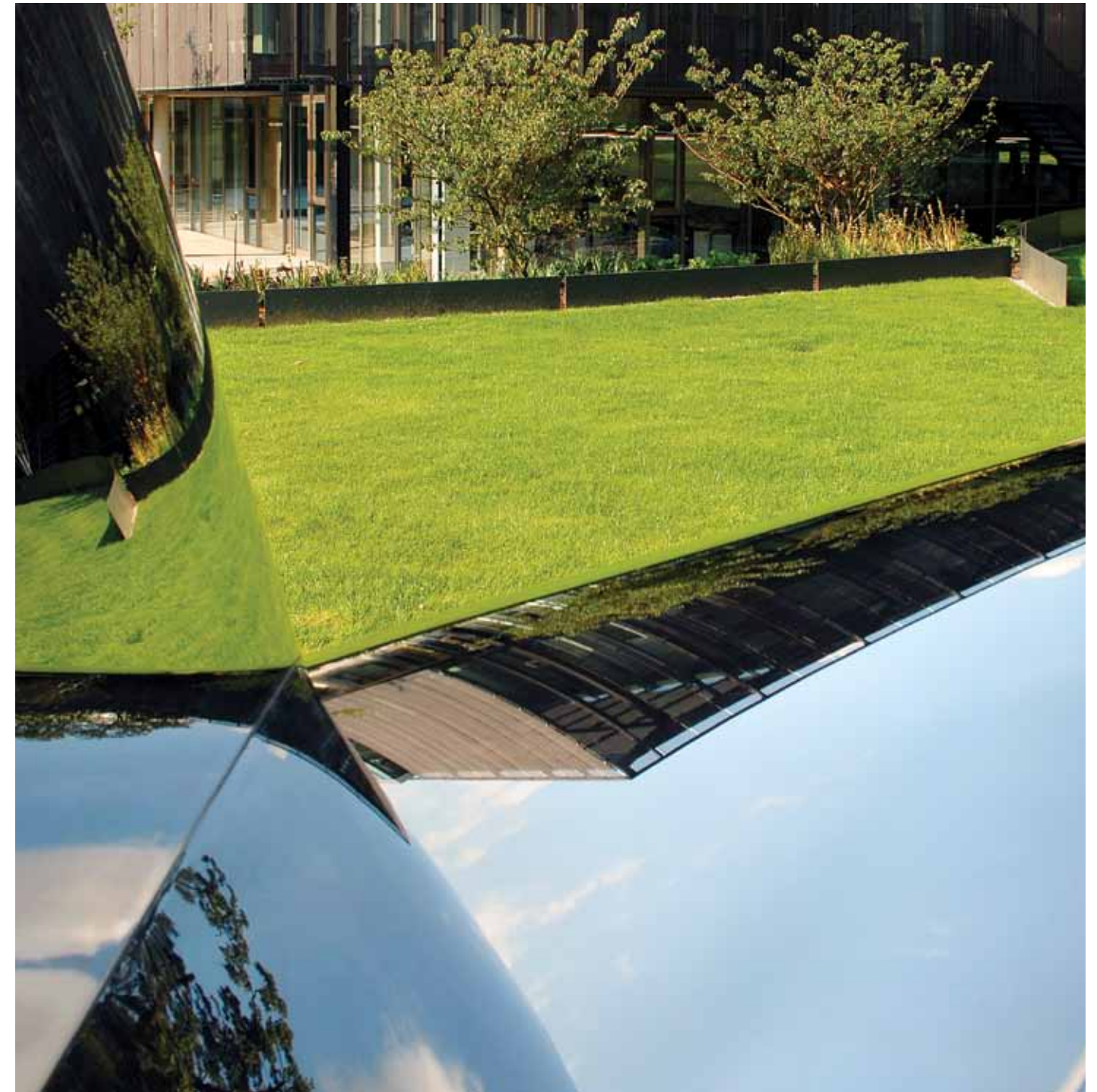


European Bioinformatics Institute Annual Scientific Report 2010

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)



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





Contents

Introduction

Foreword	2
Major Achievements 2009-2010	4

Services

Rolf Apweiler and Ewan Birney: PANDA	14
Alvis Brazma: Functional Genomics	20
Guy Cochrane: The European Nucleotide Archive	24
Paul Flicek: Vertebrate Genomics	26
Henning Hermjakob: Proteomics Services	28
Sarah Hunter: InterPro	30
Misha Kapushesky: Functional Genomics Atlas	32
Paul Kersey: Ensembl Genomes	34
Gerard Kleywegt: PDB	36
Jane Lomax: The Gene Ontology Editorial Office	38
Maria Martin: UniProt Development	40
Johanna McEntyre: Literature Resources	42
Claire O'Donovan: UniProt Content	44
John Overington: ChEMBL	46
Helen Parkinson: Functional Genomics Production	48
Peter Rice: Developing and Integrating Tools for Biologists	50
Ugis Sarkans: Functional Genomics Software Development	52
Christoph Steinbeck: Cheminformatics and Metabolism	54
Services Teams	56

Research

Paul Bertone: Pluripotency, Reprogramming and Differentiation	62
Anton Enright: Functional Genomics and Analysis of Small RNA function	64
Nick Goldman: Evolutionary Tools for Genomic Analysis	66
Nicolas Le Novère: Computational Systems Neurobiology	68
Nicholas Luscombe: Genomics and Regulatory Systems	70
Dietrich Rebholz-Schuhmann: Literature Research	72
Janet Thornton: Computational Biology of Proteins: Structure, Function and Evolution	74
The EMBL International PhD Programme at the EBI	76
Research Groups	78

Support

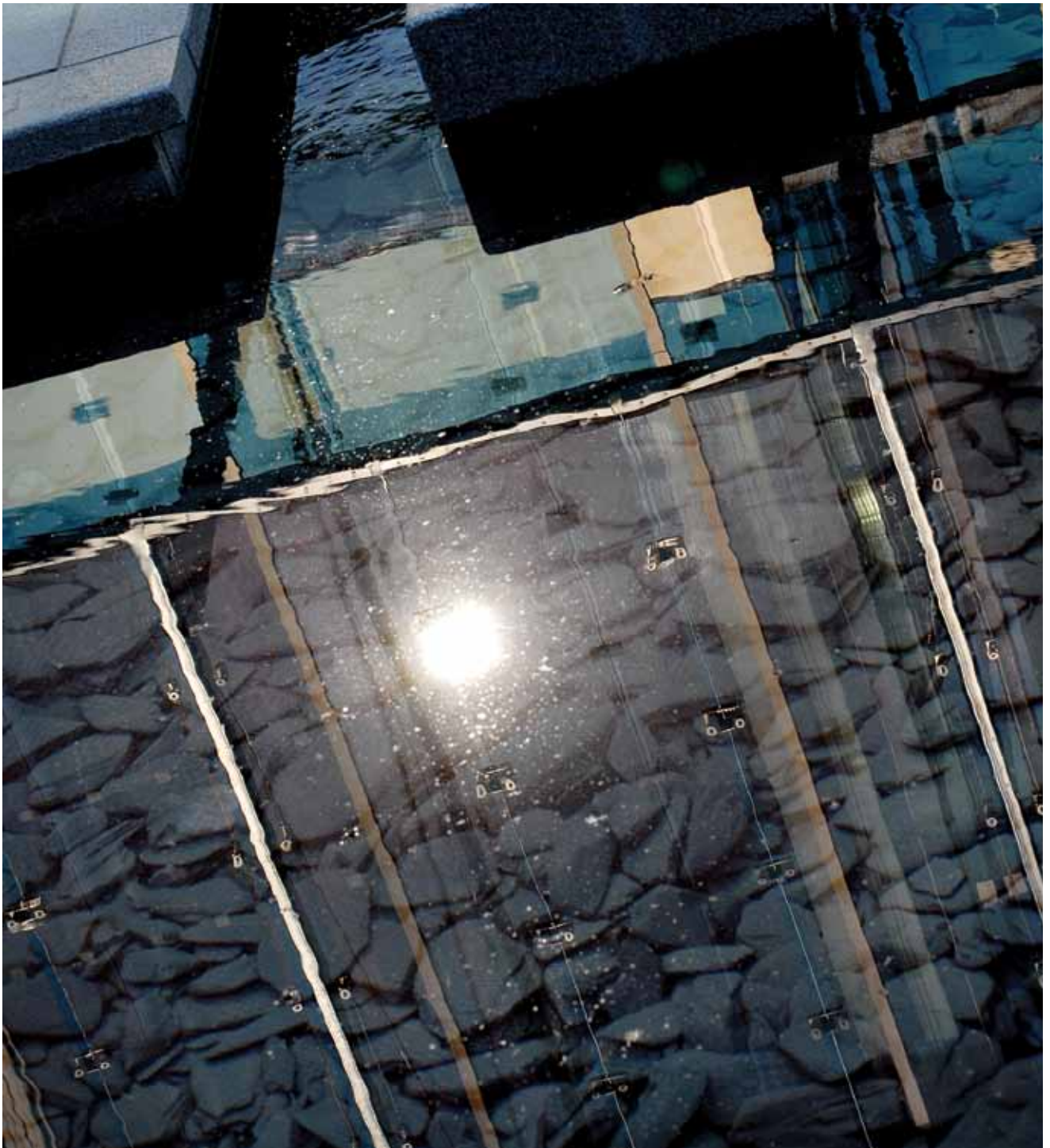
Cath Brooksbank: Outreach and Training	82
Dominic Clark: The EMBL-EBI Industry Programme	84
Petteri Jokinen: Systems and Networking	86
Rodrigo Lopez: External Services	88
Support Teams	90

Facts and Figures

A Year in Numbers	94
Major Database Collaborations	100
Scientific Advisory Boards and Committees	104
External Seminar Speakers	112
Publications	114

Index

Introduction





Janet Thornton

Director



Graham Cameron

Associate Director

Foreword

Welcome to EMBL-EBI's 2010 Annual Scientific Report.

The past year has seen substantial demands on EMBL-EBI in terms of handling the exponential growth of sequencing data, which has been made possible by impressive advances in DNA-sequencing technology. Life scientists now generate petabytes of data on a daily basis. Our mission to provide bioinformatics services, research, training and industrial support has never been more important or challenging.

For example, towards the end of 2010 the results of the international 1000 Genomes Project were published. EMBL-EBI, in partnership with the National Center for Biotechnology Information (NCBI) in the US, is responsible for storing and making available these data, which are beginning to elucidate the genetic basis of human individuality. In parallel, the extension of the Ensembl Genomes infrastructure to handle plants and fungi – as well as bacteria, protists and metazoa – now makes it possible for the EBI to support researchers making the most of new technologies to determine the genetic makeup of many different species.

But as the flood of data rises, so does the need for data integration. We are developing new portals and other mechanisms to allow data that has traditionally been held in different resources to be seamlessly integrated using both the web and programmatic interfaces. Such development, while technically demanding, is essential for meeting the specialised needs of our users. We expect to make many of these new developments publicly available in early 2011.

The increasing amount and complexity of new data underscores the need for better standardisation and integration. We have consolidated some of our service teams, creating a more layered structure better suited to handling these demands. Three new team leaders have joined our services. Helen Parkinson, under the guidance of Alvis Brazma in the Functional Genomics group, is overseeing data curation while Misha Kapushesky is leading the Gene Expression Atlas team. Tom Oldfield, under Gerard Kleywegt's leadership, is now Head of Databases and Services within the protein structure database (PDBe).

Curiosity-driven research at EMBL-EBI is flourishing, and has produced both exciting discoveries and powerful new tools. Of particular note this year are several different studies on the control of regulation and gene expression that draw on public data and data generated in multiple wet-dry collaborative experiments with colleagues at EMBL sites and throughout the world. In addition, two new research group leaders, Julio Saez-Rodriguez and John Marioni, are adding to the diversity of research at the EBI. Julio focuses on computational analysis of information transfer within signalling networks implicated in disease, while John develops computational and statistical methods to answer questions in evolutionary biology, especially those driven by new technological developments such as single-cell sequencing.

Our services are being used ever-more frequently by scientists within Europe and beyond. We provide simple web access for several hundred thousand independent users every month, and researchers are relying increasingly on programmatic access to large amounts of data through web-service technologies.

The computational and storage needs of all these projects are tremendous, and demand enormous growth in our storage capacity. Our new London Data Centre, funded by a £10 million (approximately €11.5 M) grant from the UK's Biotechnology and Biological Sciences Research Council (BBSRC), is now equipped with 3600 CPU cores and 3.8 PB of raw storage. Its first service, the ftp server, went live in June 2010. Our Systems and External Services teams have worked hard over the reporting period to carefully relocate many of our services to London. Production and Research work will remain at the Wellcome Trust Genome Campus's data centre, which we share with the Wellcome Trust Sanger Institute.

We are committed to providing training for our users. The EMBL-EBI training facility, opened only in 2007, is already operating to capacity, and demand for our courses remains very high. Our in-house courses are oversubscribed and demand for roadshows outstrips our capacity to deliver them. We are developing our e-learning programme, which will complement our face-to-face teaching and enable us to reach a larger audience. The economically driven move towards pre-competitive research in industry is generating renewed interest in working together to address our common challenges. Our Industry Programme, which delivers tailored training to our industry partners, has also grown, with three companies joining late in 2010.



The maintenance and development of our core databases and services is, and will remain, central to our mission. However, the emergence of ever more high-throughput methods and the resulting increase in diversity and quantity of data require a new model. It is clear that neither EMBL, the European Commission, nor any single nation can provide sufficient infrastructural funding to solve the entire problem.

These realisations gave rise to the ELIXIR ESFRI project (see page 8), whose consultation phase explored possible models in discussion with European scientific, financial and legal experts. During the consultation phase, it became clear that a large number of institutions distributed throughout Europe are enthusiastic and qualified to contribute to ELIXIR's mission. This naturally suggests a distributed solution. Although this poses substantial technical challenges and requires careful division of tasks, it is the only viable way to pool our expertise and collectively fund the solution. Recognising this, the ELIXIR Steering Committee has endorsed a 'hub-and-nodes' model, with the hub being at EMBL-EBI. Adoption of this model has now begun. Several European countries have committed funds to prepare for the construction of nodes, and have put ELIXIR on their national roadmaps for research infrastructure.

Indeed, an open call for suggestions for ELIXIR's nodes, issued in April 2010, received an overwhelmingly positive response. At the time of publication (November 2010), we have received suggestions for 54 nodes, from 23 countries. It is now necessary to secure commitment from further national funders who realise the importance of biological data to the future economic prosperity of their countries and are willing to contribute to this pan-European effort.

Extensive international collaborations are not new to the EBI. All our efforts rely on interaction with colleagues throughout the world. 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 built an extensive community of collaborators. It remains our privilege and pleasure to work with them.

Janet Thornton, Director

Graham Cameron, Associate Director

Major Achievements 2009–2010

SERVICES

The wide uptake of next-generation sequencing and other ultra-high throughput technologies by life scientists with interests spanning fundamental biology, medicine, agriculture and environmental science has led to unprecedented growth in data generation. It has also put the need for unrestricted access to biological data at the centre of biology. This is reflected in the use of EMBL-EBI's services, which exceeded 4.6 million requests a day (including Ensembl) by the end of June 2010. The EBI has a mandate to provide biomolecular data resources of universal relevance to biological and medical research; our services include the provision of biological databases and tools to explore them.

EMBL-EBI's constituency includes academic and commercial researchers throughout Europe and beyond, and we form a European node in many global data-sharing collaborations (see Figure 1). These service activities are accompanied by extensive outreach and training, concentrated mostly in Europe, with a dedicated team (see page 82); industry users receive targeted support through the EBI Industry Programme (page 84). Over the course of the reporting period (July 2009–June 2010) we effected some fundamental changes to our service provision. Some of the developments over the past year illustrate perfectly the scale of effort needed to stay one step ahead of our users' needs.

Nucleotides

The EBI is at the sharp end of spectacular improvements in the speed, capacity and affordability of DNA sequencing, with submission rates exceeding half a million bases per second (see Figure 2). Life scientists can now carry out experiments at rates previously undreamt of: for example, the International Cancer Genome Consortium plans to sequence 25 000 cancer genomes. But the assumption that the infrastructure will always be there to cope with this unimaginable onslaught of information (after all, it has been there since the launch of the EMBL Data Library in the mid-1980s) has become unrealistic.

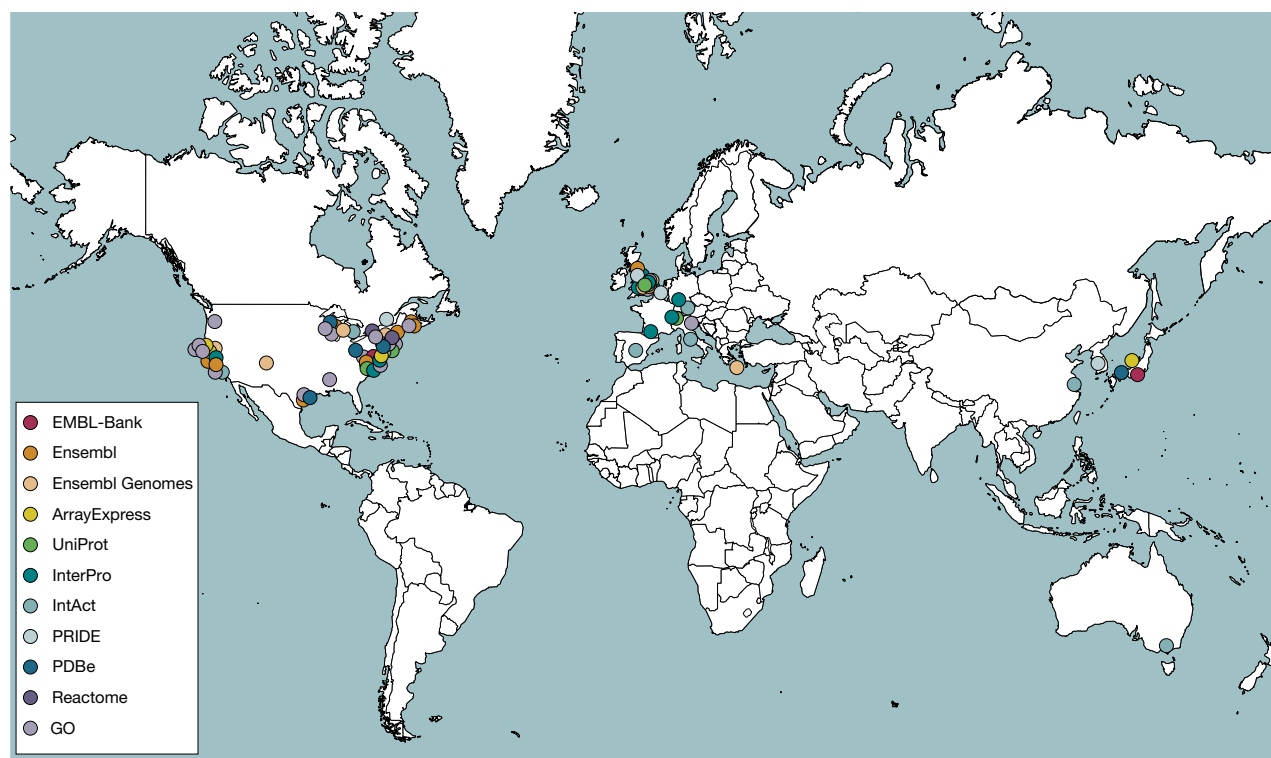


Figure 1. EMBL-EBI's collaborations for the major databases.

The **European Nucleotide Archive** (ENA) was launched in May 2010. Carefully annotated and cross-linked sequence records from the EMBL Nucleotide Sequence Database (EMBL-Bank) form the backbone of ENA, which also provides direct access to raw sequence data. Raw data from electrophoresis-based sequencing machines is held in the European Trace Archive, which was previously maintained at the Wellcome Trust Sanger Institute. The Sequence Read Archive (SRA) is a newly established repository for raw data from next-generation (array-based) sequencing platforms. Bringing together globally comprehensive, public-domain nucleotide sequence from all these data sets, ENA's services include: integrated browsing across all content; rapid, comprehensive sequence-similarity search; graphical assembly and feature visualisation; and enhancements to the Webin submissions infrastructure. In addition, the team has launched simple programmatic services through which all ENA content can be accessed.

The **1000 Genomes Project** is the largest coordinated data production and analysis project yet undertaken in genomics. The EBI leads the project's data coordination; the complete pre-publication release of the raw and processed pilot project data was announced in March 2010. This data set includes 15 million single nucleotide polymorphisms (SNPs; 8 million discovered in this project alone), 1 million small insertion and deletion polymorphisms and 20 000 larger structural variants. By releasing all of this data in advance of publication, we enabled it to be used to inform medical sequencing and other disease-related studies in a timely fashion. Several important results based on this pre-publication data have already been announced.

Ensembl Genomes is a significant new project designed to leverage the power of the Ensembl system for genome analysis and display beyond the vertebrates (for which it was originally designed) to meet the needs of communities working on species from all domains of life. These communities are generating increasing quantities of genome-scale data but lack the bioinformatics infrastructure for its integration and exploitation. In September 2009, we launched two new divisions of Ensembl Genomes: Ensembl Plants and Ensembl Fungi. These portals followed the launch of Ensembl Bacteria, Protists and Metazoa earlier in the year and completed the launch phase of Ensembl Genomes.

