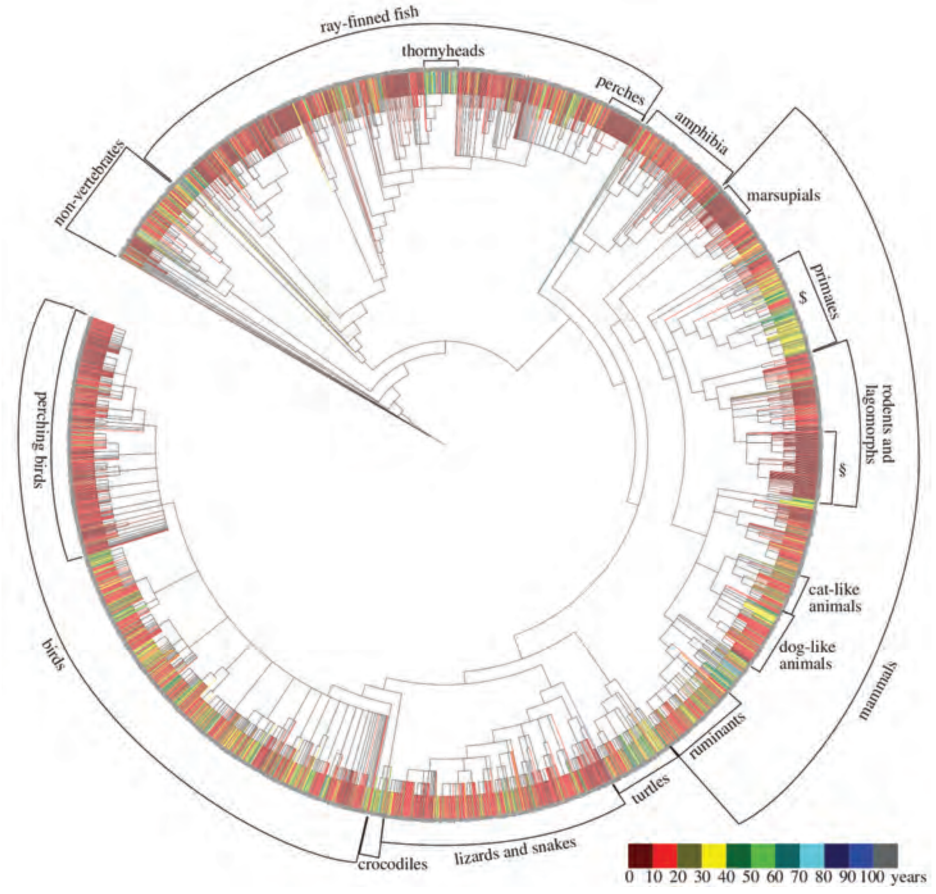


Figure. Bioinformatics analysis of phylogenetic clustering of maximal lifespan records. The figure was created with the help of the iTOLWebservice using longevity data from AnAge and phylogeny data from NCBI Taxonomy. Similar maximal lifespans can be found in phylogenetically distant species. Within the large variety of maximal lifespans in evolutionary clades, clusters of similar maximal lifespan can be observed. This hints to a strong genetic influence on ageing. $ indicates Callitrichinae (a family of New World monkeys), § indicates mouse-like animals.

The EMBL International PhD Programme at EMBL-EBI

EMBL-EBI benefitted from the presence of 39 PhD students in 2011. Of these, eight were formally awarded their PhDs and another two submitted their theses. 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 EMBL faculty, local academics and, where appropriate, industry partners. Our strong links with the University of Cambridge allow our students to obtain their degree from there.

As a global centre of excellence for bioinformatics, EMBL-EBI offers its students a unique opportunity to participate in open, interactive research groups that are defining the state of the art. In addition to their research, our students share their knowledge by organising PhD Student Symposia and an annual Science and Society event, which welcomes members of the public to debate issues in science and technology that have a tangible impact on everyday life.

During 2011 the following students successfully defended their theses and were awarded PhDs from the University of Cambridge: Florence Cavalli, Heidi Dvinge, Julia Fischer, Markus Fritz, Jacky Hess, Michele Mattioni, Dace Ruklisa and Judith Zaugg. Greg Jordan and Diva Tommei submitted their theses and will defend in 2012. The work of three of our graduating students is summarised here.

Julia Fischer obtained her degree from the University of Cambridge for her thesis, 'Characterisation, classification and conformational variability of organic enzyme cofactors' based on her work in Janet Thornton's laboratory. Julia explored the properties, evolutionary context and functional roles of organic enzyme cofactors using computational methods. She developed the CoFactor database using data extracted manually from the scientific literature as well as automatically from web resources. Julia is now a technical consultant for PA Consulting in Melbourn, UK.

Dace Ruklisa received a degree for her thesis, 'Large-scale genomic association studies in fruit fly and human,' based on her work in Ewan Birney's group. In light of the growing genetic marker density for many organisms, and recognising new challenges in modelling phenotypes, Dace investigated new ways to explain phenotype using computational methods. She developed novel methods to construct large composite models for genotype-to-phenotype association that use multiple markers. Her thesis is one of the very few fruit fly studies that exploit dense SNP maps and incorporate multiple traits. Her methods were also applied to the reanalysis of the human case–control data for type 1 diabetes from the Wellcome Trust Case Control Consortium. Dace is now working in the Genetics Institute at University College London.

Judith Zaugg obtained her degree from the University of Cambridge for her thesis, 'A computational study of promoter structure and transcriptional regulation in yeast on a genomic scale,' based on her work in Nick Luscombe's group. Her studies show how promoter structures vary extensively across a genome, and explain how these differences are often associated with specific transcriptional behaviours. Judith proposes that these differences should be taken into account when interpreting genome-wide functional data. Her future work will investigate additional levels of control such as the 3D intra-nuclear organisation of the yeast genome. Judith is now working as a postdoc at EMBL-EBI , and will take a position at Stanford University in 2012.

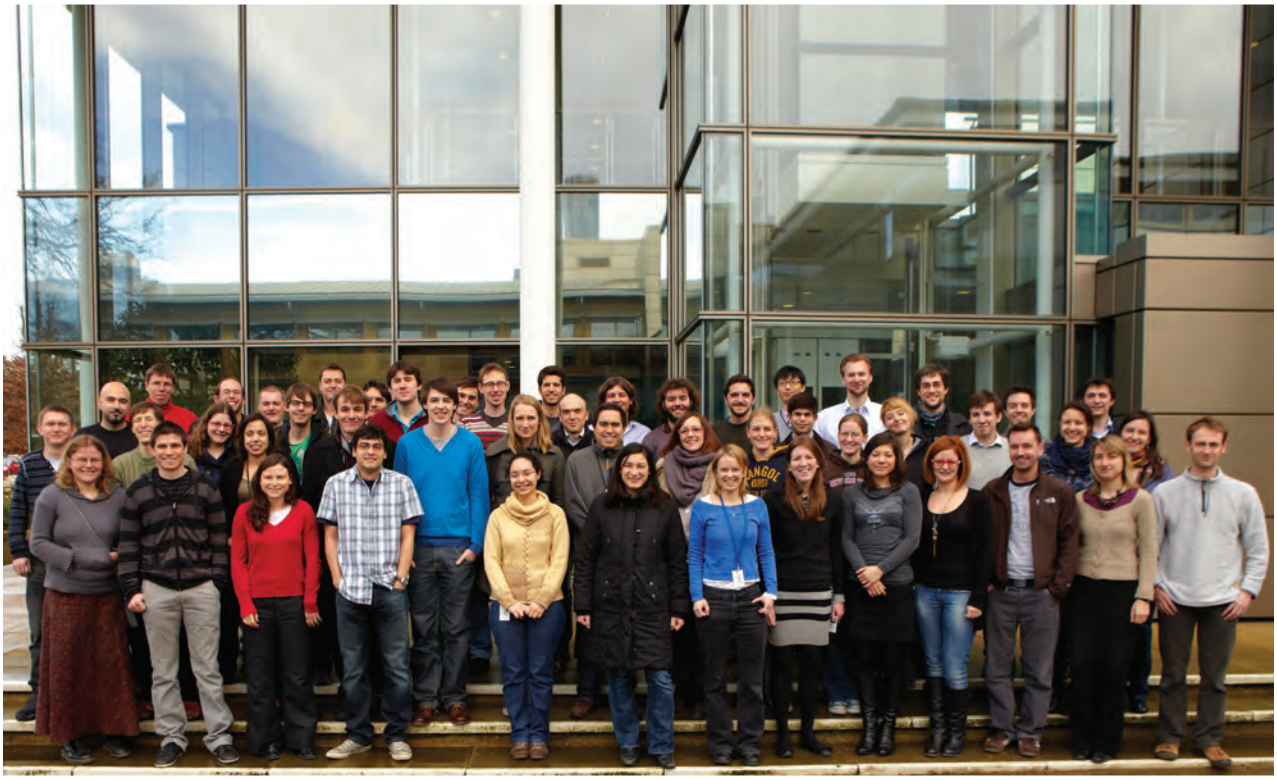


Figure 1. Postdocs and predocs at the European Bioinformatics Institute on a chilly March day in 2011.

EMBL International PhD Programme students at the EBI in 2011

Jose Assuncao

Benedetta Frida Baldi

Nenad Bartonicek

Stephan Beisken

Adam Bernard

Filipe Cadete

Samuel Croset

Andre Faure

Julia Fischer

Joseph Foster

Markus Fritz

Angela Goncalves

Mar Gonzalez-Porta

Christina Hoyer

Gregory Jordan

Myrto Areti Kostadima

Felix Kruger

Chen Li

Michele Mattioni

Inigo Martincorena

Sergio Martinez Cuesta

John May

Pablo Moreno

Anika Oellrich

Sarah Parks

Jean-Baptiste Pettit

Leonor Quintais

Dace Ruklisa

Rita Santos

Petra Schwalie

Ben Stauch

Tamara Steijger

Robert Sugar

Camille Terfve

Sander Timmer

Diva Tommei

Ying Yan

Judith Zaugg

Matthias Ziehm

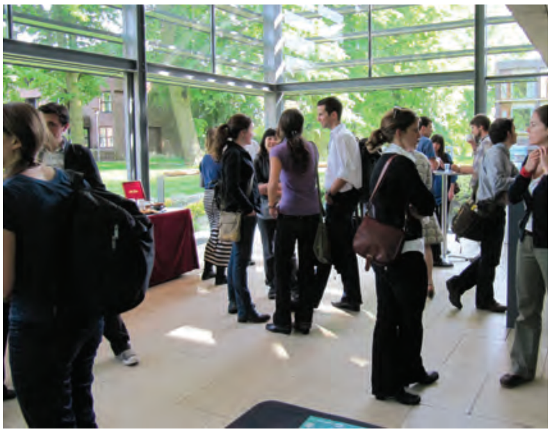


Figure 2. In 2011 PhD students organised the annual EMBL-EBI Science and Society event, “Biodiversity and endangered species: rethinking the balance of nature.” Four invited speakers from varied backgrounds presented unique points of view and shared new scientific findings. The event, held at Fitzwilliam College and the University of Cambridge and funded by EMBL’s Science and Society Programme, was widely publicised and open to all. Approximately 100 people attended.

Research Teams

BERTONE GROUP

Group Leader

Paul Bertone

Postdoctoral Fellows

Pär Engström

Remco Loos

PhD students

Ewan Johnstone*

Myrto Kostadima

Tamara Steijger

Diva Tommei

Administrative Assistant

Zoe Higney

Visitor

Beatriz Rosón*

ENRIGHT GROUP

Group Leader

Anton Enright

Bioinformaticians

Stijn van Dongen

Postdoctoral Fellows

Cei Abreu-Goodger

Mat Davis

Harpreet Saini

Iain Wallace

PhD Students

Nenad Bartonicek

Afonso Guerra-Assuncao

Leonor Quintais

Visitors

Barney Couch

GOLDMAN GROUP

Group Leader

Nick Goldman

Postdoctoral Fellows

Christophe Dessimoz*

Ari Loytynoja*

Tim Massingham

Botond Sipos

Software Engineers

Hazel Marsden

Nicolas Rodriguez (shared among multiple groups)

Team Administration

Tracey Andrew (shared among multiple groups)

Kathryn Hardwick* (shared among multiple groups)

Zoe Higney* (shared among multiple groups

PhD Students

Kevin Gori*

Greg Jordan

Sarah Parks

LE NOVERE GROUP

Group Leader

Nicolas Le Novère

Coordinator

Camille Laibe

Software Engineers

Finja Büchel*

Sarah Keating

Florian Mittag*

Stuart Moodie*

Nicolas Rodriguez

Maciej Swat*

Michael Schubert*

Curators

Vijilashkimi Chelliah

Lukas Edler*

Nick Juty

Postdoctoral Fellows

Vladimir Kiselev*

Massimo Lai

Yang Zhan

Jumei Zhu*

PhD Students

Benedetta Frida Baldi

Christine Hoyer

Michele Mattioni

Visitors

Ishan Ajmera*

Denis Brun*

Stuart Edelstein

Gael Jalowcki*

Lu Li

Michael Schubert*

Youwei Zhu*

Anna Zhukova*

LUSCOMBE GROUP

Group Leader

Nicholas Luscombe

Staff Scientists

Juanma Vaquerizas

Postdoctoral Fellows

Borbala Gerle*

Kathikeyan Sivaraman

Kathi Zarnack

PhD Students

Filipe Cadete

Florence Cavalli*

Maria Dermit*

Inigo Martincorena

Judith Zaugg

Visitors

Alessandra Vigilante*

Web Systems Administrators

Pedro Alburquerque

Philip Lewis

Robert Langlois

Dietmar Sturmayr

MARIONI GROUP

Group Leader

John Marioni

Postdoctoral Fellows

Jong-Kyoung Kim

Nuno Fonseca

Anestis Touloumis

Predocs

Jean-Baptistse Pettit

Konrad Rudolph

REBHOLZ-SCHUHMANN GROUP

Group Leader

Dietrich Rebholz-Schuhmann

Staff Scientists

Christoph Grabmüller

Jee-Hyub Kim

Vivian Lee

Ian Lewin

Software Engineers

Silvestras Kavaliauskas*

Menaka Naraysamy*

Shyamasree Saha

PhD Students

Adam Bernard*

Samuel Croset*

Chen Li*

Anika Oellrich

Ying Yan

Visitors

Senay Kafkas

Maria Liakata

Visiting Students

David Campos*

Elisabet Casanova*

Irina Colgiu*

Ernesto Jimeno Ruiz*

Electra Tapanari*

THORNTON GROUP

EMBL-EBI Director and Research Group Leader

Janet Thornton

Staff Scientists

Pedro J. Ballester

Tjaart de Beer

Nicholas Furnham

Gemma Holliday*

Roman Laskowski

Xun Li*

Irene Papatheodorou

Syed Asad Rahman

Daniela Wieser

Sophie Williams

PhD students

Sergio Martinez Cuesta

Julia Fischer*

Matthias Ziehm

Personal Assistant to the Director

Helen Barker-Dobson*

Deborah Howe*

PA to the Director’s office

Stacy Schab

ELIXIR Office

Andrew Lyall

Nicola Slater

Visitors

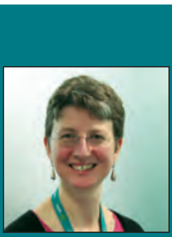
Nimish Gopel

Anna Gorel

Gilleain Torrance*

Support

Outreach and Training Programme



Cath Brooksbank

Head of Outreach and Training

PhD in Biochemistry, University of Cambridge, 1993. Elsevier Trends, Cambridge and London, UK, 1993–2000. Nature Reviews, London, 2000–2002. At EMBL-EBI since 2002.

Nick Goldman

Training Coordinator

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.

DESCRIPTION OF ACTIVITIES

As part of EMBL-EBI's mission to serve the research community, our team engages with a wide range of people who need to know about what EMBL-EBI does. For example, our user-training programme empowers scientists at all career stages to make the most of Europe's core biological data resources and liaises with the wider bioinformatics training community. We engage with new user groups in academia and in industry; showcase EMBL-EBI's career options; publicise EMBL-EBI's work via press campaigns and publications; engage with schools and the general public; and provide coherence to EMBL-EBI's growing and ever-diversifying range of events. Our team helps raise awareness of ELIXIR, Europe's emerging research- and e-infrastructure for biological data, and co-leads the development of its training strategy.

SUMMARY OF PROGRESS

- Actively involved more than 155 members of personnel in 396 events, reaching an audience of 33 000 people in 32 countries;
- Launched Train online, EMBL-EBI's free, web-based training resource (www.ebi.ac.uk/training/online);
- Ran 33 bioinformatics training courses and workshops at EMBL-EBI, serving a total of 1003 trainees (average attendance, 30);
- Delivered 24 EBI Bioinformatics Roadshows to 720 trainees in 18 countries;
- Launched the Bioinformatics Training Network (BTN), an online community of practice for bioinformatics trainers;
- Contributed to the development of shared standards for course quality in the context of EMTRAIN, an Innovative Medicines Initiative (IMI) project
- Contributed to the development of on-course®, the EMTRAIN course catalogue, and added details of more than 700 short courses;
- Redeveloped the ELIXIR website (www.elixir-europe.org; see External Relations, page 84) and created a new suite of publications;
- Issued 10 press releases and 13 research highlights, resulting in coverage of EMBL-EBI news in both the scientific and the general media;
- Developed and began to implement a social media strategy;
- Contributed to the EMBL-EBI website redesign project.

Figure 1. Students at the joint Wellcome Trust–EMBL-EBI Summer School in Bioinformatics, 2011. The Outreach and Training Programme organises a large number of hands-on courses in our purpose-built IT training room throughout the year.

Figure 2. Many of our courses involve guest faculty, for example Francis Oulette from the Ontario Institute of Cancer Research (back) and Des Higgins from Trinity College Dublin (middle). Both taught at our 2011 Summer School in Bioinformatics, a course jointly funded by the Wellcome Trust.

MAJOR ACHIEVEMENTS

A major new activity in 2011 was the launch of Train online, EMBL-EBI's free, web-based training resource. Train online provides short courses on the EMBL-EBI's most widely used data resources, created by experts in our service teams. Users do not need any previous experience of bioinformatics to benefit from this training. The goal of Train online is to help researchers to rapidly become highly competent users of EMBL-EBI data resources. Users can learn in their own time and at their own pace.

We also launched the Bioinformatics Training Network (BTN), an online community of practice for bioinformatics trainers. BTN is a completely open community resource that allows trainers to share, review and develop training materials and best practice.