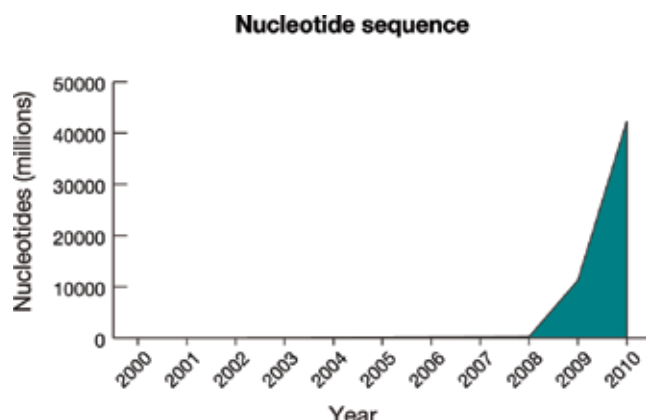


Figure 2. Growth of submissions to the European Nucleotide Archive (previously known as the EMBL Data Library then EMBL-Bank).

Functional genomics

In the era of multi-omics experiments, keeping track of the origin of the data, and being able to cross-reference, for example, a sample on which genome sequencing, expression profiling, proteomics and metabolomics information has been gathered, is becoming essential. The Functional Genomics group has been developing a pilot version of the **BioSample Database**, which will contain information about biological samples used in experiments. The data from the experiments themselves are stored in other EBI databases, such as ArrayExpress, ENA and PRIDE. A particular set of biosamples (e.g. commonly used cell lines, mouse strains) may be referenced by many experiments.

Another result of faster, cheaper sequencing technology is that gene-expression studies are moving away from microarray-based experiments and towards the sequencing of cDNA (known as RNA-seq). The Functional Genomics Production Team is now fully capable of accepting RNA-seq information, and has made available 400 sequencing-based experiments in the **ArrayExpress** database. The team has developed integrated submission processes with the short read archives of ENA and EGA to integrate data within the EBI, and has produced new submission tools to support annotation and submission of these data. In addition, they have agreed on and implemented a system for exchanging data with the NCBI Gene Expression Omnibus.

The **Gene Expression Atlas** is now available as an open-source, stand-alone version. Users can install the complete Atlas locally and use it to load and view private datasets together with the public data. The Atlas includes the global map of human gene expression (Luk et al., 2010) as well as a pipeline for processing RNA-seq studies. The R Cloud service, launched this summer, provides remote access to Atlas data and the R cloud-computing environment on EBI servers.

Small molecules

Enhancements to our cheminformatics services have continued apace in the past year. **ChEMBLdb**, the EBI's database of bioactive compounds (drugs and drug-like molecules), was publicly launched in January 2010. The database has grown substantially in size (an increase of over 50% in bioactivity content), with updates released monthly. The ChEMBL data has been integrated with other EBI resources. It has also been integrated with the US-based bioactivity resource PubChem, with further integration work proceeding as a high priority. Of particular note was the deposition of previously private datasets

into ChEMBL, including the malaria-screening results released by GlaxoSmithKline, Novartis and St Jude Children's Research Hospital. Our hope is that ready access to these data sets will expedite much-needed research on new therapies for neglected tropical diseases.

ChEBI, our chemistry database and ontology, now accepts submissions. A bulk submission facility has also been developed. As part of a collaboration with the La Jolla Institute of Allergy and Immunology in the US, we have curated around 1500 entities associated with immunology, making good use of a new facility for including citations in compound records. ChEBI has also successfully registered over 600 000 small molecules populated from the ChEMBL database. As well as making chemical data openly available, we have also made good progress in developing open-source software tools for analysing small molecules. The ChEBI team has released OrChem, an open-source package that allows users to perform substructure, similarity and exact searching of chemical structures stored in an Oracle database. Wild-card chemical-structure searching has recently been added to this toolkit.

Data integration and development of new user interfaces

Excellent progress has been made on the development of a **new search and browse** engine, designed to better serve bench-based life scientists – our end users – and is due for launch in early 2011. The new engine will provide our users with an integrated view of gene, expression, protein, family, structure and literature information and will rank search results in a researcher-oriented manner. For example, if a user enters p53 in the search box, the chances are that he or she is looking for the (human) p53 gene, not the gene encoding a p53-binding protein or a p53-like protein; the results-ranking system for the new search engine takes this into account. Jenny Cham, our User Experience Analyst, has worked with users throughout Europe, both in academia and in industry, to carry out face-to-face user testing, which has been fed back into the development of the service. Central to the new service is the EBI's general search engine. With more than 300 million entries indexed and updated daily, the search acts as a gateway to all the major EBI data providers' sites. Its programmatic interface is now used by several groups both within and beyond the EBI.

The scientific literature is the first port of call for the vast majority of life scientists when they are searching for information. Jo McEntyre's first year at the EBI as leader of the **Literature Services Team** (see page 42) has led to some exciting developments in the UK PubMed Central (UKPMC) project. UKPMC is a unique, innovative and free online resource offering access to information sources for biomedical and health researchers. It is being developed by EMBL-EBI together with the British Library and the University of Manchester and in close cooperation with the NCBI; our role has been to develop a full-text semantic index and retrieval service for UKPMC. In January 2010, UKPMC moved beyond its basic mirror function of PubMed Central USA and launched a new website, powered by this service. The service provides one-stop searching for both citation and full-text results, incorporating text-mined terms such as genes, proteins, organisms, GO terms, accession numbers, diseases and small molecules. Updated daily as new articles are deposited, the web service provides the core functions that set UKPMC apart from similar resources.

Gerard Kleywegt's first year as leader of the **PDBe Group** (see page 36) has also been a productive one, and the newly designed PDBe website, which includes several new services, has been accompanied by encouraging usage statistics.



Data standards

The development of data standards remains at the heart of the EBI's collaborative approach to bioinformatics, and there have been important milestones for molecular interactions and microbial and viral nomenclature this year. In February 2010, the International Molecular Exchange Consortium (IMEx) entered production mode, co-ordinating the curation of molecular interactions among five major interaction databases in Europe and the United States, including the EBI's IntAct database. In April 2010, researchers from life-science organisations worldwide worked together at the third NCBI Genome Annotation Workshop to develop and gain community acceptance of prokaryotic protein naming guidelines. Following this agreement, the International Sequence Database Collaboration (INSDC) and UniProt created a more generalised set of guidelines to make these standards useful for taxa beyond the cellular prokaryotes. The decision by the INSDC to offer these guidelines for adoption by all submitters to their databases will greatly enhance the annotation of complete genomes and proteomes and ensure that the user community can exploit these data to their full potential.

RESEARCH

Our research groups, which comprise around a quarter of EBI's personnel, perform computational research into many different biological questions, ranging from genome evolution and transcriptional regulation to systems modelling of basic biological processes and disease. Bioinformatics continues to diversify, often led by the development of new technologies that generate the need for new methods for data handling and interpretation. Our research groups are compact, typically with two or three students and externally funded postdocs. Their research complements the broad remit of EBI's service provision, benefitting from the in-house technical expertise provided by the larger service teams and in turn helping to identify current challenges for researchers using our data resources. Several service teams also incorporate a small research and development component.

Curiosity-driven research

While the EBI's services focus on gathering and presenting comprehensive collections of information, much of our research addresses how that information is used in living organisms to choreograph the processes of life. Some approaches adopt a genome-wide approach, whereas others zoom in on specific processes.

Measuring microRNA expression. MicroRNAs (miRNAs) constitute an important class of regulatory molecules known to play critical roles in development and disease. Paul Berton's group (see page 62) performed a large-scale analysis of leading microarray platforms, next-generation sequencing and quantitative real-time assays to determine the relative accuracy of these methods for the investigation of microRNA function (Git et al., 2010). This study provides the most comprehensive assessment of its kind and reveals fundamental differences in the application of various technologies to RNA genomics.

MicroRNAs and regulation of red blood cell development. A collaborative effort between the Enright group (see page 64) at EMBL-EBI and the O'Carroll Group in EMBL-Monterotondo (Rasmussen et al., 2010) used a combined computational and experimental approach to study a pair of miRNAs that exist as a cluster, and the influence of this cluster on the development of red blood cells. By knocking out the cluster and comparing gene-expression patterns in knockout versus wild-type mice, they found that miR-451 had a stronger effect on regulation of red blood cell development but both miRNAs contributed. A large set of predicted functional targets of the miRNAs is being further analysed for other downstream functional effects.

The relationship between nuclear pores and active regions of the genome. Nick Luscombe's group (see page 70) has shown that the nuclear pore components Nup153 and Megator bind to a quarter of the *Drosophila* genome in the form of chromosomal domains (Vaquerizas et al., 2010). These domains represent active regions of the genome. The group used extensive three-dimensional image analysis to show that pools of these proteins from both within and outside the nucleus contribute to these domains. This implies that chromosomal organisation by nuclear pore proteins could contribute to global gene-expression control.

Understanding membrane-pore specificity. Using previously developed tools for membrane-pore analysis, the Thornton group (see page 74) completed a study of the specificities of aquaporins (water channels), showing that specificity is related to the electrostatic potential along the channel. Specific electrostatic fingerprints were observed for the aquaporin subfamilies (according to their permeability to water and/or glycerol) and for the potassium-channel family. This simple approach gives a better understanding of specificity in this important family of transporters and explains the effects on function of some disease-related mutations (Oliva et al., in press).

Global analysis of gene expression. By integrating gene expression data from an unprecedented variety of human tissue samples, Alvis Brazma's team (see page 20) and their collaborators produced the world's first global map of gene expression (Lukk et al., 2010). The analysis used data collected from 163 laboratories worldwide involving 5372 human samples from various tissues, cell types and diseases. Most transcriptomics experiments compare gene expression in only a few cell types or conditions and although technically challenging, integrating this data on a large scale has created a new way for scientists to explore gene expression.

Rewiring of gene regulation. Paul Flicek's group (see page 26) contributed to a study that directly interrogated the evolutionary mechanisms of transcription factor binding using matched experiments in liver tissues across 300 million years of evolution (Schmidt et al., 2010). Contrary to expectation there is little conservation in the location of transcription factor binding, and the conservation that is observed appears associated with embryonic development. Despite this significant rewiring of transcriptional regulation, these factors still manage to maintain the largely conserved gene expression and function of liver tissue.

Resources developed for research

The EBI's research aims to develop new ways to understand biology through bioinformatics. Some of this research involves the development of new resources that enable our research groups to answer these biological questions. In the EBI's spirit of open access, these resources are made openly available to other researchers.

Phylogeny-aware multiple sequence analysis. Nick Goldman's group (see page 66) has released the webPRANK server (<http://tinyurl.com/webprank>). This is built upon the PRANK multiple sequence aligner developed by Ari Löytynoja and Nick Goldman (Löytynoja and Goldman, in press). PRANK is 'phylogeny aware': its inferences about insertion and deletion events are informed by the evolutionary relationships of the sequences being aligned. Alignments are a cornerstone of sequence analysis, and PRANK alignments have been shown to be superior for various downstream evolutionary analyses.

Circuit diagrams for biologists. Nicolas Le Novère's group (see page 68) has contributed towards the development of SBGN – a visual notation equivalent to 'circuit diagrams' for biologists (Le Novère et al., 2009). Developed by a community of biochemists, modellers and computer scientists, SBGN consists of three complementary languages: process descriptions, entity relationships and activity flows. Together they enable scientists to represent networks of biochemical interactions in a standard, unambiguous way.

Standardising the scientific literature. Dietrich Rebholz-Schuhmann's group (see page 72) has coordinated the creation of the CALBC Silver Standard Corpus (Rebholz-Schuhmann, 2010). This is a large-scale repository of documents that have been annotated with different semantic types (e.g. genes, proteins, diseases, chemical entities, species). Annotations have been harmonised through pair-wise or multiple alignment of tokens and their semantic tags. This new standard will facilitate the linking of different semantic types to their relevant primary data resources – a previously unsolved problem.

COORDINATION OF MAJOR EUROPEAN PROJECTS

Europe has always been at the forefront of bioinformatics research, but as we move towards the European Union's goal of a single European Research Area there is a greater need than ever for bioinformatics experts and experimental biologists throughout Europe to work together to achieve common goals. As well as serving the biological research community by providing Europe's core biological data resources, the EBI coordinates Europe's bioinformatics service providers, effectively adding value by distributing effort. One of our most significant projects in this respect is ELIXIR, the nascent European life sciences infrastructure for biological information.

PAN-EUROPEAN INFRASTRUCTURES: ELIXIR

The purpose of ELIXIR is to create a sustainable infrastructure for biological information in Europe. This is pivotal for academic and industrial research as Europe tries to find solutions to the Grand Challenges of providing healthcare for an aging population, securing a sustainable food supply and protecting the environment. It is also of vital importance if Europe is to retain a competitive pharmaceutical sector.

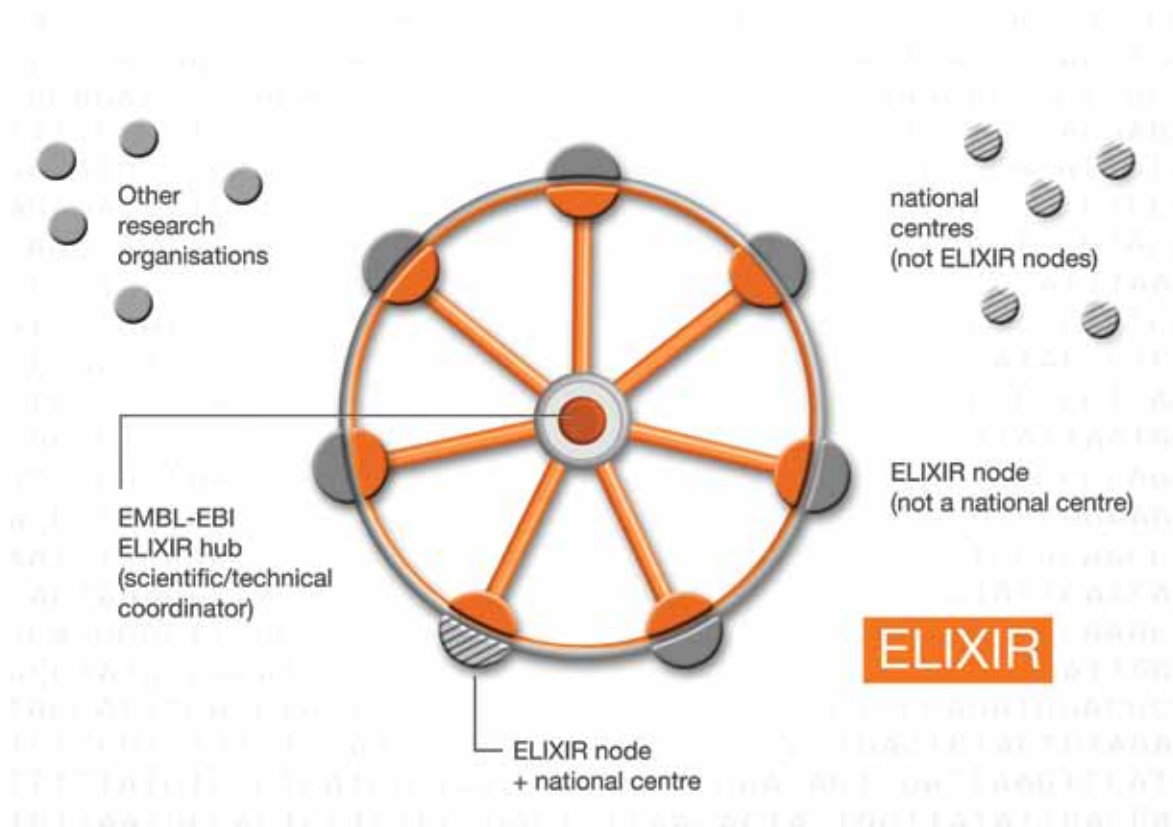


Figure 3. ELIXIR's hub-and-nodes model.

The preparatory phase of ELIXIR has been funded by the EU. In October 2006 the European Strategy Forum on Research Infrastructures (ESFRI), a body set up by 33 countries at the initiative of the European Council, published its first roadmap. This identified 35 pan-European research infrastructure projects that are of key importance for the development of science and innovation. One was an upgrade of Europe's bioinformatics infrastructure. When the European Commission (EC), through its Seventh Framework Programme (FP7), committed funds to preparatory-phase ESFRI projects, EMBL-EBI coordinated an application for ELIXIR. Funded via a €4.5 million grant, the project runs from 2008 to 2011 and involves 32 partner organisations from 13 EU Member States. ELIXIR is both a research infrastructure and an e-infrastructure, and works closely with other infrastructures (see below) to ensure that biomedical research data is managed in a consistent way. It is one of very few ESFRI preparatory-phase projects considered to be of global significance.

Major achievements

In August 2009 the UK's Biotechnology and Biological Sciences Research Council (BBSRC) awarded £10M (approximately €11.5 M) to EMBL-EBI to initiate the construction of the ELIXIR European data centre. This will enable the EBI – as the hub of a larger, distributed, pan-European infrastructure – to continue serving its core data resources while addressing the increasing demand to perform ever-more sophisticated analyses of data sets that are growing at supra-exponential rates. As part of this effort, the Systems and Networking Team (see page 86) has begun the phased relocation of major EBI services to the new data centre.

The data centre has a geographically distributed topology to protect against local disaster, provides high-level security (equivalent to that for processing credit-card payments) and is sited very close to high-bandwidth internet connections. This is a major step forward, not least because it helps secure the short-term provision of Europe's core data resources for biological research – and their deployment through ELIXIR – while ensuring the robustness and availability of our services, which is essential for our users.

The BBSRC funding sends a powerful message showing UK support for the ELIXIR concept, which will undoubtedly be critical for the construction phase. Welcome as this generous grant was, the London data centre project represents only a short-term solution to Europe's growing data-infrastructure needs. Our efforts are now focused on finding ways to ensure that funding will be available for the continuing development of the data centre so that it may fulfil the expectations of European life-science and medical researchers well into the future.

The consultation phase brought to light an almost universal desire to see ELIXIR distributed across Europe. Although technically challenging, this approach is believed to be the best way to attract funding; to optimise growth and flexibility; and to benefit from local collaborations with pre-existing centres of excellence. The 'hub-and-nodes' concept (see Figure 3), with the hub being part of the EBI, has been endorsed by the ELIXIR Steering Committee. Several European countries have put ELIXIR on their national roadmaps for research infrastructure, and some (Sweden Finland, Denmark and Spain, in order of commitment) have already committed funds to prepare for the construction of ELIXIR nodes.

The design stage of the ELIXIR Technical Hub Building was completed during the reporting period. If funded, this new building will accommodate the ELIXIR Secretariat, the staff who will operate the ELIXIR European Data Centre, the ELIXIR Industry Translational Facility and a training facility. We are seeking funding to commence construction as soon as possible. This is a crucial step for ELIXIR for without it there will be nowhere to house the staff – and therefore no ELIXIR.

An open call for suggestions for ELIXIR's nodes was issued in April 2010 and has received an overwhelmingly positive response: at the time of publication 23 countries have submitted 54 suggestions. These will be considered by the Steering Committee, not for scientific or technical review (this is the role of the member state funding agencies) but to begin mapping ELIXIR's landscape and considering the process by which it will be constructed. In 2011 ELIXIR will set up a consortium of countries that are interested in joining; this group will develop a process for deciding which sites are best suited to ensuring a stable data infrastructure, and how they will be funded.

In light of the limited EC funds available for research infrastructures, securing commitments from national funders – and impressing upon them the importance of biological data to Europe's future economic prosperity – is absolutely essential. We are therefore preparing, with input from the funding agencies, a business case and other legal documents. We continue to build awareness of ELIXIR amongst policymakers, and are dedicated to securing the commitment of the EC and EU Member States to funding ELIXIR's construction and operation.

The current version of the ESFRI roadmap includes ten biomedical science research infrastructures, and three new ones are about to be added. In addition to coordinating ELIXIR, EMBL-EBI is involved in defining data infrastructures for several others. These include InfraFrontier (the European infrastructure for phenotyping and archiving of model mammalian genomes), EU-Openscreen (the European infrastructure of open screening platforms for chemical biology) and the European Marine Biological Resource Centre.

INVOLVEMENT IN OTHER MAJOR PAN-EUROPEAN INITIATIVES

The Innovative Medicines Initiative (IMI), Europe's largest public-private initiative, aims to speed up the development of better and safer medicines for patients. IMI supports collaborative research projects and builds networks of industrial and academic experts in order to boost pharmaceutical innovation in Europe. IMI is a joint undertaking between the EU and EFPIA (the European Federation of Pharmaceutical Industries and Associations). EMBL-EBI is involved in several IMI projects.

EMTRAIN, which involves the original six ESFRI biomedical science research infrastructures, is building a platform for training in medicines research; the EBI contributes to the project as a representative of ELIXIR. We are also contributing to the IMI-affiliated eTox project, which aims to improve drug safety by applying bioinformatics and cheminformatics data-mining approaches to discover or optimise computational safety prediction methods. The EBI is also developing an infrastructure to encode, process and exchange pharmacokinetics and pharmacodynamics models used in drug discovery for the IMI project DDMoRe.