In 2011 we ran a large number of events and exhibitions, actively involving more than 155 members of personnel in 396 events and reaching an audience of 33 000 people in 32 countries. This included hands-on bioinformatics training courses at EMBL-EBI, Bioinformatics 'Roadshows', training users throughought the world, conference exhibitions, careers fairs, workshops and other events.

We are an active member of EMTRAIN, an Innovative Medicines Initiative (IMI) project to establish a pan-European platform for professional development (through education and training) covering the whole life cycle of medicines research. In 2011 we contributed to EMTRAIN's development of shared standards for course quality; these were approved by all four IMI Education and Training projects and have already gained wide acceptance from the biomedical research community. Also within the context of EMTRAIN, we contributed to the development of on-course®, a comprehensive online course catalogue, and added details of more than 700 short courses.

Our outreach team played a significant role in raising the profile of ELIXIR, the incipient research infrastructure for life science data in Europe. We completely revamped the ELIXIR website (see External Relations) and created a new suite of printed publications intended for different target audiences.

We issued 10 press releases and 13 research highlights during the year, resulting in coverage of EMBL-EBI news in both the scientific and the general media. We also contributed several career profiles and interviews to relevant publications.

Selected publication

Klech, H., Brooksbank, C., *et al.* (2011) European initiative towards quality standards in education and training for discovery, development and use of medicines. *Eur. J. Pharm. Sci.* Article in press.

Schneider, M.V. and Orchard, S. (2011) Omics technologies, data and bioinformatics principles. *Methods Mol. Biol.* 719, 3-30.

Schneider, M.V., Walter, P., *et al.* (2011) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief. Bioinform.* doi: 10.1093/bib/bbr064.

Via, A., De Las Rivas, J., *et al.* (2011) Ten simple rules for developing a short bioinformatics training course. *PLoS Comput. Biol.* 7 (10), e1002245.

Looking beyond conventional media outlets, we developed and began to implement a social media strategy. The EMBL-EBI Twitter feed now has >1800 followers and our LinkedIn page has >1300 followers. We created a Facebook page, and plan to develop this presence further in 2012. Our new YouTube channel is now a central source for video content featuring our staff and resources and, like our Facebook page, has strong links with similar EMBL Heidelberg efforts. Alongside these efforts, we contributed to a major internal project to redesign the EMBL-EBI website, especially with regards to visual design and content strategy.

FUTURE PROJECTS AND GOALS

Train online represents a significant new initiative for EMBL-EBI and could help us to bring high-quality training to many more trainees than we can train face to face. We will continue to address feedback from our beta site, develop new courses and update existing courses. We will also continue to promote Train online at events and exhibitions throughout the year.

The exciting news that the UK Government will fund the construction of the ELIXIR Hub at EMBL-EBI brings with it the opportunity to offer more hands-on training. We are looking forward to working with potential ELIXIR nodes to develop ELIXIR's training plan further. Our hope is to involve trainers from all over Europe and beyond in delivering courses at the hub, at their own nodes, and potentially also online. Our work with the Bioinformatics Training Network and with the Innovative Medicines Initiative education and training projects is helping us to start shaping these plans. Another contributing factor will be BioMedBridges, a new EU-funded project that will build 'data bridges' between Europe's different biomedical science research infrastructures. We are looking forward to managing the training workpackage for BioMedBridges over the next four years.

EMBL-EBI's newly formed External Relations Team is now poised to take on the future development of ELIXIR's public relations and communications strategy, and we look forward to working with them to ensure a smooth handover.

Social media are proving to be an effective way of interacting with our user communities and of reaching out to alumni and potential staff. We will develop our presence in these media further in 2012, alongside our contributions to developing a content strategy for the EBI website. We will also be looking to reach a wider audience through the use of video for both general outreach and training efforts.

EMBL-EBI Industry Programme

Dominic Clark

PhD in Medical Informatics, University of Wales, 1988. Imperial Cancer Research Fund, 1987–1995. UK 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).

John Overington

PhD Crystallography, Birkbeck College, London, 1989. Postdoctoral research, ICRF, 1990-1992. Pfizer 1992-2000. Inpharmatica 2000-2008. At EMBL-EBI since 2008.

DESCRIPTION OF ACTIVITIES

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. The past three years have seen an expansion in the programme in terms of the breadth of topic areas and the number of subscribing members. We support and encourage precompetitive projects amongst our members by hosting regular strategy meetings and knowledge-exchange workshops. Outputs from pre-competitive projects are made publicly available, thereby sharing the benefits with interested parties in all EMBL member states. Our programme serves as an interface between EMBL-EBI and the Innovative Medicines Initiative (IMI) and the Pistoia Alliance, and encourages the involvement of industry in ELIXIR, the emerging pan-European infrastructure for biological information.

SUMMARY OF PROGRESS

- Organised regular quarterly strategy meetings with industrial partners as well as a strategically focussed retreat in Heidelberg;
- Ran eight workshops on topics prioritised by industrial partners, one scientific retreat (co-sponsored by Wellcome Trust Scientific Conferences), and one training workshop specifically for industry partners;
- Supported pre-competitive projects within the context of the Semantic Enrichment of the Scientific Literature (SESL) pilot project and other projects;
- Promoted industrial involvement in ELIXIR;
- Organised an information workshop and hands-on training for small and medium-sized enterprises (SMEs) in Piemonte, Italy;
- Welcomed two new partners to the programme: Novartis in Switzerland and UCB in the UK.

MAJOR ACHIEVEMENTS

In 2011 we ran quarterly meetings showcasing current developments at EMBL-EBI, reviewed progress on projects, prioritised future activities and presented our service development plans. We also hosted a strategic development retreat in Heidelberg for industry partners, and co-organised a scientific retreat. Member companies can incorporate information from these activities into their internal business planning processes. We also organised and ran eight knowledge-exchange workshops on topics prioritised by the industry programme members (see Table) and one training workshop on RNA-Seq and ChiP-Seq specifically for industry partners.

A major remit of our programme is to foster precompetitive projects. To that end, we invited members to knowledge-exchange workshops, where they had an opportunity to identify and document shared needs they consider to be pre-competitive. These could relate to the development of standards, support for data resources in the public domain, public information integration activities or development of new services. Once identified as being precompetitive, a project plan is developed, including industry drivers and outcomes. Once funding is agreed, EMBLEM (www.embl-em.de) draws up a legal agreement and the partners define project governance and reporting procedures. In 2011, EMBL-EBI staff worked with several industry partners to develop and publish a new standard: Minimal Information About a Bioactive Entity (MIABE; Orchard *et al.*, 2011). Another key project to deeply enhance the scientific literature, SESL, originated in the Industry Programme and was supported by the Pistoia Alliance. In 2011 the project released a demonstrator.

To further support precompetitive work and the development of standards, we are active participants in Innovative Medicines Initiative (IMI) projects and members of the Pistoia Alliance. The IMI, funded by the EU and the European Federation of Pharmaceutical Industries and Associations (EFPIA), supports collaborative projects between the European pharmaceutical industry, academia, patient organisations and regulatory agencies. We are partners in eTOX, a project that seeks to integrate bioinformatics and cheminformatics approaches to develop expert systems that allow the in silico prediction of toxicities. We are also part of EMTRAIN, which is establishing a network to facilitate and coordinate European training and education that is relevant for stakeholders of medicines research and development. EMBL-EBI plays a key role in DDMoRe (Drug Disease Model Resources), which

Figure 1. Industry Programme Quarterly Meeting at the Wellcome Trust Conference Centre in Hinxton.

aims to establish a set of standards that allow the efficient exchange and reuse of knowledge.

The importance of ELIXIR as a key European research infrastructure cannot be understated. Its realisation promises to vastly improve the translation of research discoveries into applications that advance medicine, health, agriculture and many other fields of science for the benefit of society. Because industrial involvement is essential for its success, our programme has been working closely with companies to secure their participation.

SMEs are the major drivers of the economy; however, turnover is high and the needs of this innovative sector are often more short-term than those of the larger companies. Our programme offers individuals from SMEs opportunities to take advantage of EMBL-EBI training, services and support. SMEs benefit particularly from workshops that focus on freely available tools and information resources that can add value to their business processes immediately. With this in mind, we run an annual workshop (in different locations across Europe) covering these resources tools and services with. The 2011 SME workshop was held in Piemonte, Italy, with the co-operation of BioPmed and the Bioindustry Park Silvano Fumero. The event featured hands-on training, presentations, discussion and networking. Workshop topics were selected on the basis of detailed discussions with a regional focus group and included, for example, chemogenomics, proteomics, functional genomics, web services and patent services provided by EMBL-EBI.

Workshop title	Date
Public-Private Data Challenges	Jan 2011
Foundations for Biomedical Data and Model Interoperability	Mar 2011
Bio-therapeutics	Apr 2011
Semantic Web for Industry	May 2011
Molecular Informatics Open Source Software (Scientific retreat, organised jointly with Wellcome Trust Scientific Conferences)	May 2011
Literature Services	June 2011
Chemical Registry Systems	Oct 2011
Systems Biology in Drug Discovery and Development	Nov 2011
Patent Informatics	Dec 2011

Table. EMBL-EBI Industry Programme organised eight knowledge-exchangeage workshops, a scientific retreat and a training workshop as well as a startegic retreat (held in Heidelberg, Germany). Workshop topics are chosen by our members.

Selected publication

Orchard, S., Al-Lazikani, B., Bryant, S., *et al.* (2011) Minimum Information About a Bioactive Entity (MIABE). *Nat. Rev. Drug Discov.* 10, 661-669. Published online 31 August. doi:10.1038/nrd3503.

EMBL-EBI Industry Programme members

Astellas Pharma Inc.

AstraZeneca

Bayer Pharma AG

Boehringer Ingelheim

Eli Lilly and Company

F. Hoffmann-La Roche

Galderma

GlaxoSmithKline

Johnson & Johnson Pharmaceutical Research & Development

Merck Serono S.A.

Nestlé Research Centre

Novartis Pharma AG

Novo Nordisk

Orion Pharma

Philips Research

Pfizer Ltd

Syngenta

Sanofi-Aventis Recherche & Développement

UCB

Unilever

FUTURE PROJECTS AND GOALS

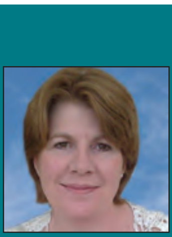
Going forward, we see our interactions with industry partners growing even stronger as the flood of data continues to rise and the need for companies to reduce costs and avoid duplication intensifies. We anticipate an increasingly pressing need for pre-competitive service collaborations, open-source software and standards development. During 2012, the programme will also be more involved in IMI projects relating to knowledge management in the area of semantic information integration. As workshops for SMEs are more and more in demand we will continue to organise these events, which provide invaluable support to an essential part of the emerging 'innovation economy'.

The award by the UK government of the capital grant for the ELIXIR hub will allow an opportunity to extend our interactions with industry from the end of 2013.

OINDUSTRY
SI
Il nostro mestiere, oggi, è inter
vvenimenti che si stanno verific
possibili sviluppi e agire di consegu
il fine di entrare da protagonisti ne

Figure 2. The annual EMBL-EBI SME Forum, co-funded by the hosting European bioregion, welcomes small and medium-sized enterprises to a series of talks and workshops. Pictured: 2011 event co-host Fabrizio Conicella of BioPmed.

External Relations



Lindsey Crosswell

BA Hons, London University, French and Environmental Science; 1997–2000, Government and Public Affairs Manager, BP plc, London; 2000–2003, Head of External Relations, Chatham House, The Royal Institute of International Affairs, London; 2003 – 2008 Director of The Oundle Society, Oundle School, Northamptonshire; 2008–2011, Educational Consultant. At EMBL-EBI since 2011.

DESCRIPTION OF ACTIVITIES

As a European, intergovernmental scientific organisation, EMBL-EBI has a broad reach throughout Europe and internationally. The role of the External Relations team, formed in September 2011 and comprising two staff with expertise in government, public affairs and EU matters, is to build and maintain strong relationships with ministries, funding bodies, policy makers and a range of academic and non-academic stakeholders within Europe. The aim of these interactions is to raise awareness about EMBL-EBI's services and research and to promote the ELIXIR research infrastructure, a flagship biological sciences research project coordinated by EMBL-EBI.

SUMMARY OF PROGRESS

- Garnered support for ELIXIR from 10 member states, who signed the Memorandum of Understanding and in so doing triggered the formation of an Interim ELIXIR Board;
- Contributed to EMBL's submission to the EU's European Research Area (ERA) Framework Consultation;
- Participated in joint activities by the Biomedical Sciences (BMS) Research Infrastructures to incorporate support for these research infrastructures in Horizon 2020;
- Continued a programme of advocacy with EU institutions to recognize the need to support the implementation of BMS Research Infrastructures in a sustainable manner;
- Launched the first edition of Informed, the ELIXIR e-newsletter;
- Hosted visiting European scientific delegations and facilitated meetings with key ministries and funding bodies in the context of ELIXIR and BioMedBridges;
- Showcased ELIXIR at three major European scientific conferences.

MAJOR ACHIEVEMENTS

Building on previous work by the Outreach and Training Team, the focus of External Relations in 2011 was firmly on ELIXIR, Europe's emerging infrastructure for biological information, which entered the fifth and final year of its preparatory phase in November. The ELIXIR management team and the steering committee for ELIXIR's preparatory phase worked with the ELIXIR stakeholders to publish the Business Case in early 2011.

Significant progress was made during the year in garnering support for the project from 10 member states, who signed the Memorandum of Understanding and in so doing triggered the formation of an Interim ELIXIR Board. The funding bodies of these member states had, by the end of 2011, committed approximately €117 million to the construction of both the Hub and Nodes of ELIXIR. The Board held its first meeting in November to appoint a Chair (Prof. Søren Brunak, Technical University of Denmark) and agree matters of governance.

Our public affairs work in 2011 included important contributions to EMBL's submission to the EU's European Research Area (ERA) Framework Consultation; participation in joint activities by the Biomedical Sciences (BMS) Research Infrastructures to incorporate support for these research infrastructures in Horizon 2020 (the EU's next framework programme for research); and continuing a programme of advocacy with EU institutions to recognize the need to support the implementation of BMS Research Infrastructures in a sustainable manner. Together with other research infrastructures, we held several meetings with Members of the European Parliament (MEPs), European Commission officials and research attachés of the Permanent Representation of Member States to the EU, in order to raise awareness of the benefits to Europe of investing in biomedical science research infrastructures – including ELIXIR.

Our team also made significant advances in communicating ELIXIR's mission and relevance to an international stakeholder audience. In September 2011 the ELIXIR website (www.elixir-europe.org) was re-launched, presenting the case for ELIXIR in a usable format for multiple audiences, and two new brochures, one speaking to the needs of the research community and another appropriate for a non-scientific audience, were produced and distributed. Continuing these efforts, our new team created Informed, the ELIXIR

Informed: ELIXIR eNewsletter

Welcome to the first edition of *Informed*, the ELIXIR eNewsletter!

ELIXIR's mission is to build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine and the environment, the bio-industries and society.

Informed will provide you with the latest news and updates on developments within the ELIXIR project.

Figure 1. *Informed*, the new ELIXIR eNewsletter.

e-newsletter, and distributed the first edition to ELIXIR's constituents. Informed will be published on a quarterly basis throughout the implementation phase of ELIXIR.

Hosting visiting scientific delegations and facilitating meetings in the context of ELIXIR and BioMedBridges is another important function of the External Relations team. EMBL-EBI hosted fact-finding meetings for, among others: a delegation of MEPs; representatives of funding bodies, including the UK's Natural Environment Research Council (NERC) and Medical Research Council (MRC); and researchers from EvaRIO, an economic impact assessment project funded under the EU's Seventh Framework Programme.

Our team also showcases ELIXIR to the European scientific community at specialist conferences. In 2011 these included the EMBO conference in Vienna, Austria; the FEBS Congress in Turin, Italy; and WIRE (Week of Innovative Regions in Europe) in Debrecen, Hungary.

FUTURE PROJECTS AND GOALS

2012 is an important year for ELIXIR, as the Interim Board begins to negotiate the International Consortium Agreement that will form the governance mechanism for ELIXIR. ELIXIR Member States will also appoint a Scientific Advisory Board, an independent body that will have responsibility for evaluating the 56 Node suggestions received from 24 countries across Europe. The External Relations team will be actively involved in the preparation of the meetings themselves and the politically complex documentation that will support this engagement.

A continued thrust of our work will be engaging with member states to support those countries working towards signature of the ELIXIR Memorandum of Understanding. We will extend our ELIXIR communication channels to include social networking. We will also present ELIXIR to a broader scientific audience, showcasing the project at a number of European conferences including ICRI in Copenhagen, Denmark; ESOF in Dublin, Ireland; EMBO in Nice, France, ECCB in Geneva, Switzerland and IUBMB/FEBS in Seville, Spain.

We anticipate considerable interest in the construction of ELIXIR and the Hub building itself at EMBL-EBI, which will begin in spring 2012. We look forward to welcoming visiting European delegations wishing to follow the physical progress of this landmark project. 2012 also marks the start of BioMedBridges, an FP7-funded project that will be coordinated by ELIXIR and makes provision for elements of ELIXIR's technical implementation. The External Relations team will help facilitate the necessary interactions between the European ESFRI BMS projects that comprise the BioMedBridges consortium.