EMBRACE

EMBL-EBI coordinates a number of EU-funded projects and is a regular partner in many others. Lack of space precludes us from discussing every project here; however, one project, EMBRACE, deserves special mention as it was successfully completed in 2010.

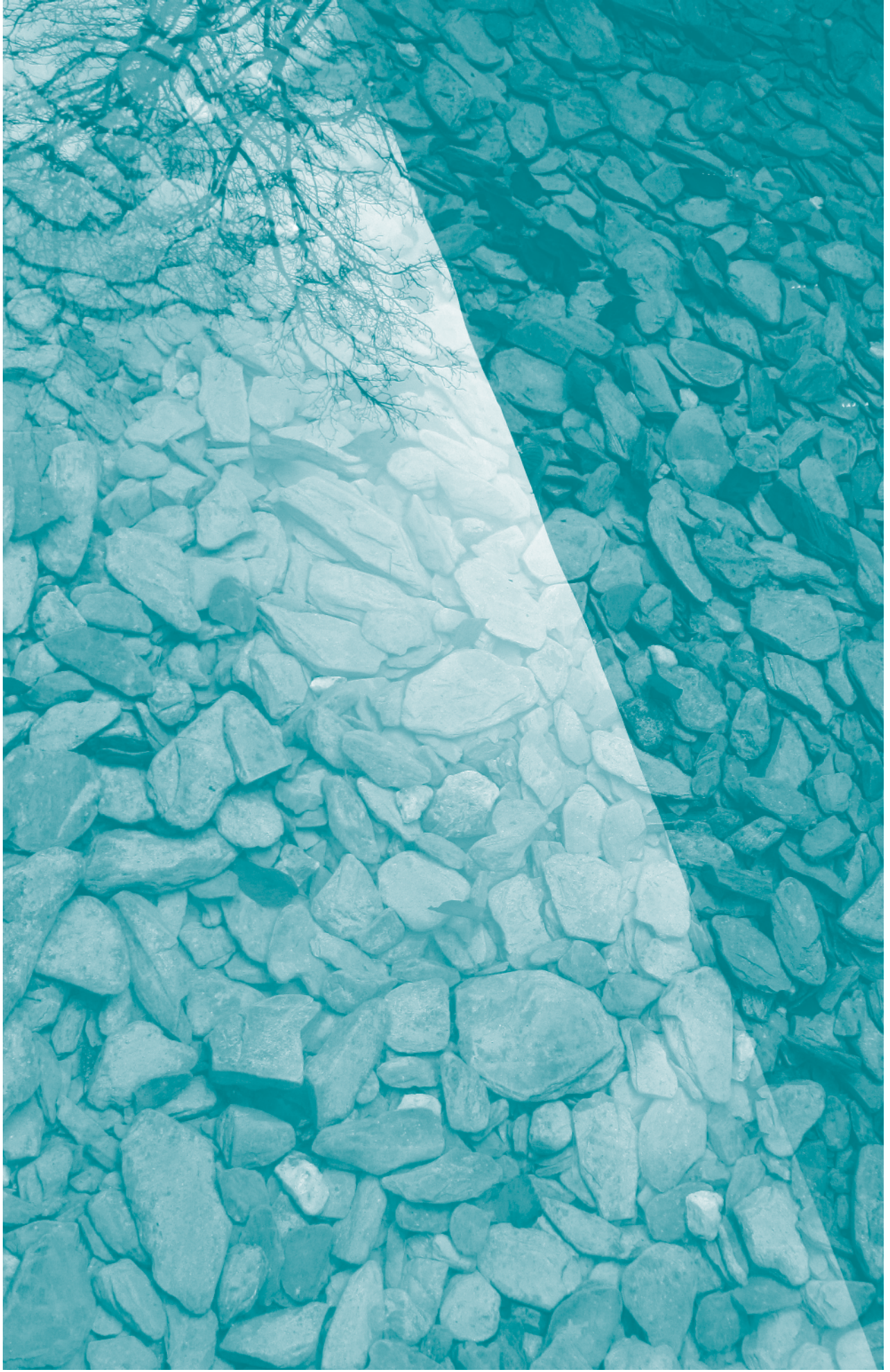
The past decade has seen a series of paradigm shifts in the life sciences. One such shift is from hypothesis-driven to data-driven research. The vast amounts of -omics data and, more importantly, the many different *types* of data (e.g. genomics, transcriptomics, proteomics, epigenomics, metabolomics) require a second paradigm shift: life scientists must collaborate on a grand scale to achieve progress. EMBRACE, an EU-funded network of excellence with 17 partners in 11 countries, began in February 2005. The EMBRACE partners realised early on that this new way of working requires novel technologies for data and software interoperability. The partners have developed standards, test cases, and fully fledged solutions that address this challenge. These require new ontologies, data-exchange formats and protocols, and ways to deal with the massive computational tasks at hand.

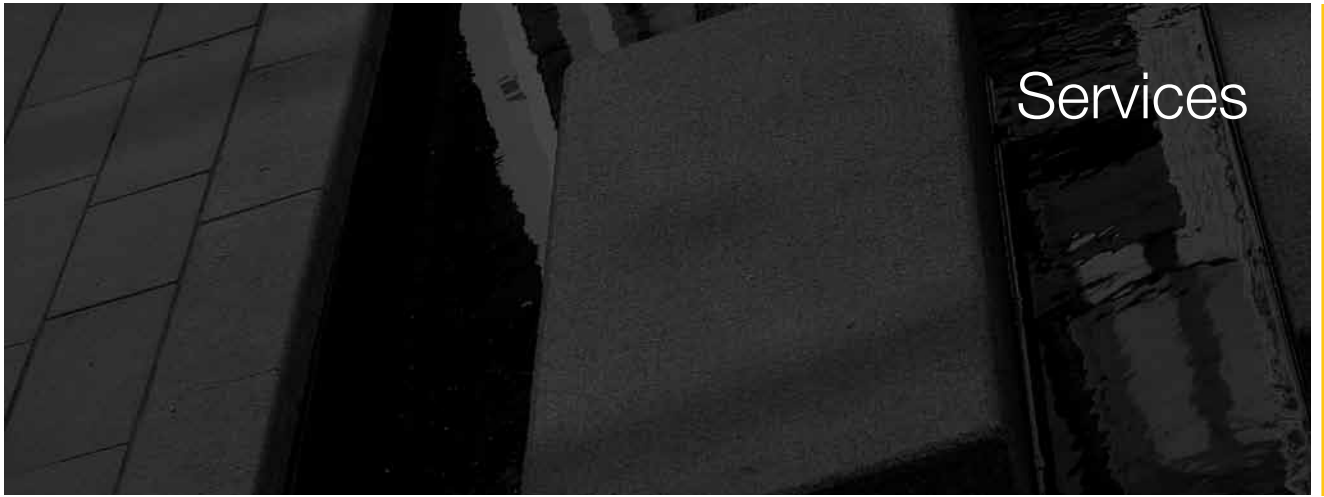
EMBRACE solutions, now widely accepted in the European bioinformatics community, include the SOAP protocol for web service usage; a large series of test cases illustrate the power of this concept. The EMBRACE web service registry, which has now been incorporated into the Biocatalogue, lists nearly 1000 freely accessible web services. The growing use of EDAM, the EMBRACE ontology for bioinformatics web services, is a testament to its importance to the development of future semantic web applications.

Perhaps the most important aspect of EMBRACE is that it got European developers in this field to discuss interoperability issues openly and to agree on a single standard, rather than working alone on a large number of incompatible ones. EMBRACE project partners are the opinion formers in this field, and will continue discussing and collaborating on these topics long after the successful completion of the project in January 2010. That will be very beneficial for bioinformatics, for the life sciences and for Europe as a whole.

SELECTED REFERENCES

- 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.
- Le Novère, N., et al. (2009) The Systems Biology Graphical Notation. *Nature Biotechnol.* 27, 735–741.
- Löytynoja, A. and Goldman, N. (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* (in press).
- Lukk, M., et al. (2010) A global map of human gene expression. *Nature Biotechnol.* 28, 322–324.
- Oliva, R., et al. (2010) Electrostatics of aquaporin and aquaglyceroporin channels correlates with their transport selectivity. *Proc. Nat. Acad. Sci.* 107, 4135–4140.
- Rasmussen, K.D., et al. (2010) The miR-144/451 locus is required for erythroid homeostasis. *J. Exp. Med.* 207, 1351–1358.
- Rebholz-Schuhmann, D., et al. (2010) CALBC Silver Standard Corpus. *J. Bioinform. Comput. Biol.* 8, 163–179.
- Schmidt, D., et al. (2010) Five vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040.
- Vaquerizas, J.M., et al. (2010) Nuclear pore proteins Nup153 and Megator define transcriptionally active regions in the *Drosophila* genome. *PLoS Genet.* 6, e1000846.







Rolf Apweiler

*Joint Team Leader,
PANDA Group (Proteins)
PhD University of
Heidelberg, 1994.
At EMBL since 1987, at
EMBL-EBI since 1994.*



Ewan Birney

*Joint Team Leader, PANDA
Group (Nucleotides)
PhD Sanger Institute, 2000.
At EMBL-EBI since 2000.*

Protein and Nucleotide Data

DESCRIPTION OF SERVICES AND RESEARCH

The Protein and Nucleotide Data (PANDA) group was created in June 2007 by merging the former Ensembl (Ewan Birney) and Sequence Database (Rolf Apweiler) groups. The PANDA 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 (ENA), the Ensembl and Ensembl Genomes resources, the UniProt protein resource, the InterPro domain resource and a range of other biomolecular databases. These efforts can be divided into three major groups: nucleotides, proteins, and cheminformatics and metabolism.

In addition to PANDA activities, both the Birney and Apweiler groups have complementary research components. Substantial training and outreach efforts are also part of the PANDA group's activities. Various external service aspects of the PANDA group's activities are described by Rodrigo Lopez (see External Services, page 88). The activities of the Vertebrate Genomics, Ensembl Genomes, ENA, Proteomics Services, InterPro, UniProt, ChEMBL, and Chemoinformatics and Metabolism groups are also described separately.

SUMMARY OF PROGRESS

- Handled a growing amount of nucleotide and protein data;
- Launched the ChEMBL database, which now contains 3 million experimental bioactivities and approximately 750 000 compounds;
- Launched the DGVA database for copy-number and structural-variation data;
- Launched Ensembl Plants and Ensembl Fungi, completing the set of Ensembl Genomes portals spanning the taxonomic space;
- Launched the European Nucleotide Archive;
- Provided three new InterPro data sources via the Distributed Annotation System (DAS) protocol – including new data from the PRIAM (predicting enzyme families) database;
- Updated the InterPro web interface to show matches to the automatically generated portion of Pfam (Pfam-B);
- Received 576 ChEBI submissions from 16 individual external users and developed a bulk submission tool to allow programmatic provision of submitted data;
- Implemented two new UniProt interfaces: BioMart and DAS;
- Released OrChem, our open-source chemical-search cartridge for Oracle™.

PANDA NUCLEOTIDES (Ewan Birney)

Strategy

DNA sequence remains at the heart of molecular biology and bioinformatics. In 2010 we continued to see a progressive shift towards using high-throughput, 'next generation' sequencing technologies; this has affected all teams. We also launched the ENA, which provides a single, integrated set of resources to nucleotide archives, from short reads through to traditional assembled sequence entries. Other highlights include the release of the 1000 Genomes Project pilot data (see Paul Flicek, page 26) and expansion of the Ensembl Genomes platform across many more species.

Ewan Birney provides strategic oversight across the three main branches of the PANDA Nucleotides Group: Vertebrate Genomics (including Ensembl), led by Paul Flicek; Ensembl Genomes, led by Paul Kersey; and the ENA, led by Guy Cochrane. In addition, the HUGO Gene Nomenclature Committee (HGNC), a smaller group coordinated by Elspeth Bruford, is part of PANDA Nucleotides and is presented here. The key organising principle across all groups is to coordinate resources for each genome sequence of a species in the best possible way.

A key difference between the groups is the provenance of the data. In the case of the ENA, the content is determined by the submitter; any added-value information is provided as an additional resource. These data can be redundant and conflicting but represents the foundational DNA dataset on which all genomic and nearly all protein sequence is based. This dataset is coordinated by virtue of the International Nucleotide Sequence Database Collaboration (INSDC), forming a single, worldwide coordinated set of information with partner groups at NCBI and DDBJ. In contrast, the Ensembl and Ensembl Genomes resources are community-led, and we aim to present a single, non-redundant view of a species' DNA information organised around its genomic sequence. In this case, decisions are made (via interactions with the community) based on the optimal representation of information that will provide the most utility to users. In the human genome, an important component of this community is the unambiguous assignment of gene symbols, which allows researchers to use memorable names for genes in scientific communication. This is provided by the HGNC group.

HUGO Gene Nomenclature Committee (HGNC)

Elsbeth 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 (abbreviation of the name) for every human gene, and to ensure this information is freely available, widely disseminated and universally used. Achieving these goals involves three key components: bioinformatic analysis of nucleotide and protein sequences, curation of online resources and communication.

Our curation of online resources involves 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). Our on-going communication efforts include consultation with researchers; coordinated naming of orthologous genes with nomenclature groups in other species; exchanging data with numerous databases; and raising awareness of the resource within the scientific community. During the reporting period HGNC staff have attended seven international conferences, including the 59th Annual American Society of Human Genetics Meeting, where we shared a booth with the Human Genome Organisation (HUGO).

Gene naming has continued to focus on the increasing number of genes identified by the CCDS project, with less than 50 of an estimated 18 650 genes in the CCDS set still awaiting an HGNC-approved gene symbol. As of October 2010 there were a total of 29 801 approved gene symbols, an increase of over 1550 in the past year. This is largely due to the increased assignment of names for non-protein-coding RNA genes and the naming of pseudogenes based on the pseudogene.org dataset.

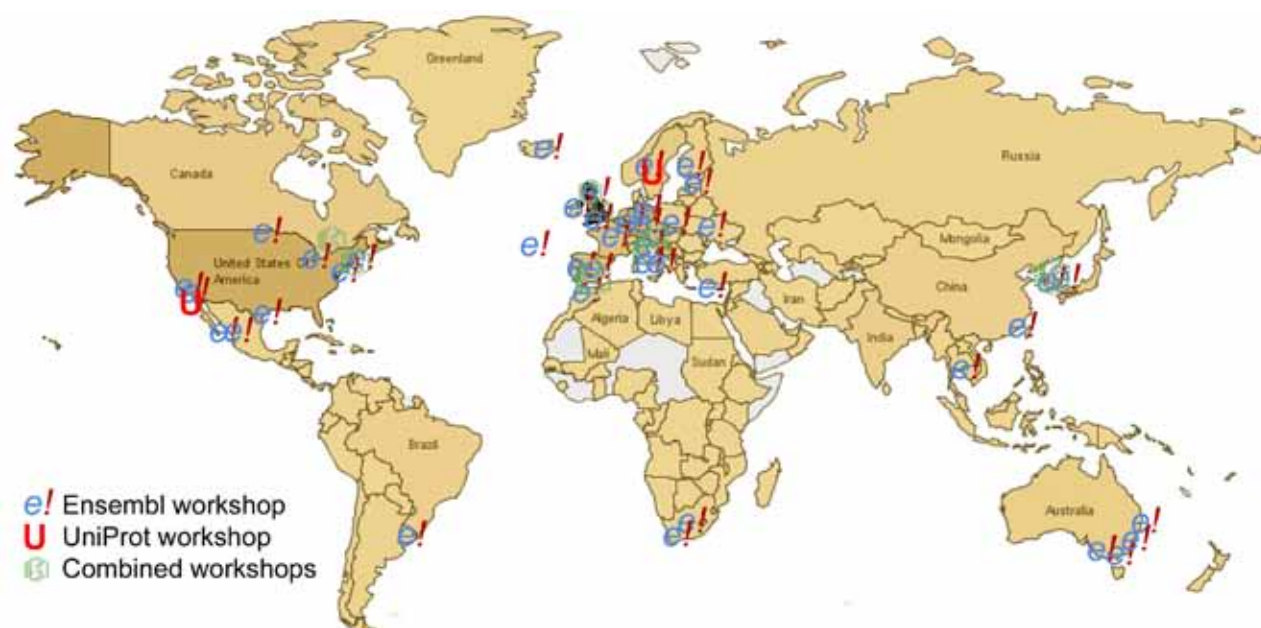


Figure 1. Heatmap: countries are coloured according to access to Ensembl and EBI resources, with the UK, USA, Germany, Spain, France, Netherlands, Italy, Sweden Japan and Switzerland in the Top 10, respectively. Locations hosting training activities between July 2009 and June 2010 are overlaid.

PANDA PROTEINS (Rolf Apweiler)

STRATEGY

The activities of the PANDA proteins teams centre on providing public access to all known protein sequences and functional information about these proteins. The UniProt resource provides the centrepiece for these activities. Most of the UniProt sequence data is derived from translation of nucleotide sequences provided by the ENA and Ensembl. All UniProt data undergoes annotation with Gene Ontology (GO) terms (see Maria-Jesus Martin, page 40, and Claire O'Donovan, page 44) and uses the classification into protein families and domains provided by InterPro (see Sarah Hunter, page 30). We add information extracted from the scientific literature and curator-evaluated computational analysis whenever possible. The combined InterPro and literature annotation forms the basis for automatic approaches to annotating all the sequence data without experimental functional data. Protein-interaction and -identification data is provided to UniProt by the IntAct protein-protein interaction database and by the Protein Identification (PRIDE) database (see Henning Hermjakob, page 28).

The UniProt, Gene Ontology Annotation, InterPro and Proteomics services are described separately by their own team leaders. The rest of this section details the RESID project, which directly reports to Rolf Apweiler.

RESID

John S. Garavelli

The RESID Database of Protein Modifications (Garavelli, 2004) 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, PDB, structure diagrams and molecular models. RESID documents the controlled vocabulary for natural protein modifications in the feature-table annotations of UniProtKB and supplements the modification descriptions with more detailed information. It was used during the initial phase of the UniProt project in merging the feature annotations of Swiss-Prot and the PIR, and in designing new standard annotations. In September 2010, quarterly release 63 contained 532 entries for chemically unique protein modifications.

Information-retrieval projects bring to light original reports for new types of modifications and for newly found modifications in additional proteins; the gathered 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. As an internet resource, RESID helps high-throughput proteomics researchers to find monoisotopic masses and mass differences, to identify known and predicted protein modifications and to suggest the modified sequences from alternative isobaric peptides that are the most consistent with current knowledge of natural modifications. RESID is a contributing component of the Proteomics Standards Initiative ontology of protein modifications (PSI-MOD), maintained by John Garavelli for the PSI and PRIDE.

OUTREACH AND TRAINING (Rolf Apweiler and Ewan Birney)

Outreach

Xosé M. Fernández, Jeff Almeida-King, Bert Overduin, Michael Schuster, Giulietta Spudich

Delivering quality workshops remains a priority within the PANDA group, and this effort enhances EMBL-EBI's training (i.e. the EBI hands-on and roadshow series). We have extended our workshops to serve more countries: 26 countries hosted PANDA training events during the reporting period (see Figure 1). This year, we conducted a series of workshops and seminars in several cities throughout Australia: Adelaide, Brisbane, Canberra, Melbourne and Sydney. We conducted our first-ever training events in Morocco, Poland, Slovenia, South Korea and Thailand. In total, we participated in 156 courses (see Table 1).

Table 1. PANDA courses, June 2009 – July 2010.

Resource	Number of Courses
Ensembl (VectorBase, Ensembl Genomes)	93
Proteomics (UniProt, InterPro, IntAct, PRIDE)	23
Sequence databases (ENA, GOA)	7
Pathways (Reactome, ChEBI)	7
EMBL-EBI (Roadshows and Hinxton-based workshops)	26

Trainee Programme

As part of EMBL-EBI's training mission, the PANDA group runs an active trainee programme. Undergraduates and postgraduate (usually Marie Curie fellows) join PANDA for a period of 3 to 12 months, applying their theoretical knowledge to practical problems. This year the group has hosted 12 trainees and 7 visitors (see Table 2), who worked on a broad variety of projects. Several project outputs have become part of EMBL-EBI's production process and resulted in co-authorships in publications and conference contributions.

Table 2. Trainee Programme participants

Trainee	Host institution	Focus/Team
Jigisha Anupama	Indian Institute of Science, Bangalore, India	ChEBI
Laura Daniels	University of California, Davis, USA	Proteomics Services
André Fauré	University of Cape Town, South Africa	Ensembl
Mirko Ferraiolo	Università degli Studi di Napoli, Italy	ChEMBL
Gavin Ha	University of British Columbia, Canada	Ensembl
Kalaivanii Jayaseelan	Vellore Institute of Technology, India	ChEBI
Duan Lian	East China University of Science and Technology, Shanghai, China	ChEBI
Thomas Maurel		Vertebrate Genomics
Nelson Ndegwa	University of Manchester, UK	Reactome, Proteomics Services
Laurence Newman	Post-secondary student	IntAct, Proteomics Services
Eric Pfeifferberger	Austria	IntAct, Proteomics Services
Andreas Schoenegger	FH Hagenberg, Austria	PRIDE, Proteomics Services
Sander Timmer	University of Amsterdam, the Netherlands	Analysis of 1000 Genomes data, Vertebrate Genomics
Jose Maria Villaveces Parda	Universidad de la Sabana, Colombia	DASTY Protein DAS client, Proteomics Services

APWEILER RESEARCH

Rolf Apweiler is currently supervising one PhD student, Joe Foster. Another PhD student working under his supervision, Garth Ilsley, successfully concluded his work and submitted his thesis during the reporting period.