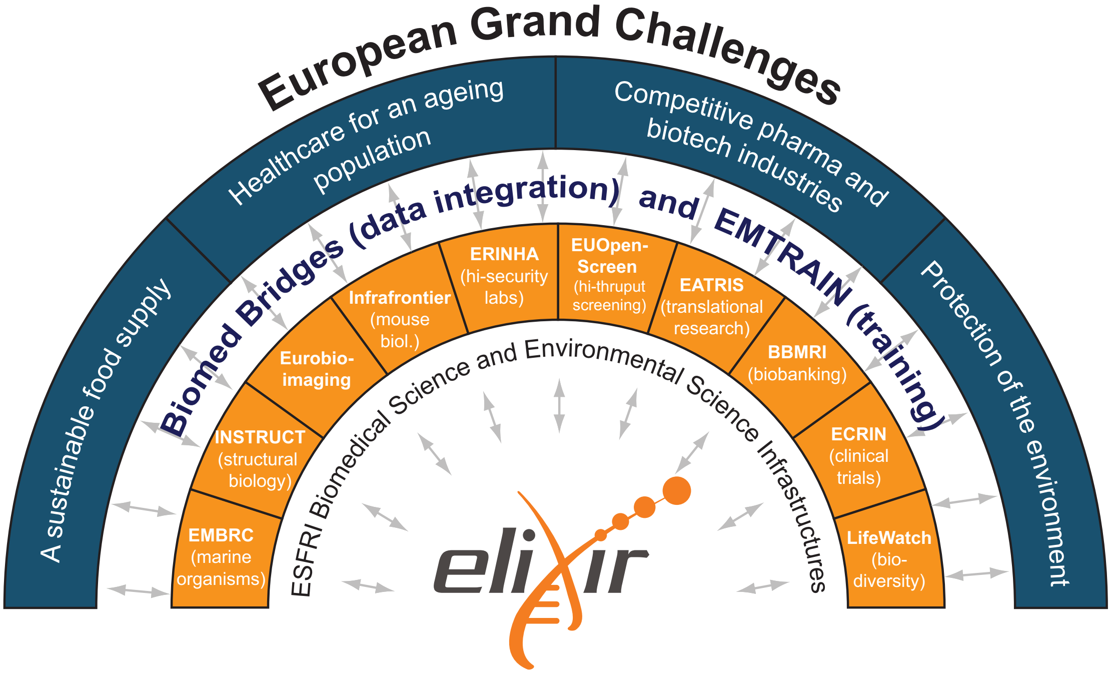

European Grand Challenges
A sustainable food supply
Healthcare for an ageing population
Competitive pharma and biotech industries
Protection of the environment
Biomed Bridges (data integration) and EMTRAIN (training)
EMBRC (marine organisms)
INSTRUCT (structural biology)
Eurobio-imaging
Infrafrontier (mouse biol.)
ERINHA (hi-security labs)
EUOpen-Screen (hi-thruput screening)
EATRIS (translational research)
BBMRI (biobanking)
ECRIN (clinical trials)
LifeWatch (bio-diversity)
ESFRI Biomedical Science and Environmental Science Infrastructures
elixir

Figure 2. The BioMedBridges project: relevance for all biomedical sciences research infrastructures.

Administration

Mark Green

Fellow of the Chartered Institute of Internal Auditors. At EMBL since 1997; joint appointment with EMBL-EBI. At EMBL-EBI since 2003.

DESCRIPTION OF ACTIVITIES

The EMBL-EBI Administration Team aims to provide a timely and efficient administrative support network for those working at EMBL-EBI. Our activities span budgetary, financial and purchasing matters; human resources; grants and external funding management; facilities management (including health and safety); as well as pre- and post-doc programmes. We coordinate and integrate administrative activities throughout EMBL-EBI in order to facilitate interactions with the wider scientific community through, for example, organising meetings and courses and arranging travel for our extremely mobile staff.

SUMMARY OF PROGRESS

- Organised and delivered the administrative aspect of the UK's Large Facilities Capital Fund Programme, which focuses on both construction of a new building on site and equipment/data facilities over the next eight years;
- Transitioned to cash budgets and streamlined the budgetary process;
- Continued efforts to attract high-quality staff through targeted recruitment and advertising;
- Started overhaul of induction process for new recruits;
- Adopted proactive approach to facilitate information flow between our grants staff and EMBL-EBI faculty;
- Introduced health and safety training for managers and addressed occupational health concerns in our computer-centric work environment (80% of ergonomic assessments completed).

MAJOR ACHIEVEMENTS

The Large Facilities Capital Funding Programme consists of two projects: the construction of a new building on the Genome Campus (estimated cost: €28 million) and the acquisition, over the next eight years, of equipment and space in a commercially run data centre (estimated cost: €60 million). The Programme is designed to help meet the growing demand for EMBL-EBI services and, in the context of ELIXIR, support life science research and its translation to medicine and the environment, the bio-industries and society. Prior to the commitment of funds by the UK government in December 2011, the Biotechnology and Biomedical Research Council (BBSRC) provided financing to prepare for these undertakings. Administration supported the teams who selected an architect and contractor to design the building (subject to contract approvals) through a tender process, and dedicated considerable efforts to overall coordination and project management as well as support for the governance structures.

Finance saw the first full year of cash budgets at EMBL-EBI in 2011. We refined our financial and budgetary process following the appointment of Senior Financial Officer Brian Nsonga in late 2010, and were able to provide budget holders with better information and more control over their resources.

The Human Resources (HR) team reviewed recruitment procedures to ensure that our efforts are directed towards the best sources of high-quality applicants. A new process of induction for new staffwas also introduced in the interests of helping newcomers get oriented to the organisation as a whole, and in many cases to adapt to life in a new country. This has already proved very popular with staff and is expected to enhance work life and increase productivity. We also participated in campus initiatives such as the 'Sex in Science' seminar series, which addresses important issues such as work–life balance.

Our Grants Office became more pro-active in their approach to communicating opportunities with grant applicants and holders. This involved holding meetings to discuss forthcoming applications including renewals, budgets, resources and other matters. The team also undertook measures to harmonise processes and procedures; efforts in this area are ongoing.

Because EMBL-EBI staff work in a computer-intensive environment, issues of occupational health are important to address regularly through measures such as ergonomic assessments. In addition to completing ergonomic assessments of the workstations of 80% of staff, we held two health and safety training courses for managers and updated the EMBL-EBI Health and Safety policy.

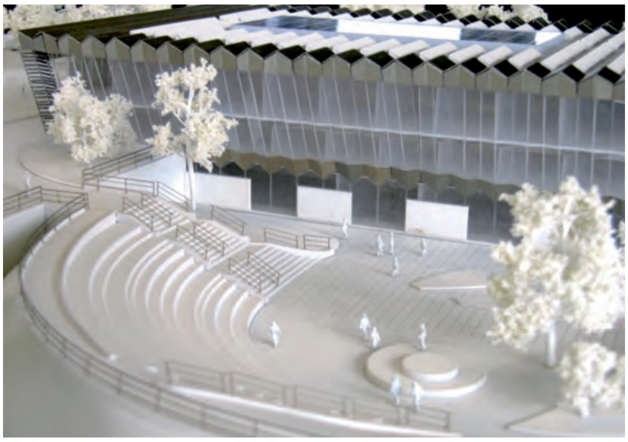


Figure 1. Model of the new EBI building, which will house the ELIXIR Hub. The Administration team is facilitating the planning process.

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.

FUTURE PROJECTS AND GOALS

In 2012 we will work to ensure the new build on campus can progress; its completion is expected by summer 2013 for occupation by Autumn 2013. We will manage the competitive tender process for establishing an approved list of suppliers for equipment purchases (to be funded by LFCF monies). We will also continue developing longer-term strategic financial plans taking account of EMBL, external and LFCF funding. We will also develop processes and procedures to facilitate the establishment of an ELIXIR Hub at EMBL-EBI. We plan to conclude the partnership agreement between EMBL-EBI and the Wellcome Trust Sanger Institute. We will continue to maintain good interactions between a wide diversity of stakeholders such as the NIH and Wellcome Trust. We will create an Alumni Working Group to facilitate interactions between current and former staff, and to help former employees become 'ambassadors' for EMBL and EMBL-EBI. We will also collaborate with our colleagues in Heidelberg to: develop and implement a grants database; provide high-quality financial reports; and complete the first phase of the EMBL risk management programme.

Figure 2. The EMBL-EBI Administration Team in 2011.

Systems and networking

Petteri Jokinen

MSc in Computer Science 1990, Helsinki University.
At EMBL-EBI since 1996.

DESCRIPTION OF ACTIVITIES

The Systems and Networking team manages EMBL-EBI's IT infrastructure, which includes compute and database servers, storage, desktop systems and networking. We also support EMBL-EBI staff in their daily computer-based activities and manage the campus Internet connection. We work closely with all project groups, maintaining and planning their specific infrastructures. As of December 2011 we handle 15 petabytes of disk storage and approximately 12 000 CPU cores.

SUMMARY OF PROGRESS

- Activated Global Server Load Balancing (GSLB) devices in London to distribute service traffic between the two London data centres;
- Drew up a new network architecture to improve redundancy, bandwidth and predictability;
- Built 'vaults' to mitigate the possibility of accidental or malicious data leaks;
- Implemented a new SSL-based virtual private network (VPN) for staff with systems-maintained laptops;
- Implemented a private cloud infrastructure to host UK PubMed Central at EMBL-EBI;
- Completed migration of all Oracle databases to version 11g and to the Linux operation system;
- Moved 18 public databases from Hinxton to the London data centres;
- Configured Legato backup software and a tape robot to back up the Oracle databases in half the usual time.

MAJOR ACHIEVEMENTS

Networking

We activated Global Server Load Balancing (GSLB) devices in London to distribute service traffic between the two London data centres, and deployed a new generation of network load balancers to the Hinxton data centre. This increased capacity, improved redundancy and reduced complexity. In Hinxton we also improved the desktop network, preparing for increased security. We worked hard to maintain a consistent data-centre network service against the pressures of rising network traffic volumes and numbers of hosts.

Planning for the future, we concluded that the current network architecture would not continue to scale adequately beyond 2012. We drew up a new network architecture and made the first procurement at the end of 2011. The benefits of the new architecture include improved redundancy, bandwidth and predictability.

We evaluated our DNS and DHCP architecture, as these fundamental services must be maintained effectively so that EMBL-EBI may continue to grow. We procured a commercial DNS/DHCP/IP address management system that will reduce the administrative burden significantly and permit more automation and flexibility.

Figure 1. Servers in the Hinxton data centre

Some groups, principally the European Genome–phenome Archive (EGA), handle large volumes of confidential data. To support this work we built 'vaults' to mitigate the possibility of accidental or malicious data leaks. In the case of EGA, we built an entirely separate hardware infrastructure separated from EMBL-EBI by a firewall, and provided strong authentication for access to any device within the EGA zone. A major benefit of this work is that the hardware and software infrastructure for secure access to the EGA vault also supports generic remote access for EMBL-EBI staff. We were able to implement a new SSL-based virtual private network (VPN) for staff with systems-maintained laptops. This also allows staff remote access to the intranet and electronic resources such as journals using just a web browser.

Desktops

We continued to support users at EMBL-EBI (>500 in 2011), and implemented a management system to handle the increasing numbers of OSX clients. We bolstered efforts to virtualise all relevant services across the Windows and Linux platforms. A VPN was made available to all users and an Oracle SGD was updated to offer secure, remote access to centralised applications. The team also implemented a private cloud infrastructure to host UK PubMed Central at EMBL-EBI.

Databases

The team completed the migration of all the Oracle databases to version 11g and to the Linux operation system. We also moved 18 public databases from Hinxton to the London data centres, and improved database resilience by configuring standby databases in the Flint Cross data centre. Provisioning of new test and development databases was greatly enhanced by the use of Delphix software, which implements 'point-in-time' virtual Oracle instances. We configured Legato backup software and a tape robot to back up the Oracle databases in half the usual time.

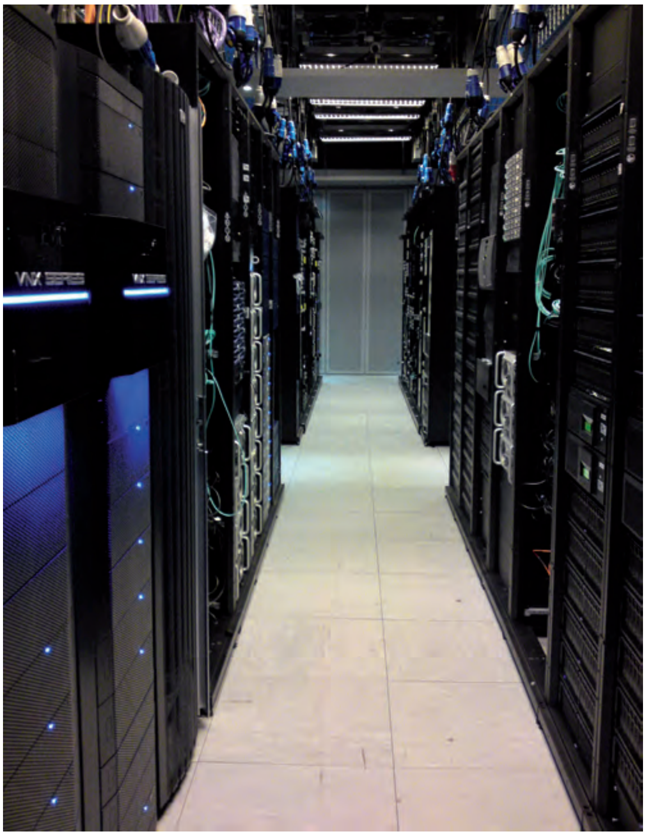


Figure 2. A section of the Hinxton Data Centre.

In 2011 we implemented comprehensive monitoring for EMBL-EBI's Mysql database instances. We audited and moved to new storage close to a third (100) of our total instances; 12 instances were successfully moved to the London data centres via a custom-developed user-directed procedure. We created 50 new Mysql instances and retired 36 old ones.

Devices, storage and statistics

The team migrated 84% of nodes on the EBI cluster from CentOS 5.4 to RedHat 6.1, with minimal impact on users; we expect to complete the process in 2012. We also obtained two large memory machines, each with 2 terabytes (TB) of RAM, which were incorporated into the EMBL-EBI farm on the Hinxton Campus. After investigating new technologies and alternatives to our current storage, we procured flash-memory-based storage to use for time-critical applications. We also obtained a new virtualised SAN storage system.

We started to collect statistics of device failures, which allows us to closely monitor the quality of various hardware components and to make better purchasing decisions. We also spent considerable efforts collecting statistics from our storage usage, which helped ensure that users at EMBL-EBI are more careful in choosing what data to store on the long term. In terms of storage performance monitoring, we saw the biggest growth in the Sequence Read Archive (SRA) system.

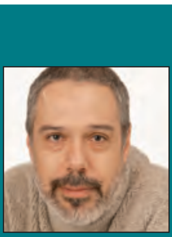
FUTURE PROJECTS AND GOALS

To ensure that the current network architecture scales adequately beyond 2012, we will deploy a new network architecture that improves redundancy, bandwidth and predictability. We will also roll out a commercial DNS/DHCP/IP address management system to reduce the administrative burden and permit more automation and flexibility.

Approximately 80% of our data is backed to tapes to enable long-term recoveries, the short term (maximum one week) recoveries are done using disk-based replication. In 2012 a new Linux-based tape backup system that uses a tape robot will be completed.

We have been investigating various ways of providing cloud services, and have two different functioning solutions. However, we are still investigating alternatives so that we can have a reasonably sized cloud service running during 2012. This will allow users to maintain their own virtual machines that run in the cloud, and provide access to the EMBL-EBI data store.

External Services



Rodrigo Lopez

Vet. Med Degree 1984, Oslo Vet.Hoyskole and NASAS Cand. Scient. Molecular Toxicology and Informatics 1987, University of Oslo.
At EMBL-EBI since 1995.

DESCRIPTION OF ACTIVITIES

The External Services (ES) Team focuses on three major areas of service: We provide platforms for web development, administration of the web infrastructure and software service frameworks. Platforms are mainly concerned with the design, deployment and maintenance of web-publishing frameworks using DRUPAL, Confluence, Jira, Wikis for the EBI web portal and various websites for EU-funded projects. Web infrastructure administration focuses on providing reliable web architectures (i.e. Clouds) for serving EMBL-EBI's databases and tools. Software services comprise core developments that allow EMBL-EBI to deploy tools using SOAP and REST web services as well as the EBI search engine.

In 2011 the continuing migration of EMBL-EBI's core services from Hinxton to the London data centres and the redesign of the EBI website were the major focus areas for External Services web administrators and web developers.

SUMMARY OF PROGRESS

- Enabled a sustained increase in the usage of EBI services;
- Launched the new EBI Search service – just one example of our work on integrating EBI's data resources for the benefit of our users;
- Unified EBI web portal technology with DRUPAL;
- Gathered usage statistics for EBI websites and facilitated their analysis (ongoing);
- Continued to improve and support EBI services.

MAJOR ACHIEVEMENTS

Increased usage of EBI services

The move of EMBL-EBI services to London data centres substantially improved international connectivity, providing more reliable access to our compute resources. The result was increased usage of EMBL-EBI's data resources and more regular, timely distribution of data sets to those who maintain local services and internal analytical pipelines. Sequence-analysis tools saw marked growth in usage; in particular those for which External Services is responsible: BLAST, InterProScan and Clustal Omega. Access to EBI resources using web services proved very popular: 58% of all job requests for services such as BLAST, FASTA and InterProScan during 2011 occurred using the SOAP and REST programmatic interfaces. This usage came primarily from laboratories in Europe, the Americas and Asia.

In 2011 the average number of jobs per month was 3 million (up from 2.2M in 2010). The number of data sets available for sequence searching reached over 4000, including many species, strains and assemblies from Ensembl Genomes.

Integration of web resources

Providing a platform for collaboration and resource sharing was an important goal for EMBL-EBI during 2011. Thanks in part to our team's efforts, groups throughout the organisation were able to work together more seamlessly through sharing resources and consolidating services.

External Services is responsible for the EBI's search technologies, which are used in portals including the European Nucleotide Archive (ENA)/EMBL-Bank, ENA's Sequence Read Archive, Ensembl Genomes, the Enzyme Portal (to launch in early 2012) and LRG (the Locus Reference Genomic sequences resource); EBI projects such as InterPro, Pombase and MetaboLights are potentially in the pipeline for take-up in 2012. Some high-profile external consumers of these services include Unilever, Qfab and UQ in Australia (EMBL-Australia).

During 2011 a new way of delivering EBI-wide search results was developed. The new Search service is modelled on the traditional central dogma of molecular biology (gene > gene expression > transcript > protein) and draws a virtual graph connecting biologically related data items from Ensembl, UniProt, InterPro, PDBe, ArrayExpress, the Expression Atlas and CiteXplore. These paths are used to generate a cohesive view of the biomolecular data, and to represent it to users in an easily navigable way. Work is underway to include cheminformatic and metabolomic resources ChEMBL, ChEBI and Reactome in the virtual graph, which will enable users to explore our resources still more deeply.

External Services also supports a suite of heavily used bioinformatics tools, for example NCBI BLAST, PSI-Search and Clustal Omega. Bioinformatics services that rely on these tools include major EBI collaborations (e.g. UniProt, Ensembl Genomes, the Protein Data Bank in Europe, international ImMunoGeneTics projects IMGT/LIGM and IMGT/HLA) as well as external users (e.g. Blast2GO, BlastStation, STRAP, T-Coffee, CCP4, Geneious, GMU-metagenomics).

On the web front, the open-source content management system DRUPAL has proven incredibly useful. EMBL-EBI web developers and content authors alike have found it fun to work with, as it allows us to integrate web content and deliver data in a logical, functional and intuitive way. In DRUPAL, code is easy to test, share and proof, making it easier to streamline coding practises and produce user-oriented portals that are adaptable to devices such as smartphones and tablets.

Web portals

User-oriented web portal design is an important component of External Services operations. Focusing on the interaction between human users, machines and contextual environments helps us design systems that address the user's experience. In 2011, user-experience design (UXD) methodologies were fully integrated into our design process so that we can continue to meet and support user needs and goals. UXD has proven beneficial to both the organisation and scientific users in a number of ways, for example through design simplification, the removal of unnecessary features, usability optimisation, streamlining design–development efforts and integrating business and marketing goals. UXD is central to the successful implementation of the new EBI web portal, which is scheduled for the second half of 2012.

External Services is responsible for the main EBI website and for more than 30 project portals. These include the 1000 Genomes Project, BioCatalogue, DVGa, the European Genome-phenome Archive, ELIXIR, ENA, ENFIN, SLING, IMPACT, INSDC, Microme, the Bioinformatics Training Network and Train online. These project portals were developed (or redeveloped) in 2011 using DRUPAL. Their success is a testament to the platform's ability to effectively integrate resources across all EMBL-EBI teams.

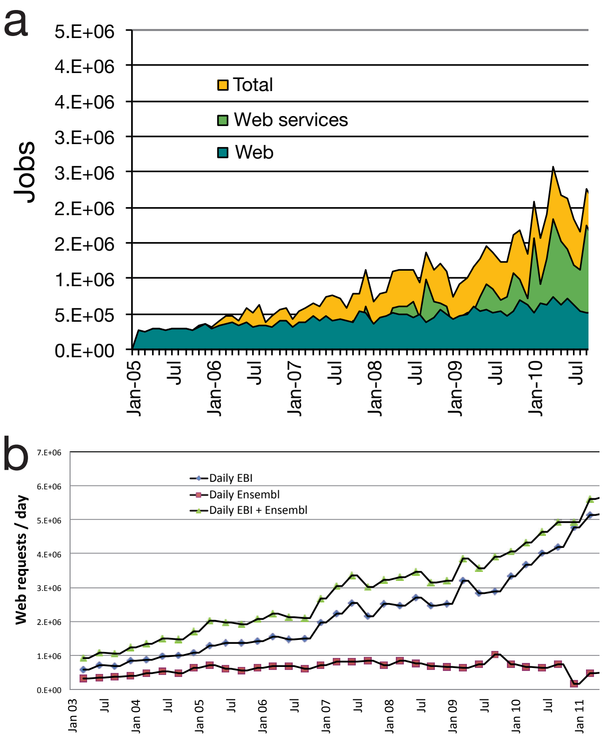


Figure. Usage statistics and supporting the use of EBI Services. (a) Jobs per month since 2005 showing Web browser usage, programmatic access using SOAP/REST Web Services and total for tools such as Blast, Fasta, ClustalOmega and InterProScan. (b) Quarterly requests per day for www.ebi.ac.uk and www.ensembl.org since Jan 2003.

Selected publications

Hunter, S., Jones, P., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40 (Database issue), D306-12.

Robinson, J., Mistry, K., *et al.* (2011) The IMGT/HLA database. *Nucleic Acids Res.* 39 (Database issue), D1171-6.

Schneider, M.V., Walter, P., *et al.* (2011) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief. Bioinform.* doi: 10.1093/bib/bbr064.

Sievers, F., Wilm, A., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7 539.

Support

External Services monitors and reports on the health and activity of the EBI portals. In 2011 our Help Desk activities diversified quite significantly, handling more than 1 500 tickets, covering: technical help for programmatic access and data acquisition; user problems with the web portals; best practices and consulting; and outreach and training. We received over 3 100 requests from internal users for: help with web node allocation; issues with web content management; new tool deployments; data production; logistics, maintenance and service change management.

Outreach and training

External Services plays a key role in many of the EBI's core activities, and as such makes a unique contribution to a variety of training events on-site in Hinxton (e.g. for programmatic access to resources) and in various roadshow programmes throughout the world. In 2011 we presented our nonredundant patent sequence databases, programmatically accessible tools and our search engine at events in Czech Republic, Finland (2), Latvia, the Netherlands (2), Romania (2) and the UK.

We also participated in EMBL-EBI's Bioinformatics Roadshow (Athens, Greece and Cuernavaca, Mexico) and many other workshops showcasing the EBI's core resources that we develop and maintain. These were held in Austria, Croatia, Greece, Italy (2), Colombia (South America) and the US.

FUTURE PROJECTS AND GOALS

The deployment of a new EBI web site is a major undertaking. Our focus in 2012 is on designing a portal that is responsive to our users' needs. We are also evolving our working practises to ensure that we remain at the vanguard of developments. Our platform-based services will serve to provide the fundamental building blocks for ELIXIR. With our web infrastructure solutions and use of virtualisation technology, we will establish a storage cloud that will allow users to work closer to the 'big data'. Improving the discoverability of our data is central to the development of the search engine; in 2012 we will continue to provide our users with better, more relevant query results, and to help them explore the data in truly meaningful ways.

Support Teams

OUTREACH AND TRAINING

Head of Outreach and Training

Cath Brooksbank

Outreach Programme

Mary Todd Bergman

Katrina Pavelin

Louisa Wood*

George Zarkadakis*

User Training Coordinator

Vicky Schneider

Scientific Training Officers

James Watson

Victoria Wright

Training Information and Liaison Officer (EMTRAIN)

Claire Johnson

Workshops and Exhibitions Organisers

Alison Barker

Holly Foster (Events Supervisor)

Johanna Langrish

Frank O'Donnell

Visitors

Filzah Ehtesham*

Abigail Wright*

Joanna Argasinska*

INDUSTRY PROGRAMME

Industry Programme Manager

Dominic Clark

Administrative Officer

Delphine Gandelin

EXTERNAL RELATIONS

Head of External Relations

Lindsey Crosswell

External Relations Officer

Andrew Smith

ADMINISTRATION

Head of Administration

Mark Green

Financial Management

Brian Nsonga

Finance and Purchase

Julie Mace

Christine Pettit

Jean Scammell

Grants Office

Nishma Chauhan

Susan Haller

Gabriele Picarella

Tom Ratcliff

Human Resources

Julia Lant

Sue Lee

Jane Stace

Facilities Management

Hilary Little

Reception

Peggy Nunn

Sally Wedlock

Research Office

Tracey Andrew

Zoe Higny

SYSTEMS AND NETWORKING

Team Leader

Petteri Jokinen

Server and Networking Team

Jonathan Barker

Elizabeth Beresford

Gianluca Busiello

Salvatore Di Nardo

Dawn Johnson

Gavin Kelman

Manuela Menchi

Pravin Patel*

Asier Roa

Radoslaw Ryckowski*

Michal Wieczorek

Desktop Team

William Barber

Richard Boyce

Karen Briggs

Andy Cafferkey

Pelham Martinez-Lainez

Systems Database Administrators

Andy Bryant

Mike Donelly

Luis Figueira

Pieter Van Rensburg

Technical Administrator

Carolina Bejar

EXTERNAL SERVICES

Team Leader

Rodrigo Lopez

Software Engineers

Mickael Goujon

Weizhong Li

Hamish McWilliam

Eric Nzuobontane*

Juri Paern

Young Mi Park

Silvano Squizzato

Franck Valentin*

Web Developers

Asif Kibria*

Thomas Laurent

Gulam Patel*

Stephen Robinson*

Francis Rowland

Brendan Vaughan

Peter Walker

Web Systems Administrators

Pedro Alburquerque

Rober Langlois

Philip Lewis

Dietmar Sturmayr

Support and Training

Andrew Cowley

Jennifer McDowall

User Experience Analyst

Jenny Cham

Facts and Figures

2011: A year in numbers

SERVICES

EMBL-EBI hosts the major, core biomolecular data resources of Europe. We collect, archive and distribute data throughout Europe and beyond. These services continued to be well used during 2011 (see External Services report, page 88).

Usage statistics

In 2011 there were on average 5.3 million requests on our services per day, not including Ensembl (compare to 4.1 million in 2010). Including Ensembl, the number of daily requests was 5.7 million on average (Figure 1), compared to 4.7 million in 2010.

We have seen steady growth not only in usage but in the number of computers accessing our services. The number of unique IPs, or web addresses, accessing our website grew by 6.1% between 2010 and 2011. We approach these figures with caution, as an IP address could represent a single person or an entire organisation.

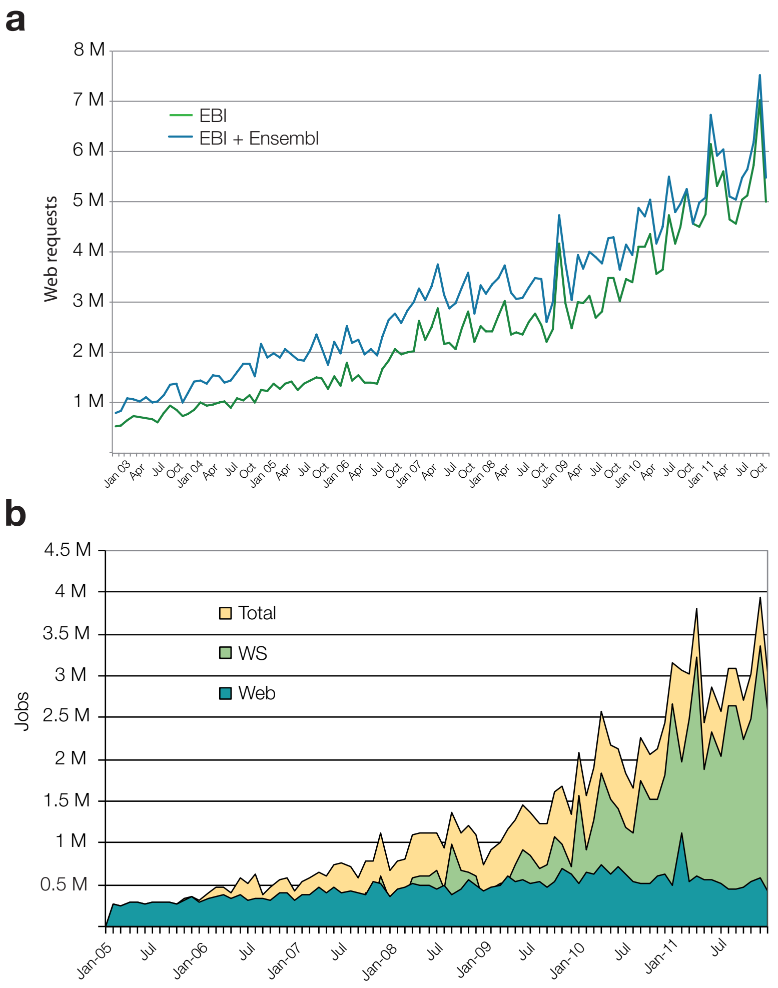


Figure 1. (a) Web requests per day fom January 2003 through December 2011. (b) Job requests per month month from January 2005 to December 2011. WS, web services: automated access to our data (pipelines); Web, requests ('hits') on our website.

Growth of core resources

In 2011 all of our core data resources grew substantially (Figure 2).

- EMBL-Bank, part of the European Nucleotide Archive (ENA), received and processed more than 8 x 10^{10} bases in 2011. Overall, ENA received and processed over 1 x 10^{14} bases. This represents a four-fold increase since last year, and means that 75% of the data have been in the archive for one year or less.
- UniProt processed over 7.4 million entries (3.3 million in 2010).
- Our Functional Genomics resources processed 310 651 assays (209 782 last year).
- PDBe processed 8994 macromolecular structures (6958 last year).
- Ensembl and Ensembl Genomes welcomed 120 new genomes (78 last year). Ensembl now holds 67 eukaryotic genomes and Ensembl Genomes holds 335 non-vertebrate genomes.

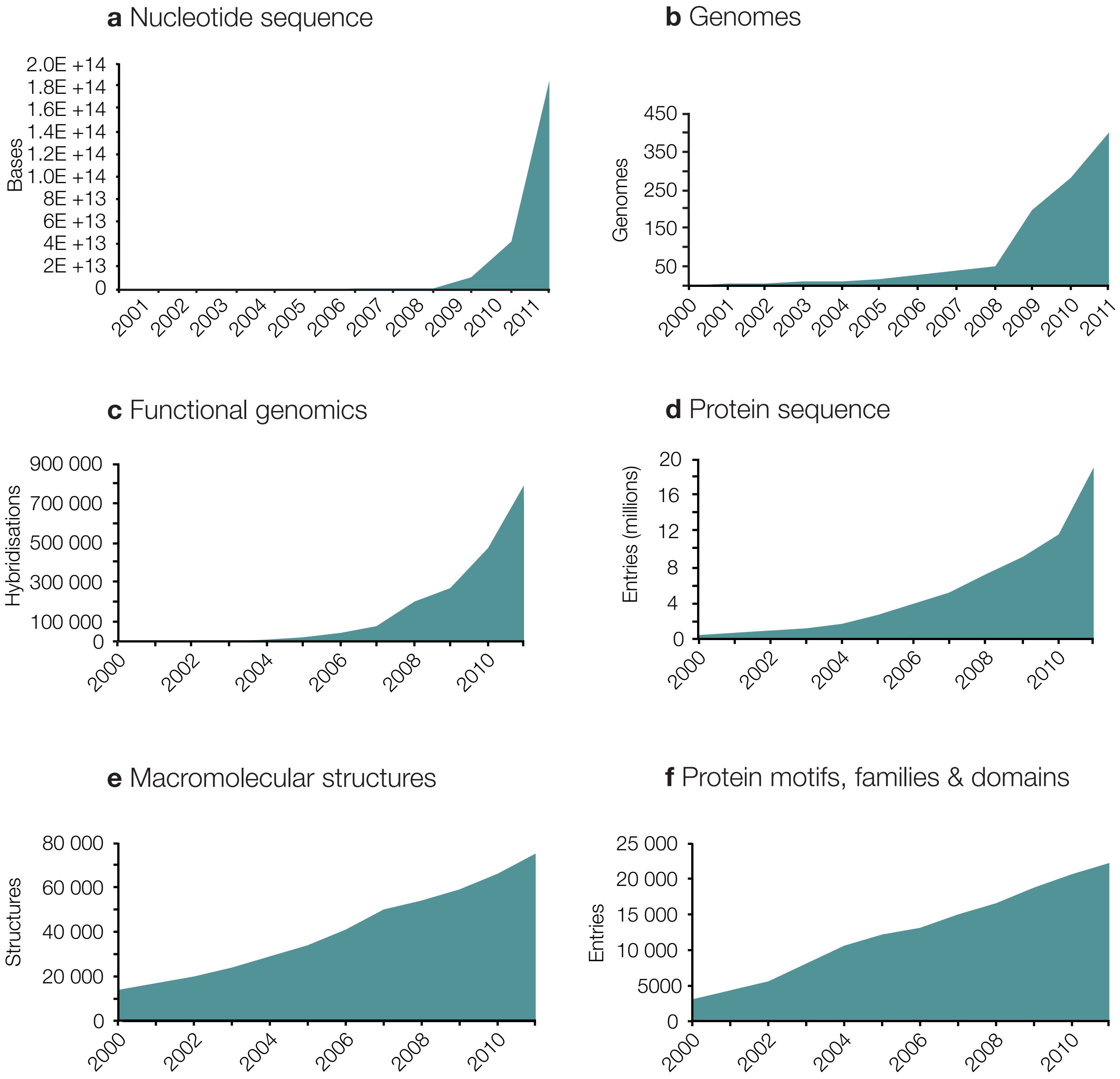


Figure 2. Growth of EMBL-EBI's core data resources from 2000 to 2011. (a) Nucleotide sequence (bases in the European Nucleotide Archive); (b) genomes (entire genomes in Ensembl plus Ensembl Genomes combined); (c) functional genomics (assays in the ArrayExpress Archive,); (d) protein sequence (protein sequences in UniParc); (e) macromolecular structures (structures in PDBe); (f) protein families, motifs and domains (entries in InterPro).

RESEARCH

EBI staff published 202 scholarly articles in 2011. Of these, 78 were basic research papers, 53 presented new methods or standards, and 71 related pertinent information about resources developed in our service and research groups to the scientific community. During the year we had 39 PhD students at the EBI (see page 74) and seven new students started the programme by attending the EMBL core course. Eight students were awarded their PhD, and two submitted a thesis.

EMBL-EBI Group and Team Leaders successfully applied for external support, receiving research funding to the tune of €6.4 million over the next two to five years (compare to €2.8 million in 2010).

COLLABORATIONS

EMBL-EBI benefits from extensive collaboration with partners throughout the world (Figure 3). Almost all of our resources are funded through collaborative agreements, and during the reporting period 90% of our publications involved collaborations with colleagues at institutes throughout the world (compared with 77% last year).

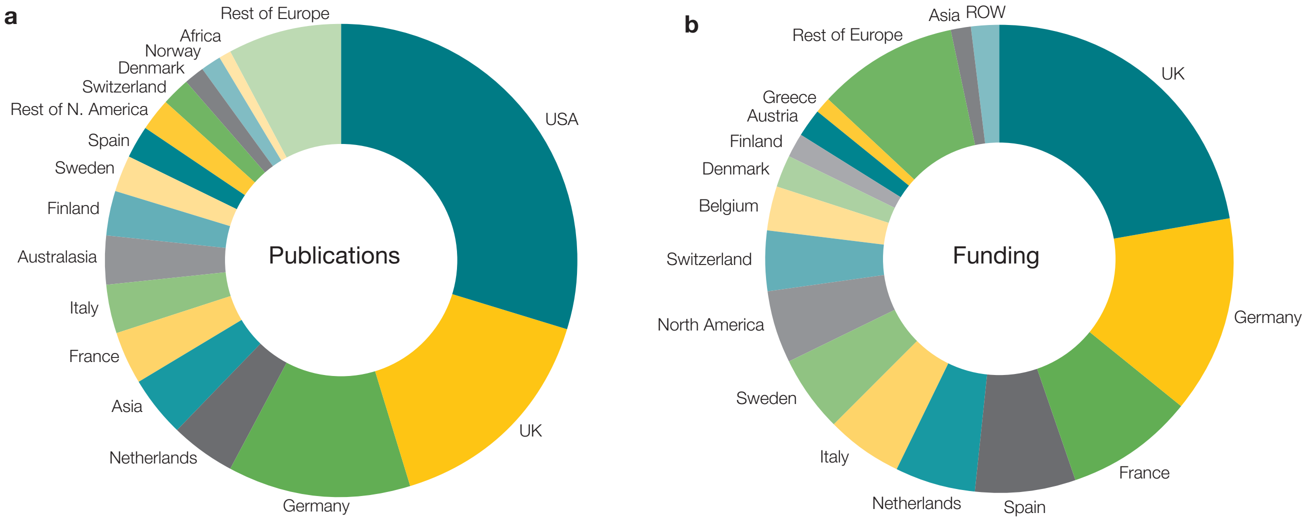


Figure 3. Collaborations as measured by (a) publications with other institutions and (b) funding shared with other institutions. Data for (a) were taken from affiliations in peer-reviewed publications, and were de-duplicated if the same institution appeared in the affiliations list of more than one paper. Data for (b) were not de-duplicated and in some cases the same institution is represented several times through different collaborations.