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

Single Reaction Monitoring (SRM) has been utilised in the small molecules field for over 30 years, and has more recently been used for proteomics. While the majority of high-abundance proteins are readily characterised by SRM, the difficulties lie with selecting likely transitions for proteins of low abundance. It is understood in the field that detection of peptides of low abundance proteins by mass spectrometry is thwarted by peptides of high-abundance proteins that can competitively interfere with the ionisation and detection of less abundant peptides. Using predicted retention time information of peptides derived from low abundance proteins of interest, and comparing these predictions to the actual Total Ion Chromatograms of the sample being measured, transition selection can be performed in context of the actual sample background. As a result, peptide candidates can be selected that represent the best chance for detection in the mass spectrometer, giving the experimentalist optimal conditions to successfully identify and quantify proteins of interest.

Garth Ilsley: Modelling gene regulation in *Drosophila* development using quantitative *in situ* hybridisation data

The idea of morphogen gradients encoding positional information for a developing organism has long been discussed in the field of developmental biology, but quantitative models that relate measured transcription factor concentrations to enhancer activity have been proposed only recently. This is largely thanks to recent methods that combine *in situ* hybridisation with image analysis to produce quantitative data at cellular resolution. One example is the Berkeley *Drosophila* Transcription Network Project, which includes the gene expression levels of almost 100 genes in approximately 6000 nuclei over a key period of development. We developed a statistical model from these data that recovers known regulatory relationships and suggests new ones. Importantly, it shows that transcription factor concentrations alone are sufficient for encoding positional information in the early *Drosophila* embryo. The best-fitting models are suggestive of an underlying *cis*-regulatory module structure. Further work building on these results will be both computational and experimental.

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. Ewan Birney is currently supervising three PhD students: Markus Hsi-Yang Fritz, Dace Ruklisa and Sander Timmer. The group also includes one joint EIPOD postdoc, Mikhail Spivakov.

Markus Hsi-Yang Fritz: Hominid segmental duplications and repeat evolution

We 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 both low-coverage data (e.g. the 1000 Genomes data) and other resources (e.g. Neanderthal information).

Dace Ruklisa: Large-scale association studies: from inference framework to effects

We have been using more involved statistical models to explore the relationship between genotype and phenotype. Using *Drosophila* oocytes, we have converted in situ images of 40 different isogenic lines into a complex array of 114 different phenotypes. We can use these phenotypes together with the known genotypes of these individuals to find associations. By carefully separating the components of variance specific to an individual (as opposed to a strain) we can discover a variety of robust genotype-to-phenotype associations. In human data, we explored statistical subsampling techniques to generate richer, multi-SNP models for associations.

Sander Timmer: Association studies in the development of *Drosophila* and humans

We are looking at genetic variation in fundamental processes of *Drosophila* development, in particular gene expression in the early embryo and in human skeleton development. To that end, we are working with the Furlong laboratory at EMBL Heidelberg to perform a large-scale expression quantitative trait loci (eQTL) study focused on early development, and obtaining robust measurements of human skeletal phenotypes using MRI scans.

Mikhail Spivakov: EIPOD project – *Drosophila* mesoderm development

Using ChIP-chip – and more recently ChIP-seq – from *Drosophila* mesoderm at a variety of time points, we are studying how the *Drosophila* developmental system specifies organogenesis. We have started to analyse the effects of individual genetic variation on developmental *cis*-regulatory networks in the fly. The distribution of variation in *Drosophila* is varied (as expected) and includes variation in fundamental genetic components, including development. In addition to clear evidence of negative selection on developmentally important motifs, there are intriguing signals of potential positive or balancing selection. This is a joint project with the Furlong laboratory at EMBL Heidelberg.

FUTURE PROJECTS AND GOALS

We intend to improve integration and synchronisation of all PANDA resources. In addition to major improvements of our current systems, we will add mining of high-throughput genomics and proteomics datasets to our automatic annotation toolset. Despite the abundance of data from large-scale experimentation on a genome-wide level (e.g. expression profiling, protein–protein interaction screens, protein localisation), the systematic and integrated use of this type of information for high-throughput annotation of proteins remains largely unexplored. We therefore intend to build on on-going research activities at EMBL-EBI to develop and assess new protocols to integrate and analyse functional genomics datasets for the purpose of high-throughput annotation of uncharacterised proteins. This will include: analysing different data types regarding their suitability for the approach; developing data structures that allow the efficient integration and mining of data of different types and quality; benchmarking the obtained results; and applying the new methodologies to UniProtKB/TrEMBL annotation.

SELECTED REFERENCES

- Apweiler, R. (2009) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38 (Database issue), D142-D148.
- Flicek, P., et al. (2009) Ensembl's 10th year. *Nucleic Acids Res.* 38 (Database issue), D557-562.







Alvis Bramza

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

Functional Genomics

INTRODUCTION

The Functional Genomics Team includes the teams led by Misha Kapushesky, Helen Parkinson, Ugis Sarkans and a number of staff reporting directly to Alvis Brazma. The teams focus on functional genomics data services; research in functional genomics data analysis; algorithms and methods; and research and development related to biomedical informatics. We run one of the EBI's core resources, ArrayExpress, which comprises the Archive of Functional Genomics Data and the Gene Expression Atlas. We have released a new core resource: the BioSample Database, which eventually will hold information about all samples and phenotypes deposited in any of the core databases at EBI. Our PhD students focus on analysing functional genomics data, building models for systems biology and developing new methods and algorithms. Integration of data across multiple platforms, including genotypes, is another important area of activity. We also substantially contribute to training in transcriptomics and the general use of EBI tools.

GENE EXPRESSION ATLAS (Misha Kapushesky, pp. 32)

The Gene Expression Atlas, part of the ArrayExpress infrastructure, allows users to query gene expression by gene names or properties (e.g. gene ontology terms) or by tissue types, cell types, disease states or other conditions where the genes are expressed (Kapushesky et al., 2010).

- Released an open-source, stand-alone version of Atlas software, accompanied by several data releases (more than 5500 experiments loaded into the Atlas for 16 species, including microRNA expression and RNA-seq data sets) – the software has been adopted by several external groups.
- Launched the R Cloud Service, providing RNA-Seq processing via the ArrayExpressHTS R package as well as a general-purpose environment for statistical analysis.



ARRAYEXPRESS PRODUCTION (Helen Parkinson, pp. 48)

The Functional Genomics Production Team manages data content, user interaction for the core EBI databases, the ArrayExpress Archive, the Gene Expression Atlas and the new BioSamples database.

- Loaded over 400 sequencing-based functional-genomics experiments into ArrayExpress, a data-loading pipeline established with the European Nucleotide Archive and the European Genome-phenome Archive.
- Curated over 2500 experiments in the Gene Expression Atlas and issued regular releases of Experimental Factor Ontology.

ARRAYEXPRESS DEVELOPMENT (Ugis Sarkans, pp. 52)

The software development team builds and maintains several major components of the ArrayExpress infrastructure.

- Built a new ArrayExpress data management infrastructure (to enter service in late 2010).
- Maintained the existing infrastructure, facilitating the 10-fold growth of ArrayExpress archive over the past four years.
- Significantly evolved the ArrayExpress Archive user interface to provide a more robust and richer service.

BIOSAMPLE DATABASE & BIOMEDICAL INFORMATICS

Julio Fernandez Bannet, Marco Brandizi, Mike Gostev, Maria Krestyaninova, Eamonn Maguire, Helen Parkinson, Philippe Rocca-Serra, Susanna-Assunta Sansone, Ugis Sarkans, Nataliya Sklyar

The BioSamples Database is designed to contain information about biological samples used in experiments (e.g. sequencing, genotyping, gene expression, proteomics, metabolomics), the data from which are stored in other EBI databases. BioSample users can reference a particular set of biological samples (e.g. commonly used cell lines or mouse strains) from many different experiments. These biosamples can be either actual physical samples (e.g. blood) or sources of such samples (e.g. cell lines, animal strains or anonymised human individuals). The data will be exchanged between databases at EBI and NCBI such that all biosample information will appear in both resources; the space of the accession numbers will be shared between the two databases. The main development work was done by Mike Gostev; however, it also involves the ArrayExpress Development and Production teams and has benefited enormously from earlier work on the SIMBioMS system, led by Maria Krestyaninova, and work on the ISA infrastructure, led by Susanna Sansone.

- Released a prototype of the BioSample Database in May 2010 and a public version in September 2010.

RESEARCH

Nils Gehlenborg, Angela Gonzales, Misha Kapushesky, Margus Lukk, Helen Parkinson, Gabriela Rustici, Holly Zhang-Bradley

Our on-going research projects are related to regulation of gene expression and analysis of large-scale functional-genomics data. The focus is on understanding how gene expression depends on molecular regulatory mechanisms as well as on genetic and experimental factors. We are also interested in transcriptomic/genomic associations with human diseases. In particular, we are interested in integrative approaches that draw on the vast amounts of public data collected in ArrayExpress and other EBI resources.

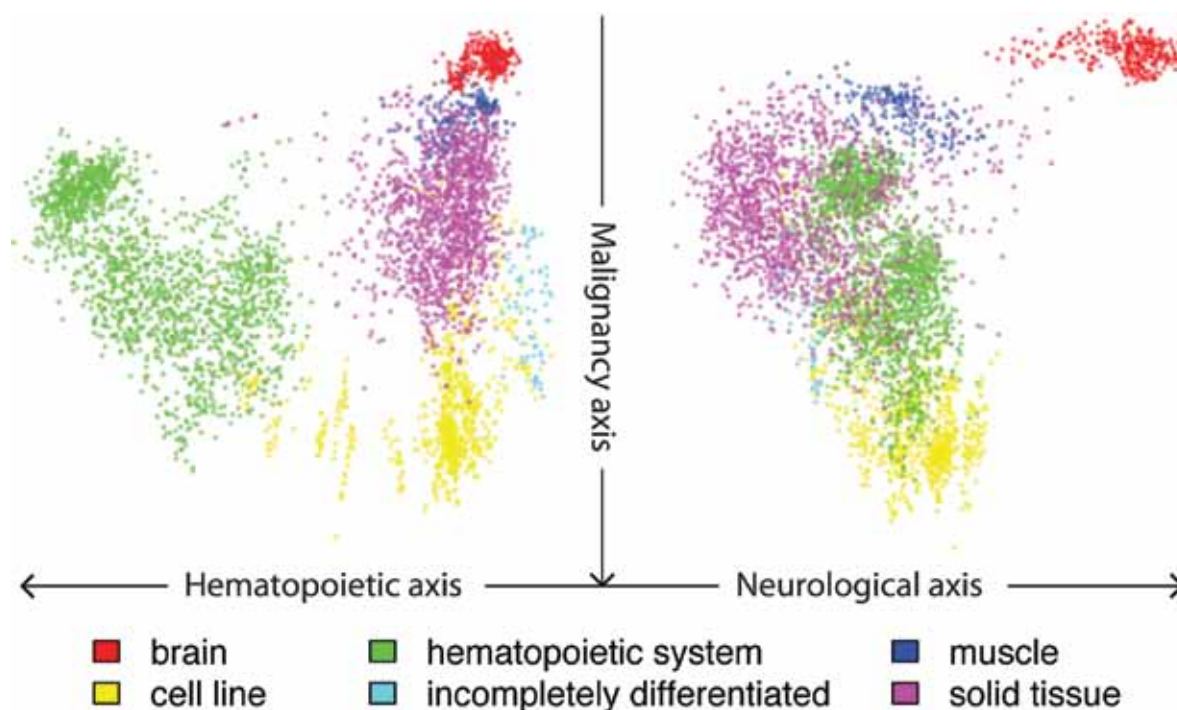


Figure. The 5372 samples are shown as dots, colour-coded for the six major clusters identified by comparing gene-expression profiles. The left and right panels are projections of the same 3D shape viewed from two different perspectives.

- Published the Human Expression Map (Lukk et al., 2010), which drew on data collected from 163 laboratories worldwide involving 5372 human samples from various tissues, cell types and diseases;
- Completed the Mouse Expression Map, analysed across more than 1000 samples and compared with the human map (Zheng-Bradley et al., submitted);
- Published research on genome-wide association to disease and genetic regulation of gene expression, in particular for type 2 diabetes (Rung et al., 2009);
- Published research on gene regulation in the fission yeast *Schizosaccharomyces pombe* (Aligianni et al., 2009) in collaboration with Jürg Bähler's group at University College London;
- Developed an R/Bioconductor package, ArrayExpressHTS, for pipeline analysis of RNAseq data from raw data to estimated transcript levels (manuscript in preparation);
- Submission of PhD thesis: Nils Gehlenborg's work on visualisation and pattern detection in large-scale biological datasets;
- Completion of PhD thesis: Katherine Lawler's work on transcriptional and post-transcriptional regulation of gene expression in fungal species;
- Published a study of CD8+ T cell transcription signatures (which predict prognosis in autoimmune disease) showing that transcriptional profiling of purified CD8(+) T cells identifies two distinct subject subgroups, predicting long-term prognosis in two autoimmune diseases (McKinney et al., 2010).

TRAINING

Tomasz Adamusiak, Ibrahim Emam, Angela Goncalves, Emma Hastings, Misha Kapushesky, Margus Lukk, Gabriella Rustici, Eleanor Williams

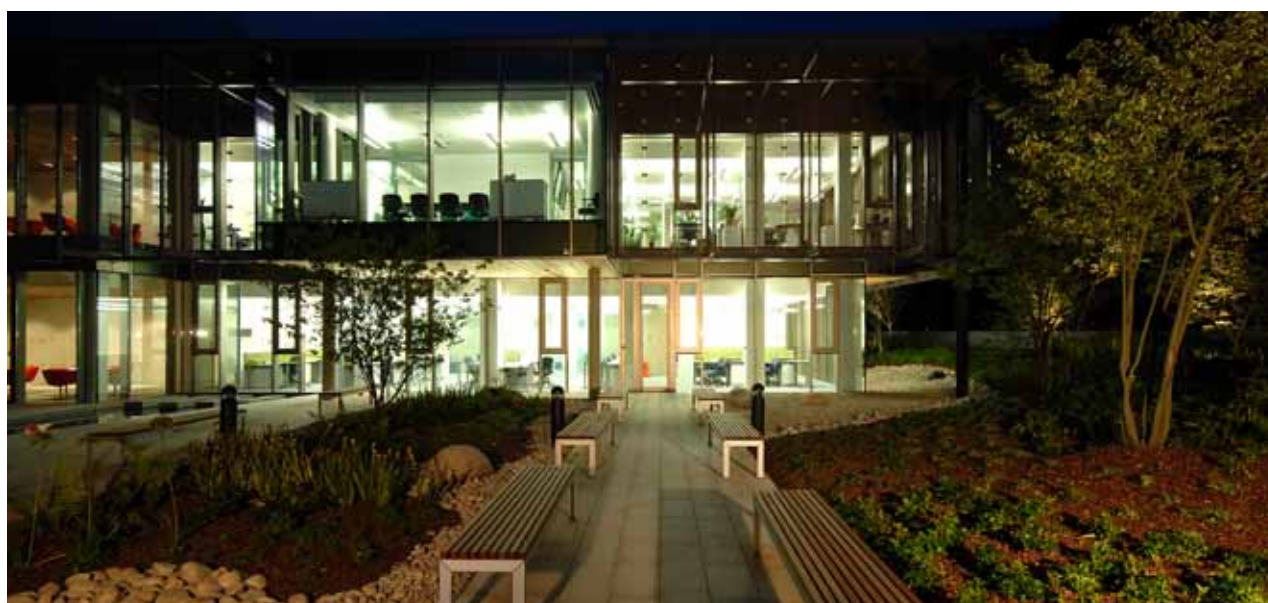
- Organised and participated in over 30 training events;
- Secured funding to repeat the EMBO course on analysis and informatics of transcriptomics data, one of the most successful training workshops at EBI in the past year;
- Contributed to the development of several training resources, including the EBI's forthcoming eLearning portal.

FUTURE PROJECTS AND GOALS

In 2011 we will work on populating the BioSample Database with all data from EBI assay databases, accepting reference layer datasets and securing data exchange with NCBI. We also plan to increase the throughput of the sequencing-based functional-genomics data, including RNAseq datasets into the Expression Atlas. We will continue our work with medically relevant collaborative projects. Further research into integrative data analysis will concentrate on using next-generation sequencing data, integrating genotype and gene expression data and building systems biology models.

SELECTED REFERENCES

- Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Database issue), D690-D698.
- Lukk, M., et al. (2010) A global map of human gene expression. *Nature Biotechnol.* 28, 322-324.
- McKinney, E.F., et al. (2010) A Cd8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nat. Med.* 16, 586-591.
- Zheng-Bradley, X., et al. (2010) Large-scale comparison of global gene expression patterns in human and mouse. *Genome Biol.* (submitted).







Guy Cochrane

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

The European Nucleotide Archive

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 – including sequence similarity search, text search, browsing and rich integration with data resources beyond ENA – provided both over the web and through an increasingly sophisticated programmatic interface. All ENA services are supported with helpdesk services and a growing training capacity. These services are focused towards 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 an on-going concentration on data presentation, integration (within ENA and with resources beyond it), tools and services development, the team's commitment is to the utility of ENA content and the broadest reach of sequencing applications.

SUMMARY OF PROGRESS

- Enhanced support for raw data obtained from both novel sequencing platforms and newly developed methodologies for existing platforms;
- Provided engineering improvements to sequence similarity search service;
- Analysed next generation sequencing trends and continued to work on projecting growth and developing strategic responses;
- Engineered core technologies to serve data to users, including data indexing and delivery services, Java object model development, XSLT and Google web toolkit development;
- Improved user interfaces and rolled out enhanced functionality;
- Developed improved and more robust validation technology for incoming data;
- Established training course and developed training materials;
- Worked to build the ENA community by running workshops and delivering training.

MAJOR ACHIEVEMENTS

In May 2010 we launched the European Nucleotide Archive, which brings together the EBI's existing and newly developed nucleotide sequence repositories into an integrated service. Underlying resources include the long-established EMBL-Bank database for nucleotide sequence and functional annotation; the Trace Archive for raw data from capillary sequencing platforms, which was established a decade ago at the Wellcome Trust Sanger Institute and is now maintained at the EBI as part of ENA; and the newly established Sequence Read Archive for raw data from next generation platforms. Services to users included in the launch cover data submission, presentation and analysis tools.

The cornerstone of the service is the ENA Browser, which was first made available in beta late in 2009 and was put into full production as part of the ENA launch. Offering intuitive and integrated access across all ENA holdings – from raw data to assembled sequence and annotation – browser functionalities include graphical assembly and feature annotation views; a navigation box that supports browsing into related ENA records and into records in third-party resources; and options to view and download data in a number of formats. Alongside the interactive browser service are programmatic services, including REST support for appropriate ENA data and access to large datasets through FTP and Aspera channels.

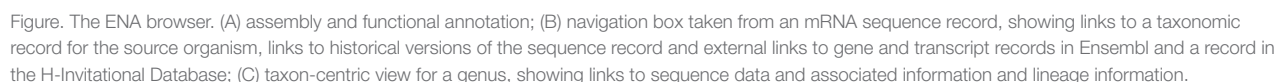
Access to and analysis of data through new search services were also featured in the ENA launch. Specifically, we have provided a single point of entry for accession retrieval across all ENA data types and have built text-based search functions

Following the ENA launch, we delivered a public prototype of the Taxonomy Portal, which provides a taxonomic point of entry into ENA data. This service provides instant access to all ENA data holdings for a given taxon and, optionally, all taxa below it.

In the coming year, we will develop extensively the ENA's data presentation services, including the browser, our programmatic services and the internal data indexing structures that feed these services. This will require work at the level of the underlying technologies (e.g. sequence similarity search) and the provision of new services and data portals, such as support for read alignments and high-level assembly services. We will take strong direction from the growing user base as to where we should focus these improvements. Our submissions services will be enhanced in a number of ways. We will develop an interactive submissions tool for smaller-scale next generation sequencing studies tailored to the needs of the smaller biology laboratory submitting community. We will develop support for third-party validation of incoming next generation sequence data to provide better integration with such submission services as that operated by ArrayExpress. Finally, we will work with our international partners to provide better coverage of genome assembly information to support the genome browser community. The many challenges and opportunities brought by next generation sequencing technologies will continue to influence our strategic thinking. Technical developments will certainly continue to be required to support the aggressive growth of data from these platforms. In addition, we will heighten our response to the ever-increasing penetrance of sequencing as a general assay platform that has arisen from rapidly falling costs; here, we will deepen our use of our model in which domain-specific services (e.g. submission and presentation tools) are delivered collaboratively on top of more generic core repository services that are maintained solely by the ENA team.

Leinonen, R., et al. (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.* 38, D39-D45.

Shumway, M., Cochrane, G. and Sugawara, H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.* 38, D870-D871.





Paul Flicek

DSc Washington University, 2004, Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008. At EMBL-EBI since 2005, Team Leader since 2008.

Vertebrate Genomics

DESCRIPTION OF SERVICES AND RESEARCH

We are a combined service and research team focusing on: genome annotation, archiving and distributing variation data, creating and deploying large-scale bioinformatics infrastructure, computational epigenomics and comparative regulatory genomics. Our major service projects include the Ensembl project, the European Genome-phenome Archive (EGA), EBI's mouse informatics team, the DGVa database of structural and copy-number variation and the data-management activities of the 1000 Genomes Project.

SUMMARY OF PROGRESS

- Issued four releases of Ensembl, incorporating the annotation of the newest version of the human genome assembly as well as comprehensive updates across the project;
- Launched the DGVa database for copy-number and structural variation;
- Completed the 1000 Genomes pilot project and released all of the resulting data prior to publication;
- Published 34 papers in peer-reviewed journals, spanning the breadth of the research and service sides of the group.

MAJOR ACHIEVEMENTS

Services

The Ensembl project creates high-quality resources for chordate genomes and provides this information in a consistent and accessible infrastructure centred on the Ensembl genome browser. Over the past year we released four comprehensive updates including the genome annotation of the updated human genome GRCh37 assembly, which provided extensive community resources for comparative genomics, variation, genome regulation and bioinformatics infrastructure. We introduced the first of a series of new Ensembl interactive tools to facilitate genome analysis and, to maximise the value and reuse of our software and data, we published a collection of Ensembl methods papers in the open-access literature.

The mouse informatics team released the CREATE project website as a central collection of all known Cre recombinase driver lines required for tissue/cell-specific and temporal control of mouse gene knockouts produced by the International Knockout Mouse Consortium (IKMC). These data, and others representing specific mouse phenotypes, are searchable via federated queries using the BioMart platform.

The EGA database has expanded and now contains data for more than 80 studies. Over 1500 researchers have registered for and been granted access to one or more EGA data sets. In early 2010, we launched the DGVarchive, a repository database for copy-number and structural variation data. The DGVa is a peer archive with the dbVar database at NCBI, and works in collaboration with the Database of Genomic Variants (DGV) project at the University of Toronto to produce curated reference sets of structural-variation data.

The 1000 Genomes Project, a flagship international project, continued to build its comprehensive catalogue of all common human variation. Over the past year the project scope expanded to a planned 2500 samples sequenced by the end of 2012. We completed the analysis of the data from the pilot phase of the project, released all pilot project data openly to the scientific community in advance of publication and publicly released data from the 700+ samples making up the main project, with more data being deposited at the EBI every day.

The emergence of new DNA sequencing technologies has led to rapid discoveries in human variation. To facilitate the sharing and understanding of these data we have been involved in the creation of relevant data standards; specifically, the development of the Locus Reference Genomic (LRG) standard for the stable reporting of genome-variation data associated with diagnostic data and locus-specific databases, and the Genome Variation Format (GVF) standard for the representation and annotation of variation data. Both standards were published in the open-access literature in 2010.

Research

The team addressed questions related to the evolution of transcriptional regulation, cell-type specificity and epigenetic phenomena. The past year has seen two major results. A project jointly led by Petra Schwalie from EMBL-EBI and Dominic Schmidt from Duncan Odom's laboratory at the University of Cambridge described a tissue-specific role for non-CTCF-associated cohesin sites in the genome. This result, demonstrated in MCF7 breast cancer cells and HepG2 liver-derived cells, suggests that the cohesin protein complex associates with tissue-specific transcription factors to ensure proper conformation of DNA in promoter regions regulated by tissue-specific master regulators (see Figure).

A second collaborative project with the Odom lab, jointly led by Benoit Ballester (EMBL-EBI), Dominic Schmidt and Michael Wilson (both at the University of Cambridge), described experimentally for the first time the evolutionary dynamics of transcription-factor binding. This was done by mapping the DNA-binding locations of two transcription factors in the livers of animals from across vertebrate evolution, from chicken to human. The results showed a surprisingly small number of shared sites through evolution, and demonstrated that the majority of all binding sites are lineage specific in the species tested. This challenges long-held opinions of the roles of conserved sequence in the genome and explains recent single-species results showing that many transcription-factor binding sites do not exhibit sequence conservation.

FUTURE PLANS

Over the next year we plan to expand our support of high-throughput DNA sequence data being generated in a number of international projects. The full 1000 Genomes Project will sequence nearly 2000 individuals by mid-2011. We will continue to be the primary data managers for this project and will incorporate the results into the Ensembl resources. Additional sequence data from the US-led ENCODE project and similar data from the International Human Epigenome Consortium will form the core of Ensembl's regulatory and functional annotation of the genome. The International Cancer Genome Consortium (ICGC) plans to deposit all data generated in Europe, Canada and Asia in the EGA and we will develop new methods to provide secure distribution this data to cancer researchers. Our research projects will continue to explore the evolution of transcriptional regulation, with experiments focused on relatively short evolutionary time scales for both human and other model organisms.

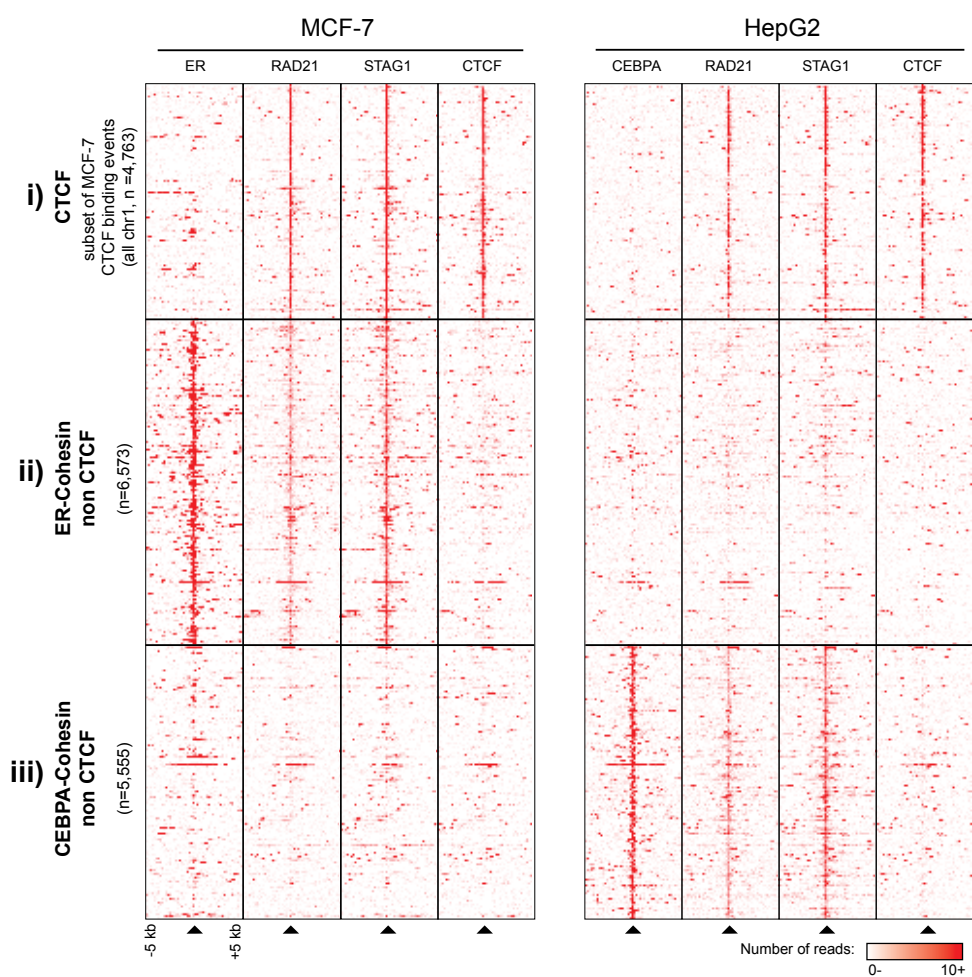


Figure. Cohesin binding events associated with CTCF are cell-type invariant (A), while cohesin locations not associated with CTCF binding are associated with tissue-specific master regulators (B, MCF7 breast cancer cells and oestrogen receptor; C, HepG2 liver cells and CEBPA). [From Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., et al. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* 20:578-588.]



Henning Hermjakob

Dipl. Inf (MSc) in Bioinformatics, University of Bielefeld, 1996. Research Assistant at the National Centre for Biotechnology (GBF), Braunschweig, Transfac Database team. At EMBL-EBI since 1997.

Proteomics Services

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 Virtual Physiological Human 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 European projects, including Apo-Sys, LipidomicNet, SLING, ENFIN and ProteomeBinders.

SUMMARY OF PROGRESS

- In the context of the HUPO Proteomics Standards Initiative (PSI), contributed to a series of community standard documents on gel electrophoresis (Hoogland et al, 2010, Gibson et al, 2010), protein-affinity reagents (Bourbeillon et al, 2010, Gloriam et al, 2010) and mass spectrometry (Martens et al, 2010);
- Co-authored a series of publications with third parties, focussing on use of PSD data in external data analysis and analysis tools (Perreau et al., 2010, Antonov et al., 2010, Lee et al., 2010);
- Oversaw the growth of the Proteomics Identifications Database (PRIDE) to more than 100 million mass spectra;
- Started the FP7 project RICORDO (coordinated by PST), which aims to integrate models and ontologies related to medical physiology and human anatomy – also contributed substantially to the EU-funded Virtual Physiological Human (VPH) project.

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 curated to common standards. InAct 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, USA), MINT (University of Rome, Italy), MatrixDB (University of Lyon, France), Molecular Connections Inc. and MPIDB (J. Craig Venter Institute, USA).

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 released in 2010, providing access to more than 15 million binary interaction evidences from 16 different sources, including protein–protein interactions, protein–small molecule interactions, and simplified pathway data (Aranda et al., submitted).

IMEX and PSICQUIC are supported by an EU grant, PSIMEX, which the Proteomics Services Team coordinates. With the PSI molecular interaction formats well established and widely used in the community, the IMEX consortium in production mode and the widespread adoption of the PSICQUIC interface, we have achieved the core target of the PSIMEX grant: the global coordination and integration of major molecular interaction data resources.

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 is under negotiation (June 2010).

Following a comprehensive review in 2009, we completely redeveloped the Reactome web site. The new Reactome (in beta testing as of June 2010) provides interactive pathway diagrams for all Reactome pathways, improved pathway-analysis tools and close integration with molecular interaction data via the PSICQUIC interface.

FUTURE PLANS

After rapid development and achievement of major milestones in the molecular interaction domain, we now need to consolidate the 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 is the complete redevelopment of the PRIDE database, necessary to cope with the rapid increase in data content but also to turn 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 the 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 REFERENCES

Martens, L., et al. (2010) mzML - a community standard for mass spectrometry data. *Mol. Cell. Proteomics* (in press). DOI: 10.1074/mcp.R110.000133.

Perreau, V.M., et al. (2010) A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease. *Proteomics* 10, 2377-2395.

Lee, K., et al. Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *Plant Cell*. 22, 997-1005.

Gloriam, D.E., et al. (2010) A community standard format for the representation of protein affinity reagents. *Mol Cell Proteomics* 9, 1-10.

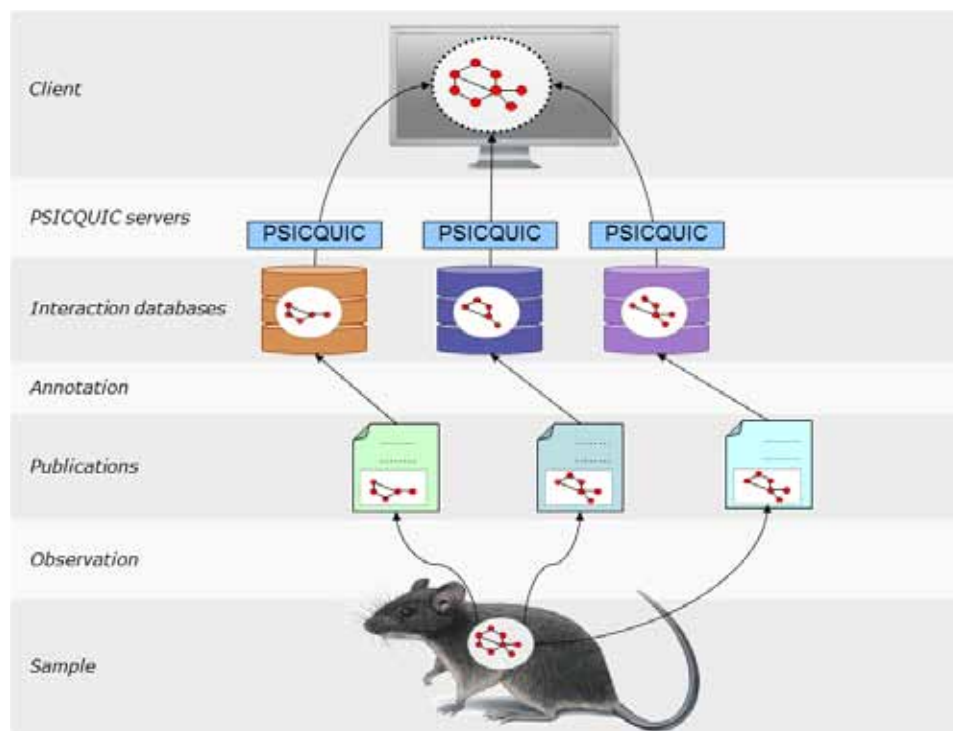


Figure. PSICQUIC and IMEX support a truly distributed provision of molecular interaction data: an experimental system is observed in multiple independent studies, resulting in multiple publications. Based on the collaboration in the International Molecular Exchange Consortium (IMEX), independent interaction databases curated these publications in work-sharing mode. Data is then released in PSI format through the PSICQUIC interface. A web client queries all PSICQUIC services and integrates the data on the client side.



Sarah Hunter

MSc. University of Manchester, 1999. At EMBL-EBI since 2005.

InterPro

DESCRIPTION OF SERVICES

The InterPro team coordinates the InterPro, CluSTr 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 the major protein signature databases into a single resource, and currently includes data from Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSE, SUPERFAMILY, CATH-Gene3D, PANTHER and HAMAP. During the integration process, InterPro rationalises instances where more than one protein signature describes the same protein family/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 table and graphical views and the InterPro Domain Architectures view.

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. The CluSTr project aims to cluster all UniProtKB proteins and protein sets from complete genomes. The resulting clusters and similarity scores are accessible via a web interface. The new metagenomics portal is intended to provide metagenomics researchers with access to EBI's functional analysis pipelines, links to data archives (e.g. the ENA) and a web interface to manage and visualise these data.

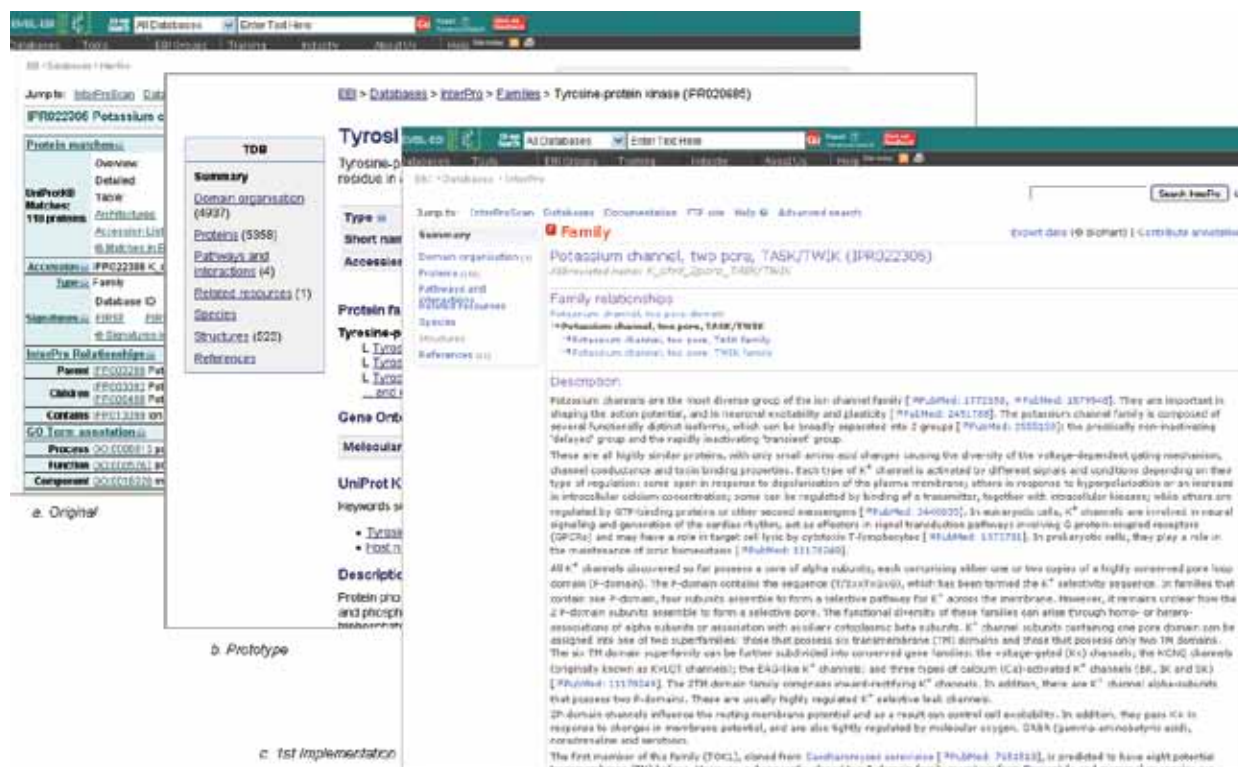


Figure. Evolution of the new designs for the InterPro web interface, from the original site (a) through to user-guided prototyping (b) and an initial implementation (c). We intend to carry out further user testing and refinement of the design during 2011.

SUMMARY OF PROGRESS

- InterPro - new data: issued seven major releases of the database, consisting of 2033 new entries incorporating 2401 signatures; also, also added new data sources to the web interface, including links to the PRIAM enzyme resource and Pfam-B (automatically generated HMMs from Pfam);
- InterPro - new interfaces: provided a new DAS source for InterPro, serving InterPro matches to UniProt and UniParc protein sequences, also extended the Biomart to include UniParc data;
- InterPro - new software: released InterProScan v4.6 (signature scanning software) and beta version of InterProScan v5 to a small numbers of testers;
- InterPro - migrated to HMMER3 for Pfam searches.
- CluSTr - added new data sets for model organism databases (TAIR and SGD);
- CluSTr - upgraded to Paralgn algorithm.

MAJOR ACHIEVEMENTS

InterPro has fulfilled its goal of releasing its data to the public in sync with alternate UniProtKB releases. Because of this, the data provided to users is more consistent and available more frequently. In the course of seven major releases, we integrated over 2400 signatures from our member databases. The last of the releases in this period (v27.0) covers 95.2% of UniProtKB/Swiss-Prot proteins, 77.7% of UniProtKB/TrEMBL and 78.5% overall for UniProtKB (Swiss-Prot and TrEMBL).

During the latter part of 2009, the Pfam database began using the HMMER3 search algorithm, which is far faster than its predecessor, to search its models against sequences. As a consequence of the algorithm's increased sensitivity to more remote homologs, significant additional curation efforts went into ensuring that the data already integrated into the InterPro database was still valid and correct. A welcome additional consequence of the move to using HMMER3 is that we are now able to calculate matches against Pfam-B. Pfam-B models are automatically generated from the ADDA database and provide coverage of protein space that the manually curated Pfam-A models do not. These Pfam-B matches are now visible in the InterPro web interface. In addition, data from the PRIAM resource was linked from relevant InterPro entries, providing more information about enzyme families. The InterProScan software package (available for download and for use via EBI services) has been updated to utilise HMMER3 and to reflect changes in signature scanning algorithms (the February 2010 release is 4.6). InterPro data is now also available to users via a number of interfaces. The existing BioMart (Smedley *et al.*, 2009), which allows users to create complex queries against InterPro data and download the results in a variety of formats, has been expanded to include UniParc proteins. The MyDAS technology has been used to develop a new DAS service for InterPro data on top of the BioMart. InterPro data available in this manner includes matches to both UniProtKB and UniParc proteins.

CluSTr had a single release during the reporting period, providing 3 636 831 744 similarities for 9 450 285 UniProtKB sequences (v15.6) and 17 616 060 clusters. The Smith-Waterman algorithm utilised by CluSTr was also updated (to Paralgn 5.0). Some additional data sets for model organisms (TAIR and SGD) were added to the protein sequences and similarities and clusters made available for them.

A metagenome represents the DNA complement of an entire community of organisms existing either in the environment or on a host species. Thanks to the advent of High-Throughput Sequencing (HTS) technologies, these organisms can now be characterised at the nucleotide level; however, the data produced is large in volume and often fragmented. EMBL-EBI aims to provide a comprehensive resource for metagenomic sequence analysis; the project to build it was begun in late 2009. The major work thus far has been to investigate the state-of-the-art algorithms available for performing analysis on these kinds of data and to piece together a pipeline to be used by researchers in characterising their metagenomics data.