OUTREACH, TRAINING AND INDUSTRY

We took part in 396 training events during the reporting period (467 last year), reaching over 33 000 participants in 32 countries.

Our hands-on training programme ran 33 courses on site in 2011 (10 last year), serving a total of 1003 trainees. The average number of participants was 30 per course. Delegates hailed from all over the world – over half came to us from outside the UK.

The training programme delivered 24 roadshows to 720 trainees in 18 countries. Of these, 13 were run under the auspices of the EU-funded SLING Integrating Action (9 last year), serving 276 trainees (average attendance, 31), the majority of whom were from Europe.

We delivered another 11 roadshows in five countries ('non-SLING funded') in the same period, serving 273 trainees. There were between six and 54 attendees at these roadshows, 88% (242) of whom hailed from outside of Europe.

The Industry Programme ran 8 workshops (11 last year) for its members, serving a total of 309 delegates (range, 20–72 per event). The Industry Programme raised approximately €250 000 through pre-competitive projects.

FUNDING AND RESOURCE ALLOCATION

We raised €23.08 million in external funding for 2011, compared to €18.40 million in the last reporting period (Figure 4a). Total internal funding to EMBL-EBI in 2011 was €24.43 million (€22.40 million in the year 2010), of which 47% was spent on salaries (58% last year; Figure 4b).

We spent €8.9 million on computing equipment (€3.0 million in 2010 when we also benefitted from LFCF equipment funds of €6 million – not shown in Figure 6), which represents 30% of our total internal spend (Figure 5).

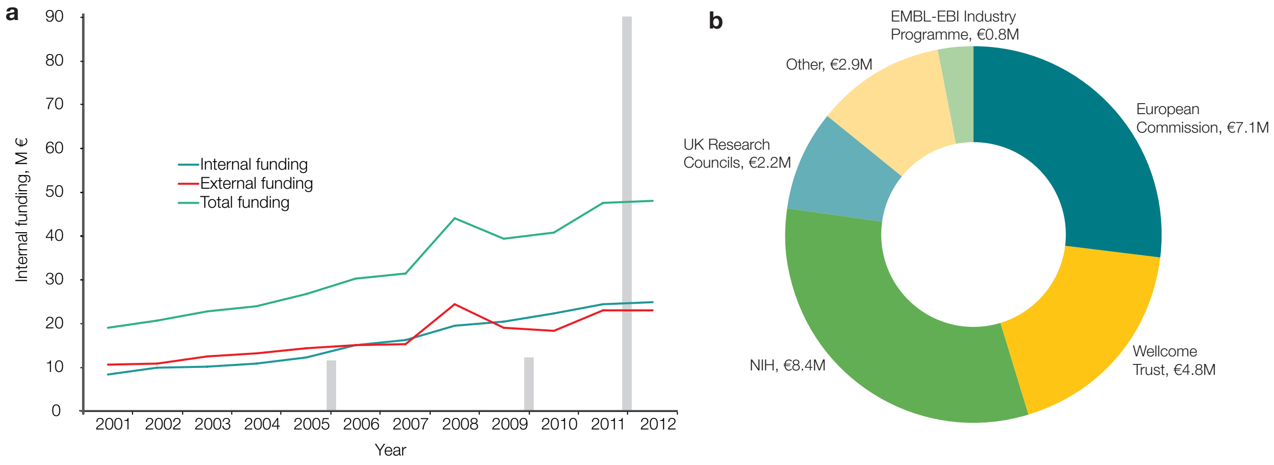


Figure 4. (a) Growth of internal, external and total funds from 2001 to December 2011, and agreed internal funds for 2012 (the start of the EMBL indicative scheme 2012–2016). Bars indicate episodic capital funds: 2005: Wellcome Trust, EMBL and Research Councils UK funding (11 M€) committed for EBI East Wing, which opened in summer 2007; 2009: UK funding for data service provision including acquisition of space and equipment (11.5 M€); 2011: the UK government committed £75 million (90 M€) as a contribution towards ELIXIR and will see the construction of a new building at EMBL-EBI (which will house the ELIXIR Hub and provide facilities for other key EBI activities) and expenses for the further acqusition of off-site data centre space and equipment. (b) Sources of external funding for 2011.

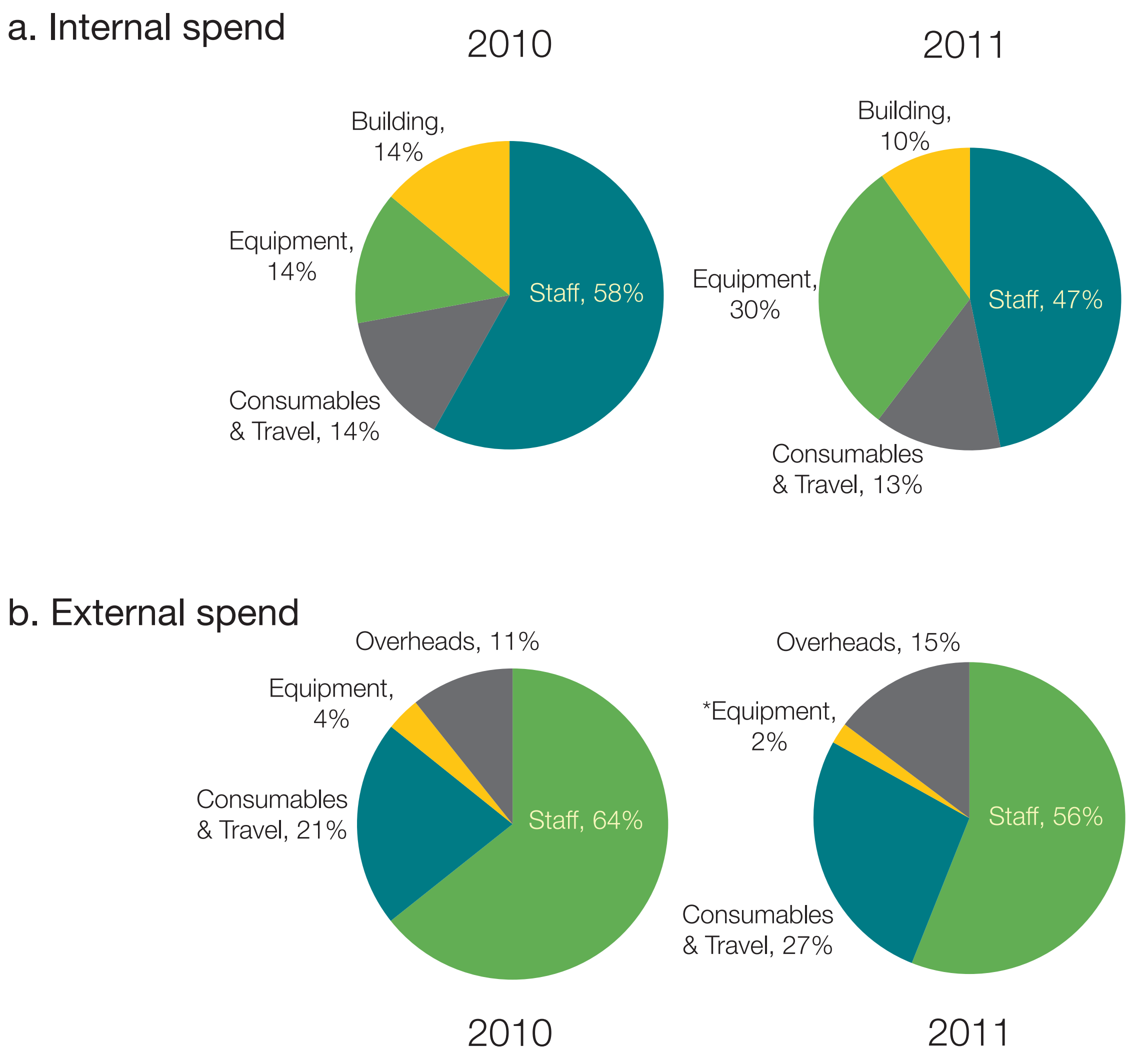


Figure 5. Breakdown of spend for 2010 and 2011. (a) Internal (including non-grant income) and (b) external spend. ‘Overheads’ in the external spend represent estate costs related to externally funded staff. *In 2010 we also spent €6 million on equipment from the UK Large Facilities Capital Fund, which is not reflected in this figure.

EMBL-EBI staff, March 2011.

STAFF

Our organisational structure reflects the different parts of our mission: services, research, training and industry support, with internal support facilitating all of these (see organogram, opposite page). EMBL-EBI personnel grew by 8.24% (Figure 6) from 461 in June 2010 to 499 in December 2011 (these figures exclude visitors).

As a European organisation we are proud to report that our personnel represent 48 countries (37 last year). In 2011 we welcomed 47 visitors who stayed with us for longer than a month (compare to 28 visitors last year). A new Team Leader was appointed within the PDBe (see page 46). We also established a new office of External Relations, which has two members of staff.

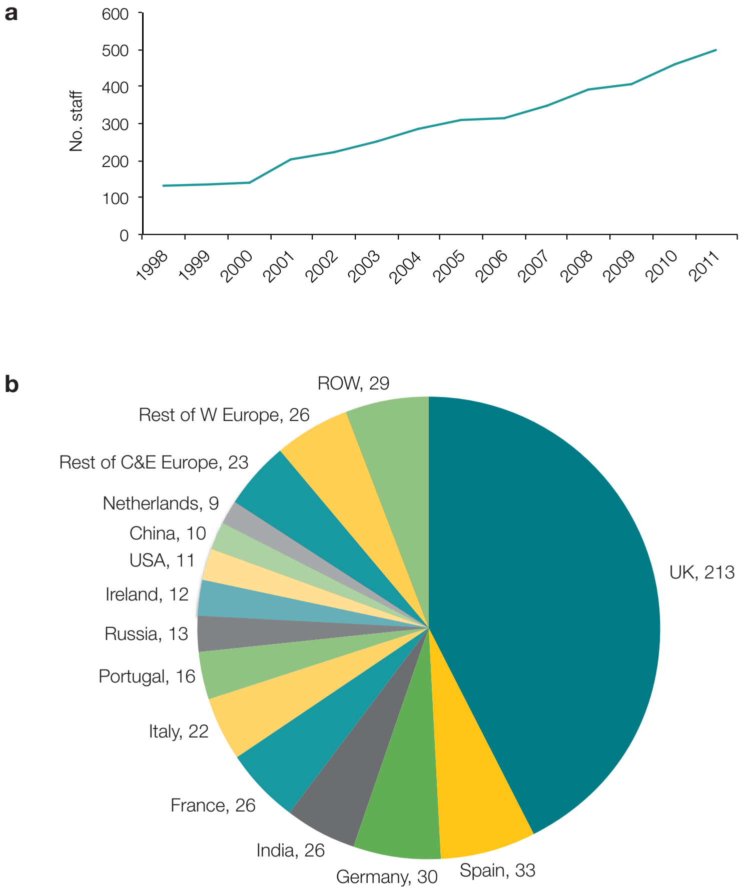
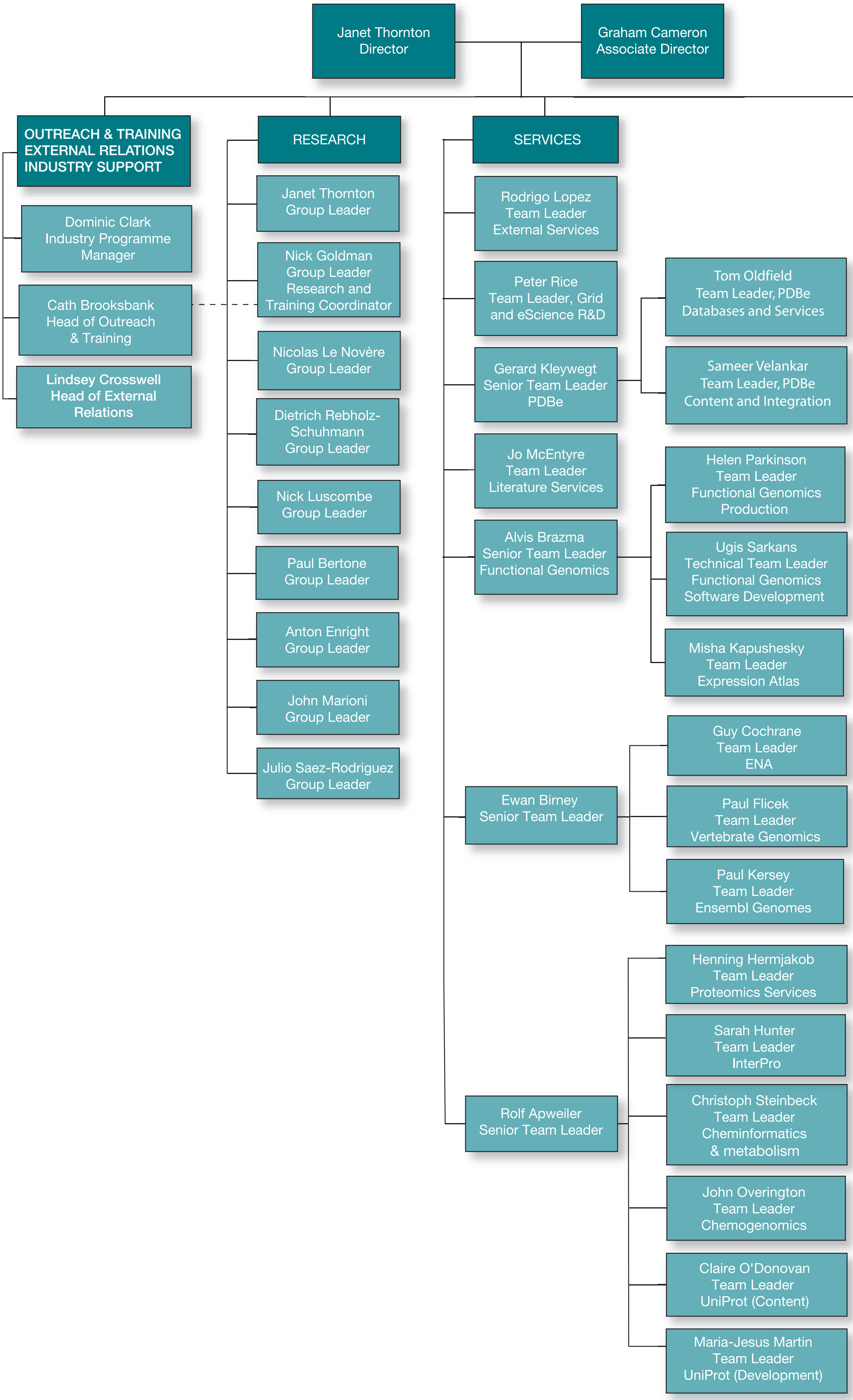


Figure 6. EMBL-EBI personnel. (a) Staff growth from 1998 to December 2011. (b) Nationalities of EMBL-EBI members of personnel as of December 2011.



- Janet Thornton
Director
- Graham Cameron
Associate Director

- OUTREACH & TRAINING EXTERNAL RELATIONS INDUSTRY SUPPORT
  - Dominic Clark, Industry Programme Manager
  - Cath Brooksbank, Head of Outreach & Training
  - **Lindsey Crosswell, Head of External Relations**
- RESEARCH
  - Janet Thornton, Group Leader
  - Nick Goldman, Group Leader, Research and Training Coordinator
  - Nicolas Le Novère, Group Leader
  - Dietrich Rebholz-Schuhmann, Group Leader
  - Nick Luscombe, Group Leader
  - Paul Bertone, Group Leader
  - Anton Enright, Group Leader
  - John Marioni, Group Leader
  - Julio Saez-Rodriguez, Group Leader
- SERVICES
  - Rodrigo Lopez, Team Leader, External Services
  - Peter Rice, Team Leader, Grid and eScience R&D
  - Gerard Kleywegt, Senior Team Leader, PDBe
    - Tom Oldfield, Team Leader, PDBe Databases and Services
    - Sameer Velankar, Team Leader, PDBe Content and Integration
  - Jo McEntyre, Team Leader, Literature Services
  - Alvis Brazma, Senior Team Leader, Functional Genomics
    - Helen Parkinson, Team Leader, Functional Genomics Production
    - Ugis Sarkans, Technical Team Leader, Functional Genomics Software Development
    - Misha Kapushesky, Team Leader, Expression Atlas
  - Ewan Birney, Senior Team Leader
    - Guy Cochrane, Team Leader, ENA
    - Paul Flicek, Team Leader, Vertebrate Genomics
    - Paul Kersey, Team Leader, Ensembl Genomes
  - Rolf Apweiler, Senior Team Leader
    - Henning Hermjakob, Team Leader, Proteomics Services
    - Sarah Hunter, Team Leader, InterPro
    - Christoph Steinbeck, Team Leader, Cheminformatics & metabolism
    - John Overington, Team Leader, Chemogenomics
    - Claire O'Donovan, Team Leader, UniProt (Content)
    - Maria-Jesus Martin, Team Leader, UniProt (Development)
- INTERNAL SUPPORT
  - Mark Green, Head of Admin
  - Petteri Jokinen, Head of Systems & Networking

Figure 7. Organisation of EMBL-EBI leadership.

Scientific Advisory Boards and Committees

EMBL Scientific Advisory Committee

Siv Andersson, Uppsala, Sweden
Naama Barkai, Rehovot, Israel
Konrad Basler, Zurich, Switzerland
Denis Duboule, Lausanne, Switzerland
Roderic Guigo, Barcelona, Spain
Reinhard Jahn, Göttingen, Germany (Vice Chair)
Daniel Louvard, Paris, France
Ron Milligan, La Jolla, USA
Tom Muir, Princeton, USA
Andrew Murray, Harvard, USA
Andrea Musacchio, Dortmund, Germany
Helen Saibil, London, United Kingdom
Sandra Schmid, La Jolla, United States (Chair)
Titia Sixma, Amsterdam, Netherlands
Michael Snyder, Stanford, USA
Alfonso Valencia, Madrid, Spain
Jean Weissenbach, Evry, France

EMBL-EBI Bioinformatics Advisory Committee

Roderic Guigo Centre de Regulació Genòmica, Barcelona, Spain
Tim Hubbard, Wellcome Trust Sanger Institute, Hinxton, UK
Olli Kallioniemi, VTT Medical Biotechnology, Turku, Finland
Jonathan Knowles, Roche, Switzerland (retired)
Philippe Sanseau, GlaxoSmithKline, UK
Anna Tramontano, University of Rome "La Sapienza" Rome, Italy (Chair)
Mathias Uhlén, Royal Institute of Technology (KTH), Stockholm, Sweden

BioModels Scientific Advisory Committee

Upinder Bhalla, National Centre for Biological Sciences, India
Michael Hucka, California Institute of Technology, USA
Pedro Mendes, Manchester Centre of Integrative Systems Biology, UK
Ion Moraru, University of Connecticut Health Center, USA
Herbert Sauro, Washington University, USA
Jacky Snoep, Stellenbosh University, South Africa (Chair)

Cheminformatics: ChEMBL and ChEBI Advisory Committee

Steve Bryant, NIH, USA
Edgar Jabcoby, Novartis, Basel, Switzerland
Andrew Leach, GlaxoSmithKline Plc, UK (Chair)
Tudor Oprea, University of New Mexico, Albuquerque, USA
Alfonso Valencia, CNIO, Madrid, Spain
Peter Willett, University of Sheffield, UK

European Nucleotide Archive Scientific Advisory Board

Mark Blaxter, University of Edinburgh
Antoine Danchin, CNRS, Institut Pasteur, Paris, France
Roderic Guigo, Centre de RegulacioÅL Geno`mica, Barcelona, Spain
Tim Hubbard, Wellcome Trust Sanger Institute, Hinxton, UK (Chair)
Jim Ostell, National Centre for Biotechnology Information, USA
Babis Savakis, University of Crete & IMBB-FORTH, Heraklion, Greece
Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin, Germany
Jean Weissenbach, Génoscope, Evry, France
Patrick Wincker, Génoscope, Evry, France

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

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

Ensembl Scientific Advisory Board

Stephan Beck, University College London, UK (Chair)
Deepak Singh, Amazon Web Services, Seattle, USA
Johan den Dunnen, Leiden University Medical Centre, the Netherlands
Michael Stratton, Wellcome Trust Sanger Institute, Hinxton, UK (Observer)
Hugues Roest, Crollius, DIOGEN, Paris, France
Toby Bloom, Broad Institute, Cambridge, USA
Robert Kunh, University of California, Santa Cruz
Cornelius Gross, EMBL Monterotondo, Italy

Ensembl Genomes Scientific Advisory Board

Detlev Arendt, EMBL, Heidelberg, Germany
Mike Bevan, John Innes Centre, Norwich, UK
Steve Oliver, University of Cambridge, UK

Julian Parkhill, Wellcome Trust Sanger Institute, UK (Chair)
Doreen Ware, Cold Spring Harbor Laboratory, USA
Jane Rogers, BBSRC Genome Analysis Centre
Chris Rawlings, Rothamsted Research

The Gene Ontology Scientific Advisory Board

Philip Bourne, University of California, San Diego, CA, USA
Richard Scheuermann, University of Texas Southwestern Meidcal Centre, Dallas, USA
Michael Schroeder, Technische Universität Dresden, Germany
Barry Smith, SUNY Buffalo, NY, USA
Olga Troyanskaya, Princeton University, Department of Computer Science and Molecular Biology, NJ, USA
Michael Tyers, Samuel Lunenfeld Research institue, Mt. Sinai Hospital, Toronto, Canada

InterPro/Pfam Scientific Advisory Board

Philip Bourne, University of California, San Diego, CA, USA
Michael Galperin, National Center for Biotechnology Information, Bethesda, MD, USA
Erik Sonnhammer, Stockholm University, Sweden (Chair)
Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain

Literature Services

Alex Bateman, Wellcome Trust Sanger Institute, Hinxton, UK (Chair)
Carolla Tilgmann, Orion Pharma, Espoo, Finland
Larry Hunter, University of Colorado Health Sciences Center, Aurora, CO USA
Mark Patterson, eLife, Cambridge, UK
Wolfram Horstmann, University of Bielefeld, Germany and University of Oxford, UK
Gianni Cesareni, University of Rome, Italy

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

Udo Heinemann, Max Delbrück Center for Molecular Medicine, Berlin, Germany
Tomas Lundqvist, AstraZeneca R&D, MoÅNlndal, Sweden
Andrea Mattevi, University of Pavia, Italy
Randy Read, University of Cambridge, UK (Chair)
Helen Saibil, Birkbeck College London, UK
Michael Sattler, TUM, Munich, Germany
Torsten Schwede, Swiss Institute of Bioinformatics, Switzerland
Titia Sixma, Netherlands Cancer Institute, Amsterdam, the Netherlands

Worldwide Protein Data Bank (wwPDB) Advisory Committee

Stephen K. Burley, Eli Lilly, USA (Chair)
Wayne Hendrickson, Columbia University, USA
Genji Kurisu, Institute for Protein Research, Osaka University, Japan
Gaetano Montelione, Rutgers University, USA
David Neuhaus, MRC Laboratory of Molecular Biology, Cambridge, UK
Randy J. Read, University of Cambridge, UK
Michael G. Rossmann, Purdue University, USA
Helen Saibil, Birkbeck College London, UK
Soichi Wakatsuki, High Energy Accelerator Research Organisation (KEK), Japan
Edward N. Baker, University of Auckland, NZ (Ex Officio)
R. Andrew Byrd, NIH, USA (Ex Officio)

EMDataBank Advisory Committee

Joachim Frank, Columbia University, USA (Chair)
Achilleas Frangakis, Goethe University Frankfurt, Germany
Richard Henderson, MRC Laboratory of Molecular Biology, Cambridge, UK
Maryanne Martone, University of California San Diego, USA
Michael Rossmann, Purdue University, USA
Andrej Sali, University of California, USA
Paula Flicker, National Institute of General Medical Sciences, USA (Observer)
Wah Chiu, Baylor College of Medicine, USA (Ex Officio)
Manju Bansal, Education Research Network, India (Associate Member)
Jianping Ding, Shanghai Institutes for Biological Sciences, China (Associate Member))

Reactome Scientific Advisory Committee

Julie Ahringer, University of Cambridge, UK
Russ Altman, Stanford University, CA, USA
Gary Bader, University of Toronto, Canada
Richard Belew, University of California, San Diego, CA, USA
Matt Day, Nature Publishing Group, London, UK
Edda Klipp, Max-Planck Institute for Molecular Genetics, Berlin, Germany
Adrian Krainer, Cold Spring Harbor Laboratory, NY, USA
Ed Marcotte, University of Texas at Austin, TX, USA
Mark McCarthy, University of Oxford, UK
Bill Pearson, University of Virginia, VA, USA
Pardis Sabeti, Broad Institute, USA
David Stewart, Cold Spring Harbor Laboratory, NY, USA

The Universal Protein Resource (UniProt) Scientific Advisory Committee

Michael Ashburner, University of Cambridge, UK
Patricia Babbitt, University of California San Francisco, USA
Helen Berman, Rutgers University NJ, USA
Judith Blake, The Jackson Laboratory, ME, USA
Ian Dix, AstraZeneca, Macclesfield, UK
Takashi Gojobori, National Institute of Genetics, Mishima, Japan
Maricel Kann, University of Maryland, Baltimore, USA
Bernhard Kuester, Technical University Munich, Weihenstephan, Germany
Edward Marcotte, University of Texas, Austin, USA
William Pearson, University of Virginia, Charlottesville, USA
David Searls (Freelancer)
Kanehisa Minoru, Institute for Chemical Research, Kyoto, Japan
Mathias Uhlén Royal Institute of Technology (KTH), Stockholm, Sweden (Chair)
Timothy Wells, Medicines for Malaria Venture, Geneva, Switzerland

PROJECTS

The following projects have their own Scientific Advisory Boards.

CALBC Scientific Advisory Board

Yves Lussier, University of Chicago and the UC Cancer Research Center, Chicago, USA
Scott Marshall, Centrum voor Wiskunde en Informatica, Amsterdam, the Netherlands
Therese Vachon, Novartis Pharmaceuticals, Basel, Switzerland
Alfonso Valencia, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

EMBOSS Scientific Advisory Board

David Bauer, Bayer Schering Pharma AG, Germany
Guy Bottu, UniversiteÅL Libre de Bruxelles, Belgium
Sarah Butcher, Imperial College Bioinformatics Support Service, UK
Sean Eddy, Janelia Farm, USA
Dirk Evers, Illumina (Computational Biology), UK
Andrew Lyall, EMBL-EBI (Observer)
Julian Parkhill, Wellcome Trust Sanger Institute, Hinxton, UK
Christopher Southan, Chris DS Consulting Ltd., UK (Observer)
John Walshaw, BBSRC John Innes Centre in Norwich, UK
Mathew Woodwark, Medimmune (a subsidiary of AstraZeneca), USA

Gen2Phen Scientific Advisory Board

Paul Burton, University of Leicester, UK
Lincoln Stein, Ontario Institute for Cancer Research, Canada
Jochen Taupitz, University of Mannheim, Germany

Vertebrate Bridging Ontology Scientific Advisory Board

Johnathan Bard, University of Oxford, UK (Chair)
Peter Holland, University of Oxford, UK
Martin Ringwald, The Jackson Laboratory, Maine, USA
Claudio Stern, University College London, UK
Monte Westerfield, Institute of Neuroscience, University of Oregon, USA

SYBARIS Scientific Advisory Board

Jane Kaye, University of Oxford, UK
Mihai G. Netea, Radboud University Nijmegen Medical Center, the Netherlands
Ken Smith, University of Cambridge, UK
Ioannis Xenarios, Swiss Institute of Bioinformatics, Switzerland

Major Database Collaborations

Most of our resources are collaborations with institutions throughout the world, and work with one another at the EBI towards a common goal of fully integrating publicly available molecular data. Here, we show representative collaborations for our major databases; it is not intended to be comprehensive, but rather to give a flavour for the global impact of our work.

ARRAYEXPRESS

- Dana Farber Cancer Institute, Boston, USA
- DDBJ Omics Archive, DNA Databank of Japan, Mishima, Japan
- Gene Expression Omnibus, NCBI, Bethesda, USA
- Functional Genomics Data Society
- Penn Center for Bioinformatics, University of Pennsylvania School of Medicine, Philadelphia, USA
- Stanford Microarray Database, Stanford University, USA

BIOMODELS DATABASE

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

ChEBI

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

ChEMBL

- BindingDB, University of California San Diego, USA
- CanSAR, Institute of Cancer Research, London, UK
- PubChem, NCBI, National Institutes of Health, Bethesda, USA

ENA - THE EUROPEAN NUCLEOTIDE ARCHIVE

The ENA is part of the International Nucleotide Sequence Database Collaboration (www.insdc.org). Other partners include:

- National Center for Biotechnology Information, Bethesda, USA (GenBank, Trace Archive and Sequence Read Archive)
- National Institute of Genetics, Mishima, Japan (DNA DataBank of Japan, Trace Archive and Sequence Read Archive)

Other ENA collaborations:

- Catalogue of Life (http://www.catalogueoflife.org/)
- Genomics Standards Consortium (http://gensc.org)

ENSEMBL

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

- Baylor College of Medicine, Houston, USA
- Broad Institute, Cambridge, USA
- DOE Joint Genome Institute, Walnut Creek, USA
- Ensembl at the Wellcome Trust Sanger Institute, Hinxton, UK
- Genome Browser at the University of California, Santa Cruz, USA
- Map Viewer at the National Center for Biotechnology Information, Bethesda, USA
- Mouse Genome Informatics at the Jackson Laboratory, Bar Harbor, USA
- Rat Genome Database at the Medical College of Wisconsin, Milwaukee, USA
- The Roslin Institute, Midlothian, Scotland, UK

ENSEMBL GENOMES

- Gramene at Cold Spring Harbor Laboratory, USA
- PomBase with University College London and the University of Cambridge, UK
- PhytoPath with Rothamsted Research, Harpenden, UK
- VectorBase: a collaboration with University of Notre Dame, USA; Harvard University, USA; Institute of Molecular Biology and Biochemistry, Greece; University of New Mexico, USA; and Imperial College London, UK
- Microme, a European collaboration with 14 partners
- transPLANT, a European project with 11 partners
- WormBase, a collaboration with the California Institute of Technology and Washington University, USA; Ontario Institute for Cancer Research, Canada; Wellcome Trust Sanger Institute and Oxford University, UK

THE GENE ONTOLOGY CONSORTIUM

- The Arabidopsis Information Resource, Carnegie Institution of Washington, Stanford, USA
- Berkeley Bioinformatics Open-source Project, Lawrence Berkeley National Laboratory, Berkeley, USA
- British Heart Foundation – University College London, London, UK
- DictyBase at Northwestern University, Chicago, USA
- EcoliWiki
- FlyBase at the University of Cambridge, UK
- GeneDB at the Wellcome Trust Sanger Institute, Hinxton, UK
- Gramene at Cornell University, Ithaca, USA
- Institute for Genome Sciences, University of Maryland, Baltimore, USA
- The J. Craig Venter Institute, Rockville, USA
- Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, USA
- Rat Genome Database at the Medical College of Wisconsin, Milwaukee, USA
- Saccharomyces Genome Database, Stanford University, Stanford, USA
- WormBase at California Institute of Technology, Pasadena, USA
- The Zebrafish Information Network at the University of Oregon, Eugene, USA

INTACT

IntAct is a member of the IMEX consortium.
Other partners include:

- Centro Nacional de Biotecnologia, Madrid, Spain
- DIP at the University of California, Los Angeles, USA
- MINT at University Tor Vergata, Rome, Italy
- MIPS at the National Research Centre for Environment and Health, Munich, Germany
- Neuroproteomics platform of National Neurosciences Facility, Melbourne, Australia
- Shanghai Institutes for Biological Sciences, Shanghai, China

INTERPRO

- CATH-Gene3D at University College London, UK
- HAMAP at the Swiss Institute of Bioinformatics, Geneva, Switzerland
- InterPro at EMBL-EBI, Hinxton, UK
- PANTHER at University of Southern California, Los Angeles, USA
- Pfam at the Wellcome Trust Sanger Institute, Hinxton, UK
- PIRSF at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, USA
- PRINTS at the University of Manchester, UK
- ProDom at INRA and CNRS, Toulouse, France
- PROSITE at the Swiss Institute of Bioinformatics, Geneva, Switzerland
- SCOP at the Laboratory of Molecular Biology, University of Cambridge, UK
- SMART at EMBL, Heidelberg, Germany
- SUPERFAMILY at the University of Bristol, UK
- TIGRFAMs at The Institute of Genome Research, Rockville, USA

THE PROTEIN DATABANK IN EUROPE

PDBe is a partner in the World Wide Protein Data Bank (wwPDB). Other partners include:

- BioMagResBank, University of Wisconsin, Madison, USA
- PDBj at Osaka University, Japan
- Research Collaboratory for Structural Bioinformatics, USA

PRIDE

- Faculty of Life Sciences, The University of Manchester, UK
- Ghent University, Ghent, Belgium
- The Yonsei Proteome Research Center, Yonsei University, Seoul, Korea.

UK PubMed Central

UK PubMed Central is part of PubMed Central International.
Other partners include:

- The British Library UK
- The University of Manchester: Mimas and Life Sciences, UK
- The National Centre for Text Mining, UK

REACTOME

- New York University Medical Center, USA
- Ontario Institute for Cancer Research, Toronto, Ontario, Canada
- Reactome at Cold Spring Harbor Laboratory, USA

THE UNIFIED PROTEIN RESOURCE - UNIPROT

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

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

External Seminar Speakers 2011

EMBL-EBI hosts a weekly seminar series that welcomes speakers from other institutions to share their perspectives on a wide range of biological questions, ranging from the fundamental mechanisms of development and metabolism to disease diagnostics and treatment. Speakers are invited by EMBL-EBI Senior Scientists, and their talks address both computational and wet-lab issues in modern biology. We list here the speakers who have kindly presented in this series.

This list is meant to give an impression of the talks that are available to our staff on a regular basis. A very large number of internal and external seminars are on offer an the Wellcome Trust Genome Campus, organised by various groups throughout the institute. They are simply too numerous to include here.

January

Claus Jørgensen, Institute of Cancer Research, London, UK
Network analysis of cell-specific Eph/ephrin bidirectional signalling in co-culture

Jerven Bolleman, Swiss Institute of Bioinformatics, Geneva, Switzerland
UniProt.rdf: what and why?

February

Marta Sanchez-Carbayo, Memorial Sloan-Kettering Cancer Center, New York, USA
Linking -omics in bladder cancer

Vincent Danos, University of Edinburgh, UK
Collective variables in biomolecular networks

Bronwyn van der Merwe and Andy Greenham, BBC
Developing a global experience language for the BBC's digital services

Tom Scott, BBC
Using linked data to describe the natural world

Sach Mukherjee,University of Warwick, UK
Network models for cancer signalling

March

Anna Divoli, Pingar, London, UK
Human factors in computational biology – from mathematical models to user interfaces

Neil Swainston, University of Manchester, UK
Exploiting semantics in metabolic systems biology

Tom Kirkwood, Newcastle University, UK
Systems models of ageing

Jens Stoye, Bielefeld University, Germany
Advanced techniques for comparative genome finishing

Jessica Mar, Harvard School of Public Health, Boston, USA
Modelling cell fate transitions and a variance-based approach to studying human disease

April

Thomas Illig, Helmholtz Zentrum, Munich, Germany
Genomics meets metabolomics

Francisco Couto, Universidade de Lisboa, Portugal
Exploring the semantics of biomedical ontologies

Masanori Arita, University of Tokyo, RIKEN Plant Science Center, Japan
Knowledge management using a Wiki-based system and its application to mass spectra

May

Cornelius Gross, EMBL Monterotondo, Italy
Pink Seminar*: Developmental remodelling of brain circuits by microglia

Matthew Bellgrad, Centre for Comparative Genomics Western Australia
A sophisticated Internet-based cross –omics analytic environment

Frank Oliver, University of Manchester, UK
Exploiting semantics in metabolic systems biology

June

Eileen Furlong, EMBL Heidelberg, Germany
Pink Seminar*: Linking transcription factor occupancy and chromatin state to gene expression during embryonic development

July

Katja Barenfaller, ETH Zurich, Switzerland
Integrating proteomics data: pep2pro, MASCP Gator and combined

Mike Atherton, Huddle, London, UK
Beyond the polar bear

Christian von Mering, EMBL Heidelberg, Germany
The STRING protein–protein interaction resource: latest developments, usage scenarios and future plans

Rob Klose, University of Oxford, UK
Interpreting the CpG island signal

September

Detlev Arendt, EMBL Heidelberg, Germany
Pink Seminar*: Duplication and divergence of neural circuits in central nervous system evolution

Nicola Osborne, EDINA, JISC National Data Centre, University of Edinburgh, UK
Science and social media

Mihaela Zavolan, Swiss Institute of Bioinformatics, Basel, Switzerland
Large-scale approaches to miRNA target identification

Claes Wadelius, Uppsala University, Sweden
Genetic and epigenetic control of transcripts in HepG2 cells

October

Maja Koehn, EMBL Heidelberg, Germany
Pink Seminar*: Investigation of phosphatases using chemical biology tools

Bernhard Knapp, Medical University of Vienna, Austria
Molecular dynamics simulations in the TCR:peptide:MHC interface

Rolf Mueller, Helmholtz Institute for Pharmaceutical Research, Saarland, Germany
Genomics, mass spectrometry, bioinformatics and biotechnology for microbial natural product discovery

November

Kiran Raosaheb Patil, EMBL Heidelberg, Germany
Pink Seminar*: From gene expression to metabolic phenotype

Des Traynor, Intercom, Dublin, Ireland
The Language of Software

Jean-Loup Faulon, ISSB, Genopole, Paris, France
Retrosynthetic biology

December

Albert Goldbeter, Universite Libre de Bruxelles (ULB), Belgium
Oscillatory dynamics of the cyclin/Cdk network driving the mammalian cell cycle

Giles Colborne, CX Partners, Bristol, UK
Secrets of Simplicity

*Pink Seminars are part of an internal EMBL series in which senior scientists discuss their current research.