To ensure that our services are reaching their target users, the InterPro team organised and actively participated in 11 different roadshows, workshops and training events in Europe and the US during the reporting period.

FUTURE PLANS

A new version of InterProScan, which has an entirely different Java- and database-based architecture at its core, is almost ready for public release. We anticipate this will happen in early 2011. We have added improved algorithms for transmembrane helix and signal peptide prediction to the software, as well as making it more modular and robust. The InterPro website is also being redeveloped to make it more intuitive for users. We have carried out prototyping and usability testing on the new designs and a first release of the interface will occur in late 2010. The focus now for the metagenomics portal is the design and implementation of a web interface to allow users who have submitted their data for analysis and archiving to be able to manage and interact with their data. We hope to have a basic interface available by the end of 2010, with further improvements and enhancements – including better submission and visualisation tools – throughout 2011.



Misha Kapushesky, PhD

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

Functional Genomics Atlas

DESCRIPTION OF SERVICES/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 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. Through collaborations with the European Nucleotide Archive (ENA) group, we have developed support for storing next generation sequencing studies in the Atlas. 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

- Released an open-source, stand-alone version of Expression Atlas software (accompanied by several data releases) that has been adopted by many external groups;
- Developed several novel visual and analytical features in the Atlas, including anatomogram displays of gene expression patterns (see Figure);
- Loaded more than 5500 experiments into the Atlas for 16 species, including microRNA expression and RNA-seq data sets and comprising nearly 140 000 samples, using the R Cloud framework for on-the-fly data pre-processing;
- Launched the R Cloud Service, providing RNA-Seq processing via the ArrayExpressHTS R package as well as a general-purpose environment for statistical analysis.

MAJOR ACHIEVEMENTS

Our main task is to create an infrastructure for organising and querying multiomics data. The Gene Expression Atlas, which can be applied to diverse transcriptomics data in the public domain, is just the beginning. During the reporting period, considerable effort went into making the Atlas available for stand-alone installations. This resulted in the deployment of a new infrastructure able to accommodate high data volumes and faster throughput in curation processes: the number of experiments loaded into the Atlas grew rapidly from about 1000 in January 2010 to more than 5500 in October 2010. Since then, we have focused our efforts on improving support for data types other than microarray transcription profiling, namely RNA-seq and proteomics.

Atlas Software and data releases occur monthly. The software, which comes with a set of interfaces for data loading and processing, is frequently downloaded from our GitHub site. Within four months of its release, several major academic organisations and pharmaceutical companies have installed and are using our freely available, stand-alone Atlas. The software is used not only to run the public EBI Atlas but also serves to manage data for the SYBARIS project.

Following the launch of the Expression Atlas, we successfully integrated several novel features in the software. These include the ArrayExpressHTS pipeline (*Bioinformatics*, under review) for processing RNA-seq data integrated with the ENA; the inclusion of a global map of a human gene expression dataset (Lukk et al., 2010) in the Atlas interface with special analytics; and the development of anatomogram visualisations and advanced data filters. The new Atlas web interface benefits from integration with the Experimental Factor Ontology (EFO; Malone et al., 2010), which is used to improve queries, result displays and underlying statistical analyses. EFO releases, developed by the Functional Genomics team, are timed to coincide with Atlas data releases, ensuring rich annotation to the latest version of the ontology.

The R Cloud, launched at the MGED 2010 conference in August 2010, powers distributed computations in the Atlas. We are now offering it as a limited service for external users. Within the R Cloud users can pre-process datasets with the RNA-seq pipeline and submit the results to ArrayExpress Archive. Our group manages public mirrors of R software archives CRAN and Bioconductor. The R Workbench provides a complete development environment for R, including a syntax-highlighting editor, console, graphics device and help browser. The R Cloud serves as a comprehensive platform for collaborative data analysis.

The end of 2009 marked the kick-off of SYBARIS, an FP7 collaborative research project coordinated by the Atlas Team. In SYBARIS, eight European partners are studying the interaction of fungal pathogens and the human and mouse immune systems with the aim of elucidating markers of antifungal drug resistance and understanding genetic components of disease susceptibility. The Atlas team is responsible for overall project management as well as for data acquisition and meta-analysis. In collaboration with SYBARIS partners we submitted for publication a study of cryptic mistranslation in yeast as well as a broad review of systems biology approaches to infectious disease.

FUTURE PLANS

Stand-alone Atlas deployment leads naturally towards the development of a distributed, federated query model. We have built a first prototype of the Distributed Atlas and plan to expand on this project, integrating diverse multiomics data types starting with transcriptomics, proteomics and eventually adding metabolomics datasets. Building on the analytics back-end of the Atlas, we plan to extend it to support complex, multifactorial experiment designs, gene set enrichment analysis with published gene signatures and a sample-based similarity search across Atlas experiments. R Cloud will be further integrated with the Atlas to promote easier online data sharing, processing and analysis. Next generation sequencing data support features prominently in Atlas plans, as well as integration with other EBI resources (e.g. Ensembl, ENA, EGA). Together with the Rebholz Group, we are also developing semantic web features (e.g. RDF export) for the Atlas.

SELECTED REFERENCES

Goncalves, A., et al. (2010) ArrayExpressHTS: distributed computing for RNA-seq data processing and quality assessment. *Bioinformatics* (submitted).

Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Suppl 1), D690-698.

Lukk, M., et al. (2010) A global map of human gene expression. *Nat. Biotechnol.* 28, 322-324.

Malone, J., et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26, 1112-1118.

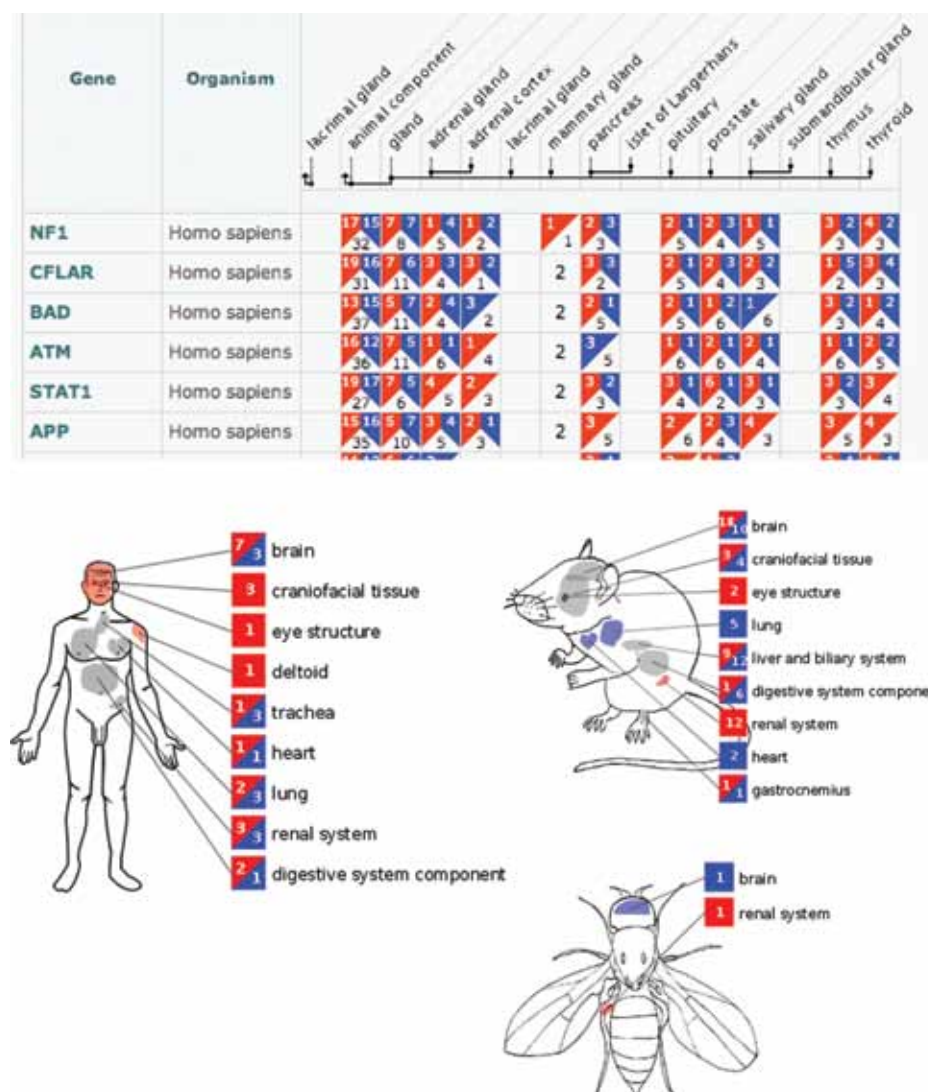


Figure. One year after the launch of the Gene Expression Atlas service at the EBI, we open-sourced the stand-alone Atlas software used to run it. Using monthly downloadable public software, data and ontology releases, users can install the complete Atlas system locally and use it to load and view private datasets together with the public data. The Atlas includes the global map of human gene expression as well as a pipeline for processing RNA-seq studies. The R Cloud service, launched in summer 2010, provides remote access to Atlas data and the R cloud-computing environment on EBI servers.



Paul Kersey

PhD University of
Edinburgh 1995. At
EMBL since 1999. Team
Leader since 2008.

Ensembl Genomes

DESCRIPTION OF SERVICES/RESEARCH

The Ensembl Genomes team is responsible for providing services based on the genomes of non-vertebrate species. The falling costs of DNA sequencing (for deciphering unknown sequences and assays of known sequences) have led to an explosion of reference genome sequences and genome-wide measurements and interpretation. Ensembl Genomes (Kersey, P.J. et al, 2010) provides five portals (for bacteria, protists, fungi, plants and invertebrate metazoa) offering access to these data through a set of programmatic and interactive interfaces, which were originally developed in the context of the (vertebrate-focused) Ensembl project. Collectively, the two projects span the taxonomic space.

The development of next generation sequencing technologies has led to the performance of complex and highly data-generative experiments, now performed even in species studied only by small communities with little informatics infrastructure. Through collaborating with the EBI and re-using our established toolset, such small communities can store, analyse and disseminate data more cheaply and powerfully. Our leading collaborators include VectorBase (Lawson et al. 2009), a resource focused on the annotation of invertebrate vectors (the EBI is a direct participant); WormBase; and Gramene. 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.

Our interest in bacteria has led us to become involved in the development of Microme, a new resource for bacterial metabolic pathways.

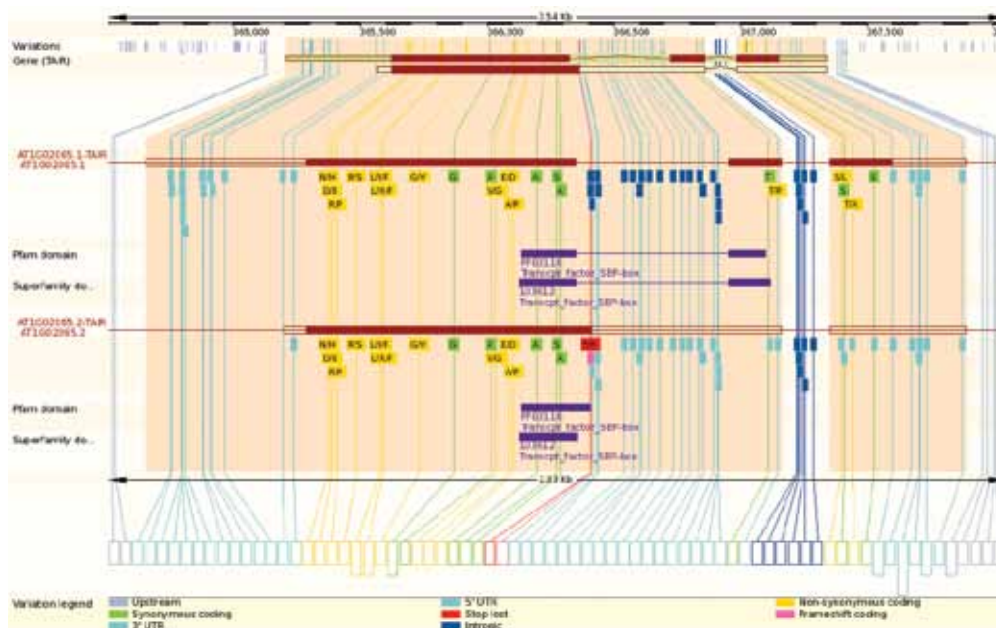
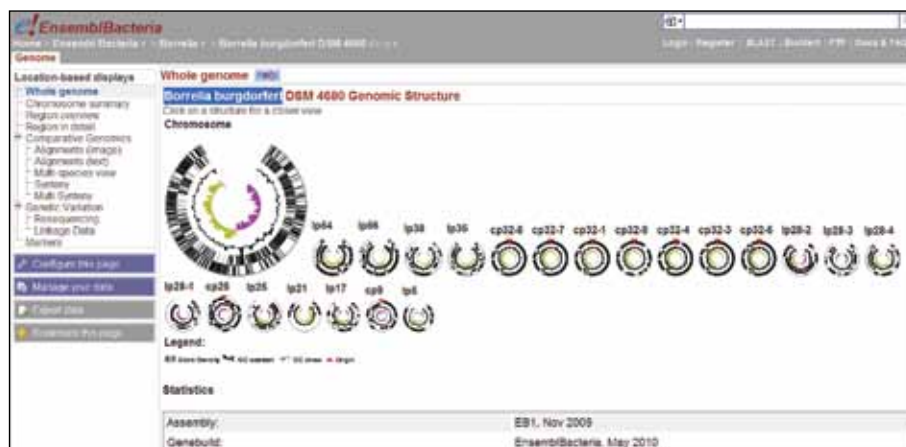


Figure. Ensembl variation image from the SPB8 gene of *Arabidopsis thaliana* (encoding a squamosa promoter binding protein-box domain protein). The data set is constructed from three programs to measure population wide variation, two of which are part of a program to completely sequence the genomes of 1001 individuals of this species (108 of which are currently available); the third is a chip-based assay of 214 000 known single-nucleotide polymorphisms (SNPs) in a further 931 genomes. These efforts have so far identified over 7 million SNPs, over 500 000 insertion/deletion loci and over 200 million individual genotypes. The Ensembl variation infrastructure includes support for visualising variation loci (and their effects on genes) in their genomic contexts (illustrated), individual sequence and links to phenotype. A data-mining tool for variation data, using the BioMart data warehousing tool, is also available.



Ensembl Genomes covers a wide diversity of species whose genomes comes in many different forms. In the past year we have implemented a fully functional circular browser, allowing users to navigate through the non-linear chromosomes and plasmids typically found in bacteria.

SUMMARY OF PROGRESS

- Launch of Ensembl Plants and Ensembl Fungi in September 2009, completing the set of five Ensembl Genomes portals;
- Introduction of 6 new genomes to Ensembl Protists (*Dictyostelium discoides*, 2 diatoms and 3 *Plasmodia*), 23 new genomes to Ensembl Metazoa, and 67 new genomes to Ensembl Bacteria;
- Development of new visualisations of bacterial genomes and multi-way comparative analysis;
- Release of databases for genomic variation for *Arabidopsis thaliana*, two strains of rice, grape, mosquito, fruit fly and yeast;
- Publication of the genome sequence and annotation of the body louse *Pediculus humanus*;
- Commencement of work on two new EU-funded projects: MICROME (focused on bacterial pathways) and INFRAVEC (focused on mosquito variation).

MAJOR ACHIEVEMENTS

The launch of the Ensembl Plants and Ensembl Fungi portals in September 2009 completed the set of five Ensembl Genomes portals (Ensembl Bacteria, Protists and Metazoa having been launched a few months earlier). In total, we have made four releases during the reporting period, expanding the number of species available through each of the portals and improving the depth of data and the quality of the associated visualisations. New genomes included that of the human body louse, *Pediculus humanus*, which was annotated by the team (Kirkness, E.F., *et al.* 2010). Areas of particular focus have included the development of variation resources, especially for plants; our collaboration with colleagues at the Universities of Oxford and Bath, who have been using next generation sequencing technologies to sample variation in the model plant *Arabidopsis thaliana*, has driven our development in this increasingly important area. We have also improved the visualisation of bacterial genomes, providing a true representation of circular chromosomes, and are developing new tools for multi-way genomic comparison. The incorporation of RNA-seq data for the *Aedes* and *Anopheles* mosquitoes within our databases is a landmark achievement, symbolic of the way that new data types are increasingly driving new developments in storage and visualisation. Another impact of NGS technologies on the team's activities has been the public release of Curtain, a new environment for sequence assembly we have developed in response to the growing needs of our collaborators to assemble and annotate genomes.

FUTURE PLANS

In the next year, we should further increase the number of genomes included in Ensembl Genomes and plug the remaining gaps in our taxonomic coverage. In particular, we have recently been awarded funding to establish PomBase, a new resource for the fission yeast genome, and PhytoPath, a resource focused on phytopathogen genomes; we have also received funding to work on the genome and variome of the wheat rust pathogen. In each case, we have partnered with leading research groups who bring their knowledge of the biological domain to our collaborations. 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. The public launch of Microme is also expected before the end of 2011.

SELECTED REFERENCES

- Kersey, P.J., et al. (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.* 38 (Database issue), D563-D569.
- Lawson, D., et al. (2009) VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res.* 37 (Database issue), D583-D587.
- Kirkness, E.F., Haas, B.J., Sun, W., et al. (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Nat Acad Sci USA* 107, 12168-12173.



Gerard Kleywegt

Team Leader

PhD University of Utrecht, 1991. Postdoctoral research, University of Uppsala. Coordinator then Programme Director of the Swedish Structural Biology Network, 1996-2009. Appointed Professor of Structural Molecular Biology, University of Uppsala, 2009. Team Leader at EMBL-EBI since 2009.



Tom Oldfield

Technical Team Leader

PhD University of York, 1990. Postdoctoral research at the University of York, 1990-1993. Principal Scientist, Accelrys, Inc., 1993-2002. Database manager, MSD, EMBL-EBI, 2002-2010. At EMBL-EBI since 2010.

The Protein Data Bank in Europe (PDBe)

DESCRIPTION OF SERVICES

The Protein Data Bank in Europe (PDBe) is one of the six core molecular databases hosted by the EMBL-EBI. PDBe is also 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 RCSB and BMRB in the United States and PDBj in Japan. PDBe is a deposition and annotation site for the two major databases containing biomacromolecular structure data: the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB). Whereas the PDB is maintained by the wwPDB partners, EMDB is a joint venture between PDBe, RCSB and Baylor College of Medicine. The work of PDBe, wwPDB and EMDB is guided by scientific advisory committees (one for each organisation) that meet annually.

The major goal of PDBe is to provide integrated resources of structural data that evolve with the needs of biologists. To that end, PDBe endeavours to: handle deposition and annotation of structural data expertly as a wwPDB and EMDB deposition site; 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, Electron Microscopy).

SUMMARY OF PROGRESS

- Processed a total of 1150 depositions to the PDB and 124 depositions to EMDB;
- Completely redesigned the website, PDBe.org, taking a more user- and problem-oriented approach – a newly developed Wizard helps novice users find the information they need;
- Launched a series of browsers of the structural archive based on biologically intuitive classifications;
- Launched PDBprints, which provides graphical summaries of PDB entries;
- Dedicated considerable efforts to training PDBe users, running several roadshows for depositors across Europe;
- Raised the profile of PDBe in the community through various communication and training activities.

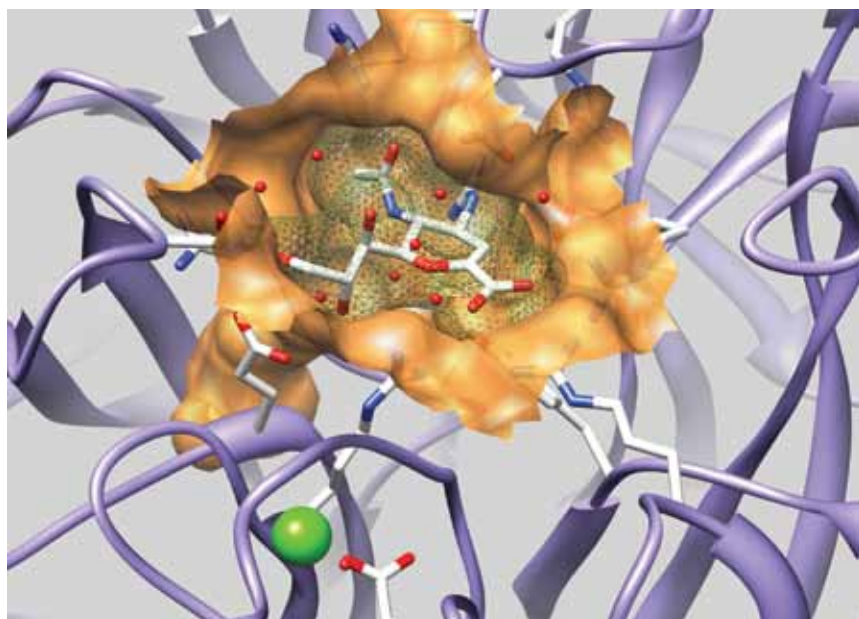


Figure. Details of the drug-binding site in the X-ray crystal structure of the complex of the 1918 influenza virus H1N1 neuraminidase and Zanamivir (PDB entry 3B7E). The development of Zanamivir was aided substantially by knowledge of the three-dimensional structure of its target, neuraminidase. The compound is currently marketed by GlaxoSmithKline under the trade name Relenza.