Publications

RESEARCH GROUPS

BERTONE GROUP

Albers, C.A., Cvejic, A., *et al.* (2011) Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat. Genet.* 43 (8), 735-7.

Reynolds, N., Salmon-Divon, M., *et al.* (2011) NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression. *EMBO J.* 31, 593-605.

ENRIGHT GROUP

Bateman, A., Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

De Fazio, S., Bartonicek, N., *et al.* (2011) The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* 480, 259–263.

Hu, M., Ayub, Q., *et al.* (2011) Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. *Hum. Genet.* Article in press.

Keane, T.M., Goodstadt, L., *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477 (7364), 289-94.

Mayoral, R.J., Deho, L., *et al.* (2011) MiR-221 Influences Effector Functions and Actin Cytoskeleton in Mast Cells. *PLoS One* 6 (10), e26133.

Puri, V., Goyal, A., *et al.* (2011) Evolutionary and functional insights into Leishmania META1: evidence for lateral gene transfer and a role for META1 in secretion. *BMC Evol. Biol.* 11, 334.

Tripathi, R., Saini, H.K., *et al.* (2011) Messenger RNA and MicroRNA Profiling during Early Mouse EB Formation. *Gene Expr. Patterns* 11 (5-6), 334-44.

GOLDMAN GROUP

Albers, C.A., Cvejic, A., *et al.* (2011) Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat. Genet.* 43 (8), 735-7.

Blanquart, S., and Gascuel, O. (2011) Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes. *BMC Evol. Biol.* 11, 70.

Hess, J., and Goldman, N. (2011) Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. *PLoS One* 6 (8), e22783.

Jordan, G. (2011) Analysis of alignment error and sitewise constraint in mammalian comparative genomics. *PhD thesis, EMBL and University of Cambridge.*

Jordan, G. and Goldman, N. (2011) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* Article in press.

Kosiol, C., and Goldman, N. (2011) Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. *J. Mol. Biol.* 411 (4), 910-23.

Lakner, C., Holder, M.T., *et al.* (2011) What's in a likelihood? simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. *Syst. Biol.* 60 (2), 161-74.

Lindblad-Toh, K., Garber, M., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

Sipos, B., Massingham, T., *et al.* (2011) PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 12, 104.

Washietl, S., Findeiss, S., *et al.* (2011) RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17 (4), 578-94.

LE NOVERE GROUP

Changeux, J.-P. and Edelstein, S.J. (2011) Conformational selection or induced fit? 50 years of debate resolved. F1000 *Biol. Rep.* 3, 19.

Courtot, M., Juty, N., *et al.* (2011) Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 7, 543.

Dräger, A., Rodriguez, N., *et al.* (2011) JSBML: a flexible Java library for working with SBML. *Bioinformatics* 27 (15), 2167-8.

Juty, N., Le Novère, N., Laibe C. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.* 40, D580-D586

Le Novère, N., Hucka, M., *et al.* (2011) Meeting report from the first meetings of the Computational Modelling in Biology Network (COMBINE). *SIGS* 5 (2), 577.

Schulz, M., Krause, F., *et al.* (2011) Retrieval, alignment, and clustering of computational models based on semantic annotations. *Mol. Sys. Biol.* 7, 512.

Stefan, M., Marshall, D., Le Novère N. (2012) Structural analysis and stochastic modelling suggest a mechanism for calmodulin trapping by CaMKII. *PLoS ONE* 7 (1), e29406.

Waltemath, D., Adams, R., *et al.* (2011) Minimum information about a simulation experiment (MIASE). *PLoS Comp. Biol.* 7 (4), Art. No. e1001122.

Waltemath, D., Adams, R., *et al.* (2011) Reproducible computational biology experiments with SED-ML -- The Simulation Experiment Description Markup Language. *BMC Sys. Biol.* 5, 198.

LUSCOMBE GROUP

Cavalli, F.M., Bourgon, R., *et al.* (2011) SpeCond: a method to detect condition-specific gene expression. *Genome Biol.* 12 (10), R101.

Ermakova, O., Piszczek, L., *et al.* (2011) Sensitized phenotypic screening identifies gene dosage sensitive region on chromosome 11 that predisposes to disease in mice. *EMBO Mol. Med.* 3 (1), 50-66.

Kahramanoglou, C., Seshasayee, A.S., *et al.* (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucleic Acids Res.* 39 (6), 2073-91.

Koepke, J., Kaffarnik, F., *et al.* (2011) The RNA-binding protein Rrm4 is essential for efficient secretion of endochitinase Cts1. *Mol. Cell. Proteomics* 10 (12), M111.011213.

Konig, J., Zarnack, K., *et al.* (2011) ICLIP - transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Visualized Exp.*(50), 1.

Prieto, A.I., Kahramanoglou, C., *et al.* (2011) Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in *Escherichia coli* K12. *Nucleic Acids Res.* Article in press.

Seshasayee, A.S., and Luscombe, N.M. (2011) Comparative genomics suggests differential deployment of linear and branched signaling across bacteria. *Mol. Biosyst* 7 (11), 3042-9.

Seshasayee, A.S., Sivaraman, K. and Luscombe, N.M. (2011) An overview of prokaryotic transcription factors: a summary of function and occurrence in bacterial genomes. *Subcell. Biochem.* 52, 7-23.

Vaquerizas, J.M., Akhtar, A. and Luscombe, N.M. (2011) Large-scale nuclear architecture and transcriptional control. *Subcell. Biochem.* 52, 279-95.

Vaquerizas, J.M., Teichmann, S.A. and LUSCOMBE, N.M. (2012) How do you find transcription factors? Computational approaches to compile and annotate repertoires of regulators for any genome. *Methods Mol. Biol.* 786, 3-19.

Zaugg, J.B. and Luscombe, N.M. (2012) A genomic model of condition-specific nucleosome behaviour explains transcriptional activity in yeast. *Genome Res.* 22 (1), 87-94.

MARIONI GROUP

Barreiro, L.B. ,Tailleux, L., *et al.* (2012) Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* 109 (4), 1204-9.

Cain, C.E., Blekhman, R., *et al.* (2011) Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* 187 (4), 1225-34.

Goh, X.Y., Rees, J.R., *et al.* (2011) Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut* 60 (10), 1317-26.

Pai, A.A., Bell, J.T., *et al.* (2011) A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.* 7 (2), e1001316.

Perry G.H., Marioni J.C, *et al.* (2010) Genomic-scale capture and sequencing of endogenous DNA from feces. *Mol. Ecol.* 19, 5332-44.

REBHOLZ-SCHUHMANN GROUP

Hoehndorf, R., Dumontier, M., *et al.* (2011) Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS One* 6 (7), e22006.

Hoehndorf, R., Dumontier, M., *et al.* (2011) A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics* 27 (7), 1001-8.

Hoehndorf, R., Ngonga Ngomo, A.C., *et al.* (2011) Ontology design patterns to disambiguate relations between genes and gene products in GENIA. *J. Biomed. Semantics* 2 (Suppl 5), S1.

Kim, J.J. and Rebholz-Schuhmann, D. (2011) Improving the extraction of complex regulatory events from scientific text by using ontology-based inference. *J. Biomed. Semantics* 2 (Suppl 5), S3.

McEntyre, J.R., Ananiadou, S., *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.* 39 (Database issue), D58-65.

Rebholz-Schuhmann, D., Rinaldi, F., *et al.* (2011) Towards mature use of semantic resources for biomedical analyses. *J. Biomed. Semantics* 2 (Suppl 5), I1.

Rebholz-Schuhmann, D., Yepes, A.J., *et al.* (2011) Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J. Biomed. Semantics* 2 (Suppl 5), S11.

Senay, K., Varoğlu, E., *et al.* (2011) Diversity in the Interactions of Isoforms Linked to Clustered Transcripts: A Systematic Literature Analysis. *J. Proteom. Bioinfo*. (4.11,S), 250-59.

Thompson, P., McNaught, J., *et al.* (2011) The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 12, 397.

SAEZ-RODRIGUEZ GROUP

Morris, M.K., Saez-Rodriguez, J., *et al.* (2011) Training signaling pathway maps to biochemical data with constrained fuzzy logic: Quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Comp. Biol.* 7 (3), Art. No. e1001099.

Prill, R.J., Saez-Rodriguez, J., *et al.* (2011) Crowdsourcing network inference: The DREAM predictive signaling network challenge. *Sci. Signal.* 4 (189), mr7.

Saez-Rodriguez, J., Alexopoulos, L.G. and Stolovitzky, G. (2011) Setting the standards for signal transduction research. *Sci. Signal.* 4 (160), pe10.

Saez-Rodriguez, J., Alexopoulos, L.G., *et al.* (2011) Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.* 71 (16), 5400-11.

THORNTON GROUP

Alic, N., Andrews, T.D., *et al.* (2011) Genome-wide dFOXO targets and topology of the transcriptomic response to stress and insulin signalling. *Mol. Syst. Biol.* 7, 502.

Ballester, P.J. (2011) Ultrafast shape recognition: *Future Medicinal Chemisty* 3 (1), 65-78.

Ballester, P.J. and Mitchell, J.B. (2011) Comments on "leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets": significance for the validation of scoring functions. *J. Chem. Inf. Model.* 51 (8), 1739-41.

Cuff, A.L., Sillitoe, I., *et al.* (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.* 39 (Database Supplement), D420-6.

Fischer, J.D. (2011) Characterisation, classification and conformational variability of organic enzyme cofactors. *PhD Thesis, EMBL and University of Cambridge.*

Furnham, N., Sillitoe, I., *et al.* (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.* 40 (Database issue), D776-82.

Hilton J M., Lewis M.A., *et al.* (2011). Exome sequencing identifies a missense mutation in Isl1 associated with low penetrance otitis media in dearisch mice. *Genome Biol.*, 12(9), R90.

Holliday, G.L., Andreini, C., *et al.* (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40 (Database issue), D783-9.

Holliday, G.L., Fischer, J.D., *et al.* (2011) Characterising the complexity of enzymes based on their mechanisms and structures using a bio-computational analysis. *FEBS J.* 278 (20), 3835-45.

Laskowski, R.A. and Swindells, M.B. (2011). LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* 51, 2778-86.

Laskowski, R.A. (2011) Protein Structure Databases. *Mol. Biotechnol.* 58 (2), 183-98.

Lee, D., de Beer, T.A.P., *et al.* (2011) 1,000 structures and more from the MCSG. *BMC Structural Biology*, 11:2.

McParland, V., Varsano, G., *et al.* (2011) The Metastasis-Promoting Phosphatase PRL-3 Shows Activity toward Phosphoinositides. *Biochemistry* 50, 7579-90.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

Partridge, L., Thornton, J.M. and Bates, G. (2011) The new science of ageing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 366 (1561), 6-8.

Thornton, J.M. (2011) Celebrating structural biology:Abstracting knowledge from the PDB over 40 years. *Nat. Struct. Mol. Biol.* 18 (12), 1304-16.

Wieser, D., Papatheodorou, I., *et al.* (2011) Computational biology for ageing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 366 (1561), 51-63.

SERVICE TEAMS

APWEILER: PROTEIN DATA

Abrahams, J., Apweiler, R., *et al.* (2011) "4D Biology for health and disease" workshop report. *Nat. Biotechnol.* 28 (4), 291-3.

Alam-Faruque, Y,. Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

Deng, N., Zhang, J., *et al.* (2011) Phosphoproteome analysis reveals regulatory sites in major pathways of cardiac mitochondria. *Mol. Cell. Proteomics* 10 (2), M110.000117.

Foster, J.M., Degroeve, S., *et al.* (2011) A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* 11 (11), 2182-94.

Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB 'complete proteome' sets. *Proteomics* 11 (22), 4434-8.

O'Donovan, C. and Apweiler, R. (2011) A guide to UniProt for protein scientists. *Methods Mol. Biol.* 694 25-35.

BIRNEY: NUCLEOTIDE DATA

Bateman, A., Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

Birney, E. (2011) Assemblies: the good, the bad, the ugly. *Nat. Methods* 8 (1), 59-60.

Birney, E. (2011) Chromatin and heritability: how epigenetic studies can complement genetic approaches. *Trends Genet.* 27 (5), 172-6.

Boyle, A.P., Song, L., *et al.* (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* 21 (3), 456-64.

Croft, D., O'Kelly, G., *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* Volume 39 (Database supplement), D691-7.

Flicek, P., Amode, M.R., *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.* 39 (Database issue), D800-6.

Fritz, M.H., Leinonen, R., *et al.* (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40.

Keane, T.M., Goodstadt, L., *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477 (7364), 289-94.

Kahlem, P. The European Network of Excellence ENFIN: Enabling systems biology. *Curr. Opin. Biotechnol.* (2011) 22 27.

Kahlem, P., *et al.* (2011) Strengths and Weaknesses of Selected Modeling Methods Used in Systems Biology. In: Ning-Sun Yang, Ed, *Bioinform. Comp. Modeling.* InTech, pp. 77-98.

Leinonen, R., Akhtar, R., et al. (2011) The European Nucleotide Archive. Nucleic Acids Res. 39 (Database issue), D28-31.

Lindblad-Toh, K., Garber, M., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9 (4), e1001046.

Seal, R.L., Gordon, S.M., *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.* 39 (Databse issue), D514-9.

Song, L., Zhang, Z., *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21 (10), 1757-67.

Vilella, A.J., Birney, E., *et al.* (2011) Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biol.* 12 (2), Art. No. 401.

Wright, M.W. and Bruford, E.A. (2011) Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics* 5 (2), 90-8.

BRAZMA: FUNCTIONAL GENOMICS

Beisvåg, V., Kauffmann, A., *et al.* (2011) Contributions of the EMERALD project to assessing and improving microarray data quality. *BioTechniques* 50 (1), 27-31.

Caldas, J., Gehlenborg, N., *et al.* (2012) Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics* 28 (2), 246-53.

Goncalves, A., Tikhonov, A., *et al.* (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 27 (6), 867-9.

Gostev, M., Faulconbridge, A., *et al.* (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.* 40 (Database issue), D64-70.

Gostev, M., Fernandez-Banet, J., *et al.* (2011) SAIL - a software system for sample and phenotype availability across biobanks and cohorts. *Bioinformatics* 27 (4), 589-91.

Kapushesky, M., Adamusiak, T., *et al.* (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40 (Database issue), D1077-81.

Kutter, C., Brown, G.D., *et al.* (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43 (10), 948-55.

ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9 (4), e1001046.

Nicholson, G., Rantalainen, M., *et al.* (2011) A Genome-Wide Metabolic QTL Analysis in Europeans Implicates Two Loci Shaped by Recent Positive Selection. *PLoS Genet.* 7 (9), e1002270.

Parkinson, H., Sarkans, U., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39 (Database issue), D1002-4.

COCHRANE: ENA

Amid, C. Birney, E., *et al.* (2012) Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res.* 40 (Database issue), D43-7.

Bateman, A. Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

Cochrane G., Karsch-Mizrachi I. and Nakamura Y. (2011) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 39 (Database issue), D15-18.

Field D., Amaral-Zettler L., *et al.* (2011) The Genomics Standards Consortium. *PLoS Biol* 9: e1001088

Fritz, M.H. Leinonen, R., *et al.* (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40.

Galperin M.Y. and Cochrane G. (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.* 39 (Database issue), D1-6.

Gaudet P., Bairoch A., *et al.* (2011) Towards BioDBcore: a community-defined information specfication for biological databases. *Nucleic Acids Res.* 39 (Database issue), D7-10.

Gaudet P., Bairoch A., *et al.* (2011) Towards BioDBcore: a community-defined information specfication for biological databases. *Database* 39 (Database issue), D7-10.

Karsch-Mizrachi, I. Nakamura, Y., *et al.* (2012) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40 (Database issue), D33-7.

Kodama, Y. Shumway, M., *et al.* (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40 (Database issue), D54-6.

Leinonen, R. Akhtar, R., *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.* 39 (Database issue), D28-31.

Leinonen R., Sugawara H. and Shumway M. (2011) The Sequence Read Archive. *Nucleic Acids Res.* 39 (Database issue), D19-21.

Yilmaz P., Gilbert J.A., *et al.* (2011) The Genomic Standards Consortium: bringing standards to life for microbial ecology. *ISME J* 5: 1565-1567

FLICEK: VERTEBRATE GENOMICS

Bateman, A., Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

Boyle, A.P., Song, L., *et al.* (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* 21 (3), 456-64.

Church, D.M., Schneider, V.A., *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.* 9 (7), e1001091.

Conrad, D.F., Keebler, J.E., *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43 (7), 712-4.

Danecek, P., Auton, A., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27 (15), 2156-8.

ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9 (4), e1001046.

Faure, A.J., Seoighe, C. and Mulder, N.J. (2011) Investigating the effect of paralogs on microarray gene-set analysis. *BMC Bioinformatics* 12 Art. No.: 29.

Flicek, P. Amode, M.R., *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.* 39 (Database issue), D800-6.

Gaudet, P., Bairoch, A., *et al.* (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Database (Oxford)* 2011 baq027.

Gaudet, P., Bairoch, A., *et al.* (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.* 39 (Database issue), D7-D10.

Gravel, S., Henn, B.M., *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* 108 (29), 11983-8.

Guberman, J.M., Ai, J., *et al.* (2011) BioMart Central Portal: an open database network for the biological community. *Database (Oxford)* 2011 bar041.

Kinsella, R.J., Kahari, A., *et al.* (2011) Ensembl BioMarts: a hub for data retrievaKohonen-Corish, M.R., Macrae, F., *et al.* (2011) Deciphering the colon cancer genes-report of the InSiGHT-Human Variome Project Workshop, UNESCO, Paris 2010. *Hum. Mutat.* 32 (4), 491-4.

Lindblad-Toh, K., Garber, M., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478 (7370), 476-82.

Lipman, D., Flicek, P., *et al.* (2011) Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biol.* 12 (3), 402.

Locke, D.P., Hillier, L.W., *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469 (7331), 529-33.

Malone, C.M.P., Domaschenz, R., *et al.* (2011) Hes6 is required for actin cytoskeletal organization in differentiating C2C12 myoblasts. *Exp. Cell Res.* 317 (11), 1590-602.

Marth, G.T., Yu, F., *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.* 12 (9), R84.

Mills, R.E., Walter, K., *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470 (7332), 59-65.

Oakley, D.J., Iyer, V., *et al.* (2011) BioMart as an integration solution for the International Knockout Mouse Consortium. *Database (Oxford)* 2011 bar028.

Pignatelli, M., and Moya, A. (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One* 6 (5), Art. No.: e19984.

Reimundo, P., Pignatelli, M., *et al.* (2011) Genome sequence of Lactococcus garvieae UNIUD074, isolated in Italy from a Lactococcosis outbreak. *J. Bacteriol.* 193 (14), 3684-5.

Renfree, M.B., Papenfuss, A.T., *et al.* (2011) Genome sequence of an Australian kangaroo, Macropus eugenii, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.* 12 (8), R81.

Ringwald, M., Iyer, V., *et al.* (2011) The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. *Nucleic Acids Res.* 39 (Database issue), D849-55.

Smedley, D., Salimova, E. and Rosenthal, N. (2011) Cre recombinase resources for conditional mouse mutagenesis. *Methods* 53, 411-6.

Song, L., Zhang, Z., *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21 (10), 1757-67.

Spudich, G.M. and Fernandez-Suarez, X.M. (2011) Disease and phenotype data at ensembl. *Curr. Protoc. Hum. Genet.* Chapter 6 Unit6.11.

Vilella, A.J., Birney, E., *et al.* (2011) Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biol.* 12 (2), Art. No.: 401.

HERMJAKOB: PROTEOMICS SERVICES

Aranda, B., Blankenburg, H., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8 (7), 528-9.

Croft, D., O'Kelly, G., *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* Volume 39 (Database supplement), D691-7.

Gel Moreno, B., Jenkinson, A.M., *et al.* (2011) EasyDAS: Automatic creation of DAS servers. *BMC Bioinform.* 12 Art. No.: 23.

Gieger, C., Radhakrishnan, A., *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature* 480, 201–208.

Griss, J., Cote, R.G., *et al.* (2011) Published and Perished? The influence of the searched protein database on the long-term storage of proteomics data. *Mol. Cell. Proteomics* 10 (9), M111.008490.

Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB 'complete proteome' sets. *Proteomics* 11 (22), 4434-8.

Haw, R., Hermjakob, H., *et al.* (2011) Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* 11 (18), 3598-613.

Haw, R.A., Croft, D., *et al.* (2011) The Reactome BioMart. *Database (Oxford)* 2011, bar031.

Kerrien, S., Aranda, B., *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40 (Database issue), D841-6.

Kinsinger, C.R., Apffel, J., *et al.* (2011) Recommendations for Mass Spectrometry data quality metrics for open access data (corollary to the Amsterdam Principles). *Mol. Cell. Proteomics* 10 (12), O111.015446.

Martens, L., Chambers, M., *et al.* (2011) mzML - A community standard for mass spectrometry data. *Mol. Cell. Proteomics* 10 (1), R110.000133.

Ndegwa, N., Cote, R.G., *et al.* (2011) Critical amino acid residues in proteins: a BioMart integration of Reactome protein annotations with PRIDE mass spectrometry data and COSMIC somatic mutations. *Database (Oxford)* 2011, bar047.

Orchard, S., Albar, J.P., *et al.* (2011) Enabling BioSharing - a report on the Annual Spring Workshop of the HUPO-PSI April 11-13, 2011, EMBL-Heidelberg, Germany. *Proteomics* 11 (22), 4284-90.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

Orchard, S. and Hermjakob, H. (2011) Data standardization by the HUPO-PSI: how has the community benefitted? *Methods Mol. Biol.* 696, 149-60.

Orchard, S. and Hermjakob, H. (2011) Preparing molecular interaction data for publication. *Methods Mol. Biol.* 694 229-36.

Salazar, G.A., Jimenez, R.C., *et al.* (2011) DAS writeback: a collaborative annotation system. *BMC Bioinformatics* 12, 143.

Villaveces, J.M., Jimenez, R.C., *et al.* (2011) Dasty3, a WEB Framework for DAS. *Bioinformatics* 27 (18), 2616-7.

HUNTER: INTERPRO

Hunter, C., *et al.* (2011) The EBI metagenomics archive, integration and analysis resource. In: Frans J. de Bruijn, Ed, *Handbook of Molecular Microbial Ecology I: Metagenomics and complementary approaches.* Wiley-Blackwell.

Hunter, S., Jones, P., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40 (Database issue), D306-12.

Jones, P., Binns, D., *et al.* (2011) The InterPro BioMart: federated query and web service access to the InterPro Resource. *Database (Oxford)* 2011, bar033.

McDowall, J. and Hunter, S. (2011) InterPro Protein Classification. *Methods Mol. Biol.* 694, 37-47.

KAPUSHESKY: EXPRESSION ATLAS

Adamusiak, T., Burdett, T., *et al.* (2011) OntoCAT - simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics* 12, 218.

Culhane, A.C., Schroder, M.S., *et al.* (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.* 40 (Database issue), D1060-6.

Goncalves, A., Tikhonov, A., *et al.* (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 27 (6), 867-9.

Kapushesky, M., Adamusiak, T., *et al.* (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40 (Database issue), D1077-81.

Kurbatova, N., Adamusiak, T., *et al.* (2011) ontoCAT: an R package for ontology traversal and search. *Bioinformatics* 27 (17), 2468-70.

Santamaria, R., Rizzetto, L., *et al.* (2011) Systems biology of infectious diseases: a focus on fungal infections. *Immunobiology* 11 (1212), 1227.

KERSEY: NON-VERTEBRATE GENOMICS

Bateman, A., Agrawal, S., *et al.* (2011) RNAcentral: A vision for an international database of RNA sequences. *RNA* 17 (11), 1941-6.

Earl, D., Bradnam, K., *et al.* (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res.* 21 (12), 2224-41.

Gan, X., Stegle, O., *et al.* (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* 477 (7365), 419-23.

Guberman, J.M., Ai, J., *et al.* (2011) BioMart Central Portal: an open database network for the biological community. *Database (Oxford)* 2011 bar041.

Kersey, P.J., Staines, D.M., *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.* 40 (Database issue), D91-7.

Kinsella, R.J., Kahari, A., *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* 2011, bar030.

Megy, K., Emrich, S.J., *et al.* (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.* 40 (Database issue), D729-34.

Robinson, G.E., Hackett, K.J., *et al.* (2011) Creating a buzz about insect genomes. *Science* 331 (6023), 1386.

Wood, V., Harris, M.A., *et al.* (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res.* 40 (Database issue), D695-9.

Yook, K., Harris, T.W., *et al.* (2011) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res.*, Published online 8 November. doi: 10.1093/nar/gkr954

Youens-Clark, K., Buckler, E., *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* 39 (Database issue), D1085-94.

KLEYWEGT: PDBE

Aranda, B., Blankenburg, H., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8 (7), 528-9.

Bernard, A., Vranken, W.F., *et al.* (2011) Bayesian estimation of NMR restraint potential and weight: A validation on a representative set of protein structures. *Proteins: Structure, Function, and Bioinformatics* 79 (5), 1525-37.

Doreleijers, J.F., Vranken, W.F., *et al.* (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res.* 40 (D1), D519-24.

Henderson, R., Sali, A. *et al.* (2012) Outcome of the First Electron Microscopy Validation Task Force Meeting. *Structure* 20 (2), 205-14.

Krissinel, E. (2011) Macromolecular complexes in crystals and solutions. *Acta Crystallogr. D Biol. Crystallogr.* 67 (Pt 4), 376-85.

Lawson, C.L., Baker, M.L., *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39 (Database issue), D456-64.

Lemak, A., Gutmanas, A., *et al.* (2011) A novel strategy for NMR resonance assignment and protein structure determination. *J. Biomolecular NMR* 49 (1), 27-38.

Ludtke, S.J., Lawson, C.L., *et al.* (2011) Workshop on the validation and modeling of electron cryo-microscopy structures of biological nanomachines: Workshop introduction. *Pac. Symp. Biocomput.* 369-73.

Morris, C., Pajon, A., *et al.* (2011) The Protein Information Management System (PiMS): a generic tool for any structural biology research laboratory. *Acta Crystallogr. D Biol. Crystallogr.* 67 (Pt 4), 249-60.

Pei, X-Y., Hinchliffe, P., *et al.* Structures of sequential open states in a symmetrical opening transition of the TolC exit duct. *Proc. National Acad. Sciences*; published online 18 January 2011.

Read, R.J., Adams, P.D., *et al.* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* 19 (10), 1395-412.

Sahakyan, A.B., Vranken, W., *et al.* (2011) Using side-chain aromatic proton chemical shifts for a quantitative analysis of protein structures. *Angew. Chem. Int. Ed Engl.* 50 (41), 9620-3.

Sahakyan, A.B., Vranken, W.F., *et al.* (2011) Structure-based prediction of methyl chemical shifts in proteins. *J. Biomol. NMR* 50 (4), 331-46.

Vandormael, B., De Wachter, R., *et al.* (2011) Assymetric synthesis and conformational analysis by NMR spectroscopy and MD of Aba- and α-MeAba-containing demorphin analogues. *Chem. Med. Chem.* 6 (11), 2035-47.

Velankar, S., Alhroub, Y., *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 39 (Database issue), D402-10.

Velankar, S., Alhroub, Y., *et al.* (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 40 (Database issue), D445-52.

Velankar, S. and Kleywegt, G.J. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. *Acta Crystallogr. D Biol. Crystallogr.* 67 (Pt 4), 324-30.

Von Der Lieth, C., Freire, A.A., *et al.* (2011) EUROCarbDB: An open-access platform for glycoinformatics. *Glycobiology* 21 (4), 493-502.

LOMAX: GENE ONTOLOGY

Alam-Faruque, Y., Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

Gaudet, P., Livstone, M. S., *et al.* (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinformatics*, 12 (5), 449-62.

Khodiyar, V.K., Hill, D.P., *et al.* (2011) The representation of heart development in the gene ontology. *Dev. Biol.* 354 (1), 9-17.

Leonelli, S., Diehl, A.D., *et al.* (2011) How the gene ontology evolves. *BMC Bioinformatics* 12, 325.

Meehan, T.F., Masci, A.M., *et al.* (2011) Logical development of the cell ontology. *BMC Bioinformatics* 12, 6.

Mungall, C.J., Bada, M., *et al.* (2011) Cross-product extensions of the Gene Ontology. *J. Biomed. Inform.* 44 (1), 80-6.

Mungall, C.J., Batchelor, C. and Eilbeck, K. (2011) Evolution of the Sequence Ontology terms and relationships. *J. Biomed. Inform.* 44 (1), 87-93.

Tirmizi, S.H., Aitken, S., *et al.* (2011) Mapping between the OBO and OWL ontology languages. *J. Biomed. Semantics* 2 (Suppl 1), S3.

MARTIN: UNIPROT DEVELOPMENT

Alam-Faruque, Y., Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB 'complete proteome' sets. *Proteomics* 11 (22), 4434-8.

Magrane, M. and UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, bar009.

Patient, S. and Martin, M. (2011) Annotating UniProt metagenomic and environmental sequences in UniMES. In: *Bioinformatics 2011 – Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. SciTePress, pp. 367-8.

The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39 (Database issue), D214-9.

Villaveces, J.M., Jimenez, R.C., *et al.* (2011) Dasty3, a WEB Framework for DAS. *Bioinformatics* 27 (18), 2616-7.

MCENTYRE: LITERATURE SERVICES

McEntyre, J.R., Ananiadou, S., *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.* 39 (Database issue), D58-65.

O'DONOVAN: UNIPROT CONTENT

Alam-Faruque, Y., Huntley, R.P., *et al.* (2011) The impact of focused gene ontology curation of specific Mammalian systems. *PLoS One* 6 (12), e27541.

Griss, J., Martin, M., *et al.* (2011) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB 'complete proteome' sets. *Proteomics* 11 (22), 4434-8.

Klimke, W., O'Donovan, C., *et al.* (2011) Solving the problem: Genome annotation standards before the data deluge. *Stand. Genomic Sci.* 5 (1), 168-93.

Magrane, M. and the UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011 bar009.

O'Donovan, C. and Apweiler, R. (2011) A guide to UniProt for protein scientists. *Methods Mol. Biol.* 694 25-35.

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39 (Database issue), D214-9.

OLDFIELD: PDBE DATABASES AND SERVICES

Lawson, C.L., Baker, M.L., *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39 (Database issue), D456-64.

Velankar, S., Alhroub, Y., *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 39 (Database issue), D402-10.

OVERINGTON: CHEMBL

Aranda, B., Blankenburg, H., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8 (7), 528-9.

Bellis, L.J., Akhtar, R., *et al.* (2011) Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* 39 (5), 1365-70.

Gaulton, A., Bellis, L.J., *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40 (Database issue), D1100-7.

Gleeson, M.P., Hersey, A., *et al.* (2011) Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discov.* 10 (3), 197-208.

Hopkins, A.L., Bickerton, G.R., *et al.* (2011) Rapid analysis of pharmacology for infectious diseases. *Curr. Top. Med. Chem.* 11 (10), 1292-300.

Kruger, F.A. and Overington, J.P. (2012) Global analysis of small molecule binding to related protein targets. *PLoS Comput. Biol.* 8 (1), e1002333.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

van der Horst Peironcely, J.E., *et al.* (2011) Chemogenomics approaches for receptor deorphanization and extensions of the chemogenomics concept to phenotypic space. *Curr. Top. Med. Chem.* 11 (15), 1964-77.

PARKINSON: FUNCTIONAL GENOMICS PRODUCTION

Adamusiak, T., Burdett, T., *et al.* (2011) OntoCAT - simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics* 12, 218.

Gostev, M., Faulconbridge, A., *et al.* (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.* 40 (Database issue), D64-70.

Kapushesky, M., Adamusiak, T., *et al.* (2012) Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40 (Database issue), D1077-81.

Kurbatova, N., Adamusiak, T., *et al.* (2011) ontoCAT: an R package for ontology traversal and search. *Bioinformatics* 27 (17), 2468-70.

Parkinson, H., Sarkans, U., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39 (Database issue), D1002-4.

Swertz, M.A., Dijkstra, M., *et al.* (2010) The MOLGENIS toolkit: Rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* 11 (SUPPL. 12), Art. No.: S12.

Travillian, R.S., Adamusiak, T., *et al.* (2011) Anatomy ontologies and potential users: bridging the gap. *J. Biomed. Semantics* 2 (Suppl 4), S3.

SARKANS: FUNCTIONAL GENOMICS DEVELOPMENT

Nicholson, G., Rantalainen, M., *et al.* (2011) A Genome-Wide Metabolic QTL Analysis in Europeans Implicates Two Loci Shaped by Recent Positive Selection. *PLoS Genet.* 7 (9), e1002270.

Nicholson, G., Rantalainen, M., *et al.* (2011) Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.* 7, 525.

Parkinson, H., Sarkans, U., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39 (Database issue), D1002-4.

STEINBECK: CHEMINFORMATICS & METABOLISM

Alcantara, R., Axelsen, K.B., *et al.* (2012) Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res.* 40 (D1), D754-60.

De Matos, P., Adams, N., *et al.* (2012) A Database for Chemical Proteomics: ChEBI. *Methods Mol. Biol.* 803, 273-96.

Griffin, J.L., Atherton, H.J., *et al.* (2011) A Metadata description of the data in "A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human.". *BMC Res. Notes* 4, 272.

Hastings, J., Chepelev, L., *et al.* (2011) The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One* 6 (10), e25513.

O'Boyle, N.M., Guha, R., *et al.* (2011) Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *J. Cheminform.* 3 (1), 37.

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

Truszkowski, A., Neumann, S., *et al.* (2011) CDK-Taverna 2.0: Migration and enhancements of an open-source pipelining solution. *Journal of Cheminformatics* 3 (Suppl. 1), Art. No. P5.

VELANKAR: PDBE CONTENT & INTEGRATION

Aranda, B., Blankenburg, H., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8 (7), 528-9.

Lawson, C.L., Baker, M.L., *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39 (Database issue), D456-64.

Velankar, S., Alhroub, Y., *et al.* (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 40 (Database issue), D445-52.

Velankar, S. and Kleywegt, G.J. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. *Acta Crystallogr. D Biol. Crystallogr.* 67 (Pt 4), 324-30.

SUPPORT TEAMS

BROOKSBANK: OUTREACH AND TRAINING

Klech, H., Brooksbank, C., *et al.* (2011) European initiative towards quality standards in education and training for discovery, development and use of medicines. *Eur. J. Pharm. Sci.* Article in press.

Schneider, M.V., and Orchard, S. (2011) Omics technologies, data and bioinformatics principles. *Methods Mol. Biol.* 719, 3-30.

Schneider, M.V., Walter, P., *et al.* (2011) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief Bioinform.* doi: 10.1093/bib/bbr064.

Via, A., De Las Rivas, J., *et al.* (2011) Ten simple rules for developing a short bioinformatics training course. *PLoS Comput. Biol.* 7 (10), e1002245.

CLARK: INDUSTRY PROGRAMME

Orchard, S., Al-Lazikani, B., *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10 (9), 661-9.

LOPEZ: EXTERNAL SERVICES

Hunter, S., Jones, P., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40 (Database issue), D306-12.

Robinson, J., Mistry, K., *et al.* (2011) The IMGT/HLA database. *Nucleic Acids Res.* 39 (DatabasSchneider, M.V., Walter, P., *et al.* (2011) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief Bioinform.* doi: 10.1093/bib/bbr064.

Schneider, V., Walter, P., *et al.* (2011) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. *Brief Bioinform.* doi: 10.1093/bib/bbr064.

Sievers, F., Wilm, A., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539.*e issue), D1171-6.*

EMBL-EBI

Annual Scientific Report 2011

Designed and Produced by PickeringHutchins Ltd www.pickeringhutchins.com

EMBL member states:
Austria, Croatia, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg,
the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom. Associate member state: Australia

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