MAJOR ACHIEVEMENTS

The PDBe team changed dramatically at the onset of the reporting period when Professor Gerard Kleywegt took over from Dr Kim Henrick and many staff members reached their nine-year contract limit. After familiarising himself with the team, its operations and its many collaborations, Gerard initiated and oversaw the introduction of new services and a complete overhaul of the website. Although the turnover presented some challenges, the new composition of the group provided an opportunity to shift the focus in new directions. The appointment in 2010 of Dr Tom Oldfield as a technical team leader with responsibility for the PDBe databases and services further strengthened the leadership of the team.

In addition to processing 1150 depositions to the PDB, the team is collaborating with partners in the US and Japan to create a common tool for handling deposition and annotation of structural data – using any technique or combination of techniques – by all wwPDB and EMDB partners. One major outcome was the development of a prototype module to handle sequence information, a substantial effort that required much of the underlying code and protocols to be in place. The new tool will be used by all deposition sites from early 2012. The wwPDB partners also undertook a new round of remediation of the PDB archive, focusing on biological assembly data, residual B-values and peptide-like inhibitors and antibiotics. As the legacy PDB format can no longer properly describe large entries, a draft specification of a new and more flexible format was completed in June 2010. Validation task forces were set up to advise wwPDB how best to assess the quality of deposited X-ray and NMR structures; PDBe will carry out the implementation of the recommendations. Another task force was established to advise wwPDB regarding the possible inclusion in the PDB of models based on small-angle scattering data.

The joint EMDB portal, EMDatabank.org, was redesigned and a new full-text search implemented. The maps currently held by EMDB were remediated, which led to a much improved archive and a better match between maps and fitted models. In addition, the visualisation program OpenAstexViewer was extended to cope with EM maps, tomograms and masks. We are working with the community and relevant journals to make deposition of EM maps and models mandatory for publication. An EM-specific validation task force was established and had its first meeting.

Our redesign of the PDBe website – and the addition of features specifically geared to non-expert users – has taken the service up a notch. We launched a series of structural archive browsers that use biologically intuitive classifications such as EC, Pfam and CATH. We also launched PDBprints, a service that provides graphical summaries of PDB entries that can be incorporated on webpages and are used by PDBe on Atlas pages as well as in lists of search results.

We collaborate with a number of academic groups specialising in NMR, which has informed the development of new tools and led to several publications. The deposition software was modified to support the mandatory deposition of chemical-shift data. These relationships have also helped us to reorganise and vastly expand the content of PDBe's NMR-specific web pages.

The PDBe team independently organises and runs several outreach and training activities. During the reporting period we held nine roadshows for depositors in Finland, Germany, Slovenia, Sweden and the UK. Team members regularly presented at professional meetings both in the UK and abroad (21 lectures, 8 posters) and published nine papers in the scientific literature. We also issue news items routinely (particularly since the launch of the new website) to structural biology-related mailing lists and bulletin boards, and have established a PDBe presence in social media.

FUTURE PLANS

We have several ambitious goals for the coming years: improving the image and awareness of PDBe in the biomedical community; becoming the logical first stop on any quest for structural information; and transforming the structural archive into a truly useful resource for biomedical and related disciplines. We will focus on the development of new services, tools and resources in our five strongest or most promising areas: refinement of our advanced services such as PDBePISA, PDBeFold, PDBeMotif and the new browsers; annotation, validation and visualisation of ligand data; integration with other resources; validation and presentation of information about the quality and reliability of structural data; and exposing experimental data in ways that help experts and non-experts alike understand the extent to which it supports the structural models produced. New resources, tools and services as well as improvements to existing ones will be released in six-month cycles.

SELECTED REFERENCES

Velankar, S., et al. (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 38, D308-D317.

Berman, H.M., et al. (2010) Safeguarding the integrity of protein archive. *Nature* 462, 425.



Jane Lomax

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

The Gene Ontology Editorial Office

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 assimilate and disseminate annotation data; and 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 GO Editorial Office plays a key role in managing the distributed task of developing and maintaining the GO vocabularies. We contribute to a number of other GO project efforts, including web presence, software testing, user support and education.

SUMMARY OF PROGRESS

- Added logical definitions (a.k.a. cross-products) to GO for the first time, enabling the development of a tool for automatic term addition;
- Progressed toward alignment with and creation of cross-products to an external chemical ontology;
- Developed ontology content in the areas of signalling, transcription and kidney development;
- Established relationships between the biological process and molecular function ontologies.

MAJOR ACHIEVEMENTS

The Gene Ontology is dynamic: existing terms and relationships are augmented, refined, and reorganised as biological knowledge advances. Major improvements have been made over the lifetime of the GO project in several areas of the ontology, usually in consultation with experts in relevant subject areas. Table 1 shows the size (as of November 2010) of each of the four ontologies maintained by the GO Consortium.

Table 1. Status of the GO vocabularies as of November 2010

Total GO Terms	32935
Molecular Function Terms	8893
Cellular Component Terms	2771
Biological Process Terms	19819

Significant changes introduced to GO in 2010 affect both biological and logical aspects of the ontologies: logical definitions (also known as 'cross-products') have been added to GO for the first time, which has allowed for the development of a tool for automatic term addition. We have made much progress toward aligning with and creating cross-products to an external chemical ontology. We have also developed ontology content in the areas of signalling, transcription and kidney development. In addition, relationships have been established between the biological process and molecular function ontologies.

We made considerable progress toward creating cross-products for GO terms (Mungall et al., 2010). These definitions help improve computability and support more sophisticated tool development. Our work has concentrated on 'internal' cross-products, i.e. those that define GO terms by referring to other GO terms. The first set of cross-products – between regulatory processes and regulated processes or functions – were added to the GO file in early 2010. Subsequently, two further sets have been added: biological processes involved in other biological processes, and cellular components that are part of other cellular components.

As a result of these changes, we developed a tool – TermGenie – 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. This tool reduces the workload for ontology editors and helps reduce human error in the ontologies.

We generated cross-products to externally maintained ontologies that intersect with GO. To this end, we are active members of the OBO Foundry (Smith et al., 2007), a collaboration to establish a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. Earlier this year, GO became one of the founder sets of OBO Foundry ontologies.

The biggest effort over the past year went into aligning GO with the Chemical Entities of Biological Interest (ChEBI) ontology, with the aim of generating cross-products between GO and ChEBI. This involved a two-day meeting with the ChEBI ontology developers in September 2010 to reconcile some of the critical differences between the two ontologies. We hope the first ChEBI cross-products will be added to GO early in 2011.

GO has traditionally comprised three orthogonal ontologies, but we have been working to add relationships between these ontologies to enrich the biological representation. In 2010 we have added 'part_of' relationships between the molecular-function and biological-process ontologies. For example, we have made many transporter functions 'part_of' their corresponding transport process.

2010 has seen major improvements to the biological content of several areas of the ontologies; transcription and transcription factors; signalling; and kidney development. The changes in these areas were developed in collaboration with biological experts, often culminating in a face-to-face meeting such as the kidney development meeting, held at EMBL-EBI in January 2010.

FUTURE PLANS

The GO Editorial Office will continue to work closely with the rest of the GO Consortium and with biological experts to ensure that the ontologies are comprehensive, logically rigorous and biologically accurate. Improvements begun or continued in 2010 on signalling, kidney development and other topics will therefore continue, and we intend to start developing terms in the area of neurobiology. We will continue adding further sets of cross-products to GO, allowing us to improve TermGenie so that more routine term addition can be done automatically. This will free up editing time for more complex, biologically detailed work. These cross-product sets will include links to ChEBI, the first set of external cross-products to be added to GO. We also hope to start making cross-products to the Cell Ontology on 2011. Additional links between the biological process and molecular function ontologies will be created using new process-specific function terms.

SELECTED REFERENCES

- Mungall, C.J., et al. (2010) Cross-product extensions of the Gene Ontology. *J. Biomed. Inform.* (in press). Published online 10 February; DOI: 10.1016/j.jbi.2010.02.002.
- Smith, B., et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251-1255.

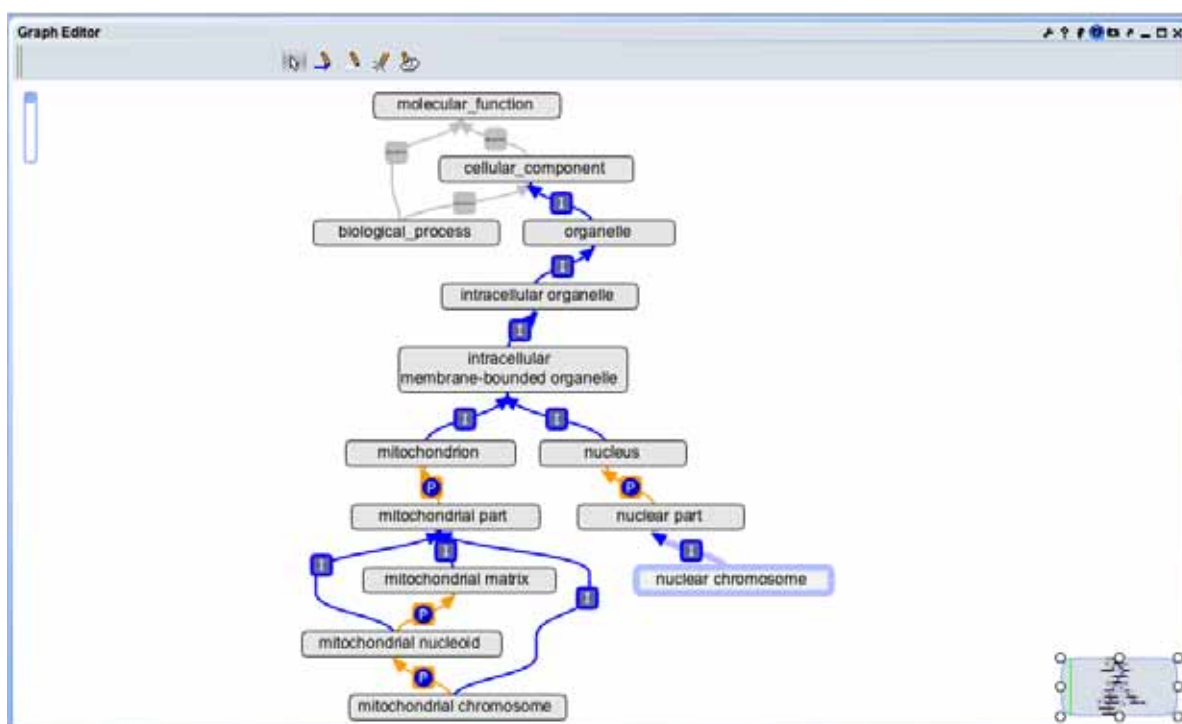


Figure. Structure of the Gene Ontology, shown in OBO Edit.



Maria J. Martin

BSc in Veterinary Medicine, Universidad Complutense, Madrid, 1990. PhD in Molecular Biology (Bioinformatics), Universidad Autonoma, Madrid, 2003. At EMBL-EBI since 1996. Technical Team Leader since 2009.

UniProt Development

DESCRIPTION OF SERVICES/RESEARCH

The Universal Protein Resource (UniProt) Development Team provides the bioinformatics infrastructure for this resource. It is also responsible for maintaining and developing tools for the UniProt Curation Team. UniProt provides the scientific community with a central repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation. It comprises four focused database layers (see Box), which incorporate, interpret, integrate and standardise data from large and diverse sources; this makes it the most comprehensive catalogue of protein sequence and functional annotation. During the reporting period, two Project Leaders were appointed: Sam Patient for software infrastructure and Alexander Fedotov for database operations.

UniProt database layers

- UniProt Knowledgebase (UniProtKB) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation;
- UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the newly expanding area of metagenomic and environmental data;
- UniProt Archive (UniParc) provides stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data;
- UniProt Reference Clusters (UniRef) provide non-redundant data collection based on the UniProt Knowledgebase and UniParc in order to obtain complete coverage of sequence space at several resolutions.

SUMMARY OF PROGRESS

- Implemented new interfaces – BioMart and Distributed Annotation System (DAS) – to serve specific user requirements;
- Distributed new data sets of interest to our user community: UniParc and proteome data set downloads;
- Implemented new annotation tools (automatic annotation, Gene Ontology and proteome editors) to support the Curation Team;
- Developed a solid data-import infrastructure to achieve consensus sequence annotation in collaboration with Ensembl, PDB and RefSeq;
- Consolidated software under a consistent Java infrastructure to accommodate and maintain growing amounts of data.

MAJOR ACHIEVEMENTS

The UniProt website provides documentation and tools to facilitate the search, identification and analysis of gene products. The team is responsible for the database release pipelines as well as for the way data is represented on the public website. Special emphasis is put on improving website usability, navigation and interfaces. Entry views have been revamped and fine-grained customisation has been substantially improved.

The team has been working on various dissemination strategies to ensure that the information in UniProt is useful to a broad user community with diverse scientific interests, requirements and levels of computational expertise. The BioMart service was developed and released to give access to UniProt data while allowing users to formulate integrated queries across related data in different resources (e.g. PRIDE, Ensembl and InterPro).

The UniProt protein DAS server has also been modified to accommodate the added functionality of the improved DAS protocol. New commands such as ‘sources’ allow DAS clients to better understand the capabilities of the UniProt protein DAS server. Error handling is now much more transparent, making it easier for DAS clients to communicate problems to its users. Annotations have become less ambiguous, with the addition of ontology tags within the UniProt DAS responses. The UniProt team has also concentrated on improving the quality of the information it supplies; changes made to the GOA data source allow users to access previously unavailable, electronically inferred gene ontology terms.

QuickGO is a fast, web-based browser that provides access to all information about GO terms and the GO annotations released by the UniProtKB-GOA group. The tool has undergone a number of major internal architectural changes that allow it to support standardised file formats (i.e. Gene Association and gp2protein files); link to external information sources; and lay the foundations for a new, more user-friendly and intuitive user interface. QuickGO's chart drawing module has been significantly enhanced so that it now displays all relationship types and inter-ontology links.

The team works to improve the accuracy and representation of the UniProt databases and services. UniParc offers a complete catalogue of protein sequences with their corresponding cross-references on a single site. UniParc is now available for text and similarity searches through the UniProt web site. We also produced downloadable files in XML and FASTA formats in response to user requests.

Complete, non-redundant proteome data sets of interest to our user community have been generated using UniProt and Ensembl data. The provision of these data sets required new import pipelines from Ensembl and RefSeq; we integrated data from the more frequently requested eukaryotic genomes (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Canis familiaris*, *Danio rerio* and *Gallus gallus*).

The UniRule system for the automatic annotation of large volumes of uncharacterised proteins is of paramount importance and represents a key development for the UniProt Consortium. To assist curators with the management of prediction rules, which are central to the annotation process, we developed and tested a user-friendly rule curation tool. To maximize computer-assisted rule development, this tool comprises components for information retrieval, model generation, automated-update monitoring, statistical assessment, rule editing and visualisation.

A proteome annotation platform is under development to support the curators in efficiently monitoring completely sequenced genomes as well as in handling annotations of their encoded proteins. With thousands of new genomes expected to be submitted to the public databases, this editor is critical for automating the tracking of genome/proteome records and the standardisation of their common annotations. In this context, the proteome editor will represent an annotation platform for the organisation and management of the information related to a set of proteins encoded by completely sequenced genomes.

UniProt curators contribute annotations to the GOA project using the web-based Protein2GO curation tool. The Development Team enhanced this tool so that it provides a variety of context-sensitive sanity checks, designed to ensure that all annotations conform to guidelines laid down by the Gene Ontology Consortium. A second mode of operation has been introduced that allows curators to work in a reference-centric fashion, rather than a protein-centric one. In this new mode, curators are able to search for (and edit) all annotations that are associated with a given PubMed or DOI reference.

FUTURE PLANS

The team plans to finish integrating the automatic annotation systems developed by the three consortium members under a single, unified database, rule-annotation tool and pipeline infrastructure. Once this has been established we will explore data exchange mechanisms to provide the annotation communities with both annotation rules and the means to annotate in our system. We will develop automatic systems to organise and visualise complete proteomes, which will allow users to have a global genome/proteome and gene-product-centric view of the sequence space from which they can drill down to the variations and annotations specific to each protein. The availability of large functional-genomics and proteomics datasets requires improved data integration approaches. The team will cooperate with diverse data providers to develop and assess new protocols for the exchange and integration of information in UniProt. We will also explore novel approaches to increasing community participation, for example by providing easy-to-use mechanisms (e.g. DAS) for making data contributions.

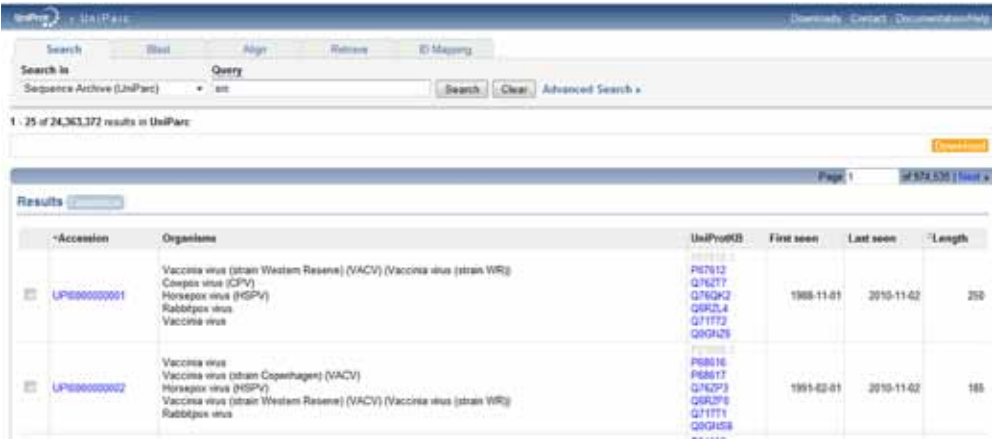
SELECTED REFERENCES

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142-D148.

Jain, E., et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10, 136.

The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38, D331-D335.

Cochrane, G., Martin, M.J. and Apweiler, R. (2010) Public data resources as a foundation for a worldwide metagenomics data infrastructure. In: *Metagenomics: theory, methods, and applications*. Marco, D., Editor. Norwich, UK: Caister Academic Press, 212 pp.



The screenshot shows the UniParc search interface. The search bar contains 'Vaccinia virus' and the results show two entries. The first entry is for 'Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))' with UniProtKB accession P07612 and a length of 250. The second entry is for 'Vaccinia virus (strain Copenhagen) (VACV) (Vaccinia virus (strain HSPV))' with UniProtKB accession P07611 and a length of 185.

Accession	Organism	UniProtKB	First seen	Last seen	Length
UP000000001	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	P07612	1988-11-01	2010-11-02	250
UP000000002	Vaccinia virus (strain Copenhagen) (VACV) (Vaccinia virus (strain HSPV))	P07611	1991-02-01	2010-11-02	185

Figure. Querying UniParc, the sequence archive.



Johanna McEntyre

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

Literature Resources

DESCRIPTION OF SERVICES

The scientific literature is a key component of the biomedical research life cycle. Providing new ways to access the literature, within the context of other biomedical data resources, will be essential as the scientific community endeavours to organise and make the best use of the flood of data promised by emerging sequencing technologies. The literature holds great promise as a force for integrating information, as it contains the formal record of the community's collective understanding of the biomedical and related sciences. Meaningful links between data resources and the literature 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.

One approach to deeper integration is to identify terms of interest within research articles and use these to link similar papers and related data resources. Text mining is a high-throughput approach to identifying biological terms in large volumes of text data, providing a basis for the development of new search and browse applications. The continued refinement of text-mining techniques, along with the growing portion of articles that are published as open access, will stimulate precise, deep linking in the future.

The goal of the 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 the citations database CiteXplore, which contains around 25 million biomedical abstracts from sources such as PubMed (from the US National Library of Medicine), Agricola (from the US National Agriculture Library), Patents (from the European Patent Office), Chinese Biological Abstracts (CAS-SICLS) and CiteSeer. The database is updated daily and links to a number of EBI data resources, including UniProt, the European Nucleotide Archive, InterPro, Intact and PDB. We also calculate citation information for the records we hold: about 9.5 million of these articles have been cited at least once; as such, ours represents one of the largest public-domain citation networks in the world.

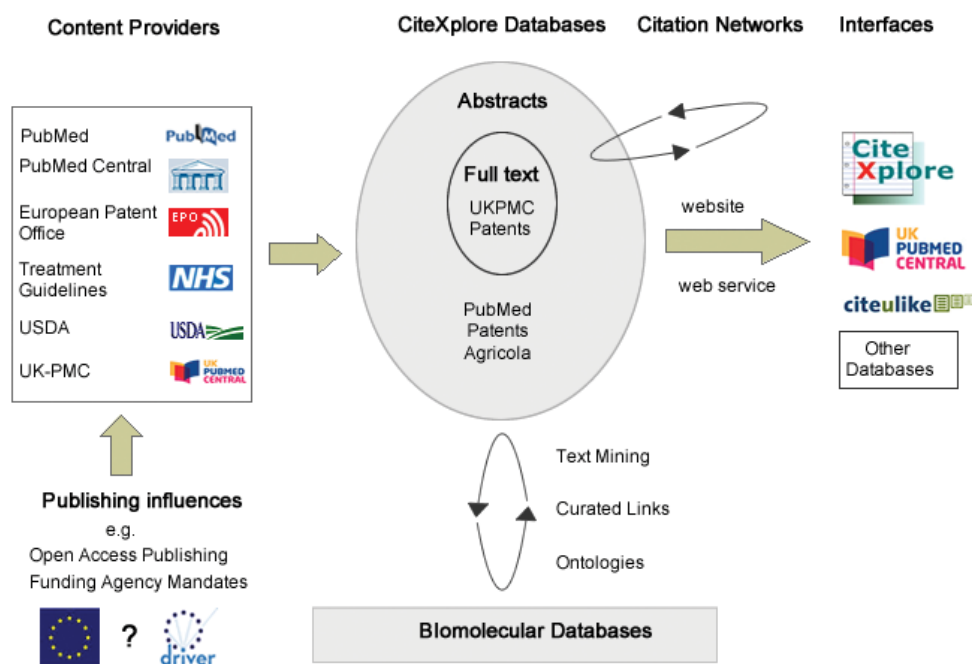


Figure 1. Overview of the activities of the Literature Services group.

Our group has been a partner in UK PubMed Central (UKPMC), a project to build a repository of life-science research articles. UKPMC was developed by EMBL-EBI, the British Library and the University of Manchester in close cooperation with PubMed Central USA, and was produced by the NCBI. UKPMC currently contains close to 2 million full-text documents. The role of our group in the project is to provide the citation and full-text search function, delivering counts, metadata and abstracts to UKPMC via a web service. Also included in the package are the links and citation information mentioned above as well as a 'key terms summary' consisting of named entities – genes, proteins, organisms, GO terms, accession numbers, diseases and chemicals – identified by the Rebholz-Schuhmann group. These data populate the web pages of UKPMC, providing the functions that set it apart from other similar resource and presenting opportunities for improved information-retrieval and knowledge-discovery functions.

SUMMARY OF PROGRESS

- Constructed UKPMC full-text index;
- Integrated text-mined biological terms into full-text content;
- Launched UKPMC website in January 2010.

MAJOR ACHIEVEMENTS

The major achievements of the Literature Services are centred on the launch of the new UKPMC website in January 2010. Moving UKPMC beyond its basic mirror function to PubMed Central, the new website is powered by the EBI Literature Web Services. The truly novel aspect of development in the project (supplied by EMBL-EBI) is the ability to search both the full text of articles in UKPMC and the citations in CiteXplore from one search box. The full-text index is enriched semantically with biological entities, providing a basis for the future development of new search approaches. Since January, the group has expanded the service to include more semantic enrichment and improved the search functions in response to user feedback.

FUTURE PLANS

Our group will build on the current programme of work, evolving UKPMC into a European-based resource that represents European science, promotes open-access publishing and provides an alternative to PubMed Central. Key to realising this vision is the development of a fast and reliable search, and the integration of the research articles with related data resources used in scientific and clinical workflows. Furthermore, engagement with the European scientific community, both directly and via existing publishing mechanisms, will help us to move towards building a public-domain content network across Europe.

Figure 2. CiteXplore and UK PubMed Central.



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. Technical Team Leader since 2009.

UniProt Content

DESCRIPTION OF SERVICES

The Universal Protein Resource (UniProt) is a collaboration of the EMBL-EBI, the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource group at Georgetown University Medical Center and the University of Delaware. Its purpose is to provide the scientific community with a single, centralised, authoritative resource for protein sequences and functional annotation. The primary mission of the consortium is to support biological research by maintaining a freely accessible, high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledge base with extensive cross-references and querying interfaces. The UniProt databases consist of four database layers optimised for different purposes (see Box).

Gene Ontology (GO) is a well-established, structured vocabulary that has been successfully used in gene product functional annotation. The UniProt–Gene Ontology Annotation (UniProtKB–GOA) database was created at the EMBL-EBI in 2001. The aim of the UniProtKB–GOA project is to provide high-quality manual and electronic annotations to the proteins stored in UniProtKB using GO vocabulary.

UniProt database layers

- UniProt Knowledgebase (UniProtKB) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation;
- UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the newly expanding area of metagenomic and environmental data;
- UniProt Archive (UniParc) provides stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data;
- UniProt Reference Clusters (UniRef) provide non-redundant data collection based on the UniProt Knowledgebase and UniParc in order to obtain complete coverage of sequence space at several resolutions.

SUMMARY OF PROGRESS

- Manual annotation has been ongoing for UniProtKB/Swiss-Prot with particular focus on the human proteome and close collaboration with other resources;
- Automatic annotation has been a key focus in the past year as the EMBL-EBI has led the consolidation of the rule systems of all participating consortium members;
- Both manual and electronic GO annotation has been ongoing, with almost 62 million GO annotations to 7.6 million UniProtKB entries, covering more than 237 000 taxonomic groups.

MAJOR ACHIEVEMENTS

UniProt is a member of the Consensus CDS (CCDS) project, a collaborative effort to identify a core set of consistently annotated and high-quality human and mouse protein-coding regions. The long-term goal is to support convergence towards a standard set of gene and protein annotations. By the end of the reporting period, UniProt had 17 512 human entries (out of 20 251 records) in synch with the RefSeq annotation group at the NCBI and the Ensembl and HAVANA teams at the EMBL-EBI and the Wellcome Trust Sanger Institute.

A highlight of the third NCBI Genome Annotation Workshop in Washington, DC in April 2010, where researchers from life science organisations worldwide collaborated to establish minimal standards for prokaryotic and viral annotation, was the development and acceptance by the community of prokaryotic protein-naming guidelines based on an initial proposal from the INSDC and UniProt. Following this agreement, INSDC and UniProt also created a more generalised protein guideline to

make this useful for taxa outside cellular prokaryotes. The decision by the INSDC to provide these guidelines for adoption by all submitters to their databases will greatly enhance the annotation of complete genomes and proteomes and ensure that the user community can exploit these data to its full potential.

The EMBL-EBI led the consolidation of its consortium members' three major automatic annotation systems, which resulted in agreement on manual rule-creation specifications along with the relevant SOP development and the setting of goals. This has led to increased coverage of UniProtKB/TrEMBL from 30% to nearly 40% and a widening of the taxonomic range and increase in the annotation depth achieved. Training in the UniRule tool is ongoing for our consortium partners.

Our curators continue to be key members of the GO Consortium Reference Genomes Initiative for the human proteome, working to provide high-quality annotations for human proteins. The GOA renal project has been very successful, providing 745 proteins with 5942 annotations; 426 new GO terms were created (1.4% of the whole of GO) as a result of a kidney ontology development workshop hosted by UniProtKB-GOA.

FUTURE PLANS

UniProt is committed to providing a 'gold standard' data set that will enable users to easily identify all experimental data for a given protein from a particular strain of a particular organism as well as all experimentally characterised annotations/proteomes from a proteome or protein family. This will involve the de-merging of existing UniProtKB/Swiss-Prot entries and an extension of the scope of the evidence-attribution system. This has the added benefit of providing experimental annotation for the benchmarking of sequence analysis tools (e.g. predictors of post-translational modifications or topology). UniProtKB/Swiss-Prot has historically 'merged' 100% identical protein sequences from different genes in the same species into a single record. As the availability and usage of genomic information has greatly increased in recent years, UniProtKB is modifying its merging policy. We will de-merge entries containing multiple individual genes coding for 100% identical protein sequences into individual UniProtKB/Swiss-Prot entries, which will give a gene-centric view of protein space and allow a cleaner, more logical mapping of gene and genomic resources to UniProtKB. We plan to extend our nomenclature collaborations to include higher-level organisms. As well as expanding the UniRule approach with additional curator resources, we will completely review the SAAS approach in the next reporting period in order to incorporate more recent developments in this field. We also expect to work closely with various internal and external resources to investigate ways of helping other institutes to use our rules, as has been requested. We intend to completely review the biology and resources behind the electronic GO annotation pipeline, with a particular focus on other UniProt-controlled vocabularies, taxonomic range and InterPro member databases.

SELECTED REFERENCES

The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142-D148.

Binns, D., et al. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 25, 3045-3046.

Alam-Faruque, Y., et al. (2010) The Renal Gene Ontology Annotation Initiative. *Organogenesis* 6, 71-75.

The Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38 (Database issue), D331-D335.

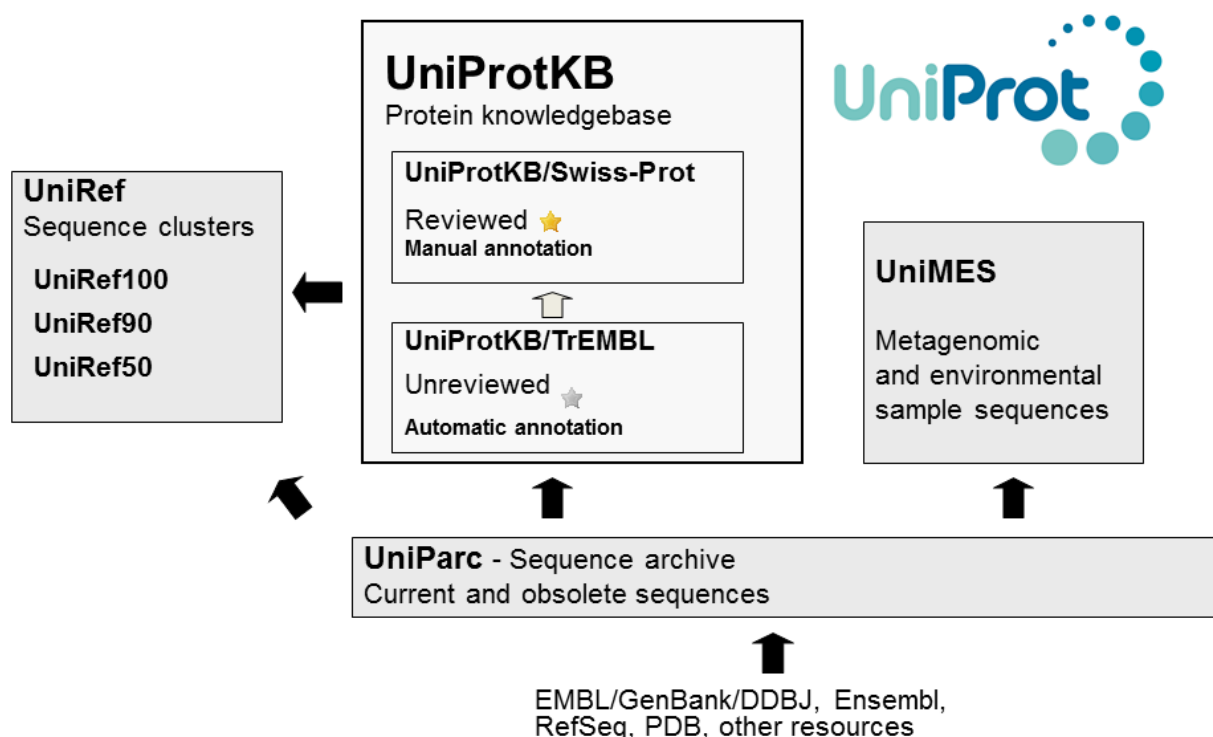


Figure. UniProt database organisation.



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.*

ChEMBL

DESCRIPTION OF SERVICES

The ChEMBL group develops and manages the EBI's database of bioactive, drug-like small molecules, which contains two-dimensional structures, calculated properties and abstracted bioactivities such as binding constants, pharmacology and ADMET data. ChEMBL data are abstracted and curated from the primary scientific literature, and cover a significant fraction of the structure-activity relationship and discovery of modern drugs. 2010 marked the first full year of staffing for the group, and has seen a number of milestones for the ChEMBL resource. Its first public release – in January 2010 – achieved broad coverage in the press and was extremely well received by the scientific community. The data is widely accessed via the web interface and via download of the entire database for local searching, and advanced tools developed by the ChEMBL team for interactive filtering and data selection provide added value to users.

SUMMARY OF PROGRESS

- Switched all ChEMBL resources to run under the secure, industry standard https: internet protocol;
- Established a robust, monthly update cycle for new data;
- Established a mechanism for the rapid upload, archival and searching of deposited datasets;
- Launched SARfari drug-discovery integration systems;
- Achieved integration of ChEMBL into other large-scale chemistry resources, including PubChem and the ChemSpider system of the Royal Society of Chemistry;
- Started to implement a fully featured and open infrastructure for large-scale scoring of targets for their 'drugability';
- Pursued research activities in two major areas.

MAJOR ACHIEVEMENTS

Usage of the resource, in particular downloads of the data, has been strong and steady. We accompanied the ChEMBL launch with a series of talks, webinars, on-campus training courses and site visits for local training. Alongside more traditional approaches to promoting resource awareness, we maintain a group blog to report on progress with the database, new drug launches and various analyses of the resource.

We switched all ChEMBL resources to run under the secure, industry standard https: internet protocol, ensuring that all traffic to our services is encrypted and secure. This is a key concern for researchers given the high confidentiality of small molecule structural data. We also established a robust monthly update cycle for new data, giving the community rapid access to new chemotype and target information. A network of specialist curators has been engaged to curate key portions of the data (e.g. ADMET data). During the reporting period the number of data records within ChEMBL grew by more than 50%.

The group established a mechanism for the rapid upload, archival and searching of deposited datasets. This year, three deposited datasets on whole-cell malaria screening – contributed by GlaxoSmithKline, Novartis and St. Jude's – featured ca. 20 000 novel compounds that are active in a relevant model of malaria infection. A specific portal, ChEMBL-NTD (Neglected Tropical Diseases), was constructed to serve this important subset of contributed data.

We went live with our SARfari drug discovery integration systems, two of which integrate data for bioassays, phylogenetic information, three-dimensional structural data and binding-site data for protein kinases and rhodopsin-like G-protein coupled receptors (GPCRs). In addition, the ChEMBL database began to be widely integrated into other large-scale chemistry resources. Of particular note is the integration of its chemical structure and bioactivity data into PubChem, as well as compound-level integration with the ChemSpider system of the Royal Society of Chemistry.

One of the areas of most immediate application for the data contained within ChEMBL is in the assessment and scoring of proteins as targets for drug discovery. We have started to implement a fully featured and open infrastructure for large-scale scoring of targets for their 'drugability'. The first released component is an analysis of properties of the binding sites for their suitability to bind drug-like molecules.

We currently have two active research areas. The first is the building of a computational system to analyse functional and binding data for peptides, and then to propose their optimisation in order to improve pharmaceutical properties, stability, affinity and selectivity. We published a paper on the analysis of ligand efficiency measures for the content of ChEMBL as well as a series of similarity maps for natural and unnatural amino acids. This work is funded under the EIPOD scheme, with the designed peptides planned for synthesis and bioassay in the lab of Maja Koehn (EMBL-Heidelberg).

The second area of research is a comprehensive analysis of 'tool compounds' or 'chemical probes'. We have assembled a number of sets of compounds that are generally considered to be chemical tools, that is, small molecules that are used to probe the function of specific proteins in either a cell or an in vivo model system. These compounds have been characterised for various properties (e.g. affinity, molecular size); approaches to predict the affinity variances across model organism species have been developed (i.e. across rat, mouse, and human orthologues).

We have participated in two significant EU-funded projects: eTox and EU-OPENSREEN. 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. EU-OPENSREEN is a large-scale infrastructure; EMBL-EBI is involved defining its data-standards (see the Steinbeck group, page 54) and database design and content (ChEMBL group). This project will provide an open-access and open-data infrastructure for screening a large compound collection and disseminating the collected data.

FUTURE PLANS

This coming year, we will release the drugability prioritisation and analysis tools, and also populate the database with biotherapeutic and clinical candidate development data. Also of high priority will be completing integration with core EMBL-EBI resources such as Ensembl, UniProt, PDB and ArrayExpress.

SELECTED REFERENCES

Abad-Zapatero, C., et al. (2010) Ligand efficiency indices for an effective mapping of chemo-biological space: the concept of an atlas-like representation. *Drug Discov. Today* 15, 804-811.

Gaulton, A. and Overington, J.P. (2010) Role of open chemical data in aiding drug discovery and design. *Future Med. Chem.* 2, 903-907.

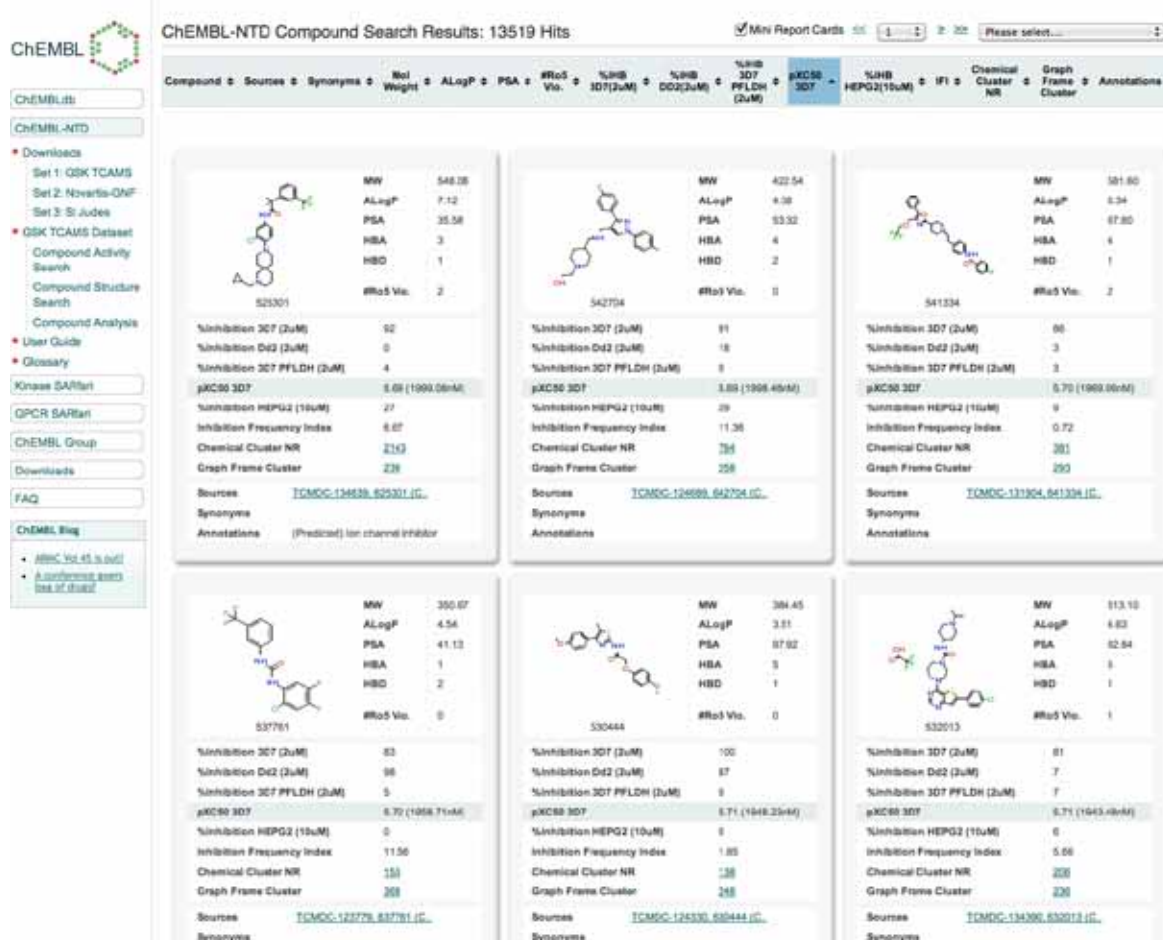


Figure. Representative screen-shot of the ChEMBL database showing flexible querying and powerful analysis routines for bioactivity data.



Helen Parkinson

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

Functional Genomics Production

DESCRIPTION OF SERVICES/RESEARCH

The Functional Genomics Production Team manages data content and user interaction for the core EBI databases: the ArrayExpress Archive (Parkinson, 2009), the Gene Expression Atlas (GXA; Kapushesky et al., 2010) and the new Biosamples Database. All three resources have complex metadata representing experimental types, variables and sample attributes for which we require semantic markup in the form of ontologies. We develop 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, the Ontology for Biomedical Investigation and the Vertebrate Anatomy Ontology (VBO). We collaborate with international partners to develop MAGE-TAB based data management infrastructure and annotation tools for gene expression data. The team has expanded its remit to deal with the change in technology from arrays to RNA sequencing experiments; this has resulted in collaboration with the EBI databases ENA and EGA to provide data flow and integration between these sequence databases and ArrayExpress.

SUMMARY OF PROGRESS

- Agreement with the Gene Expression Omnibus for data exchange of high-throughput sequencing functional genomics data;
- Monthly EFO releases (consistent over the past 28 months);
- Four open source software releases, supporting MAGE-TAB infrastructure (Limpopo and Annotare) and ontology query and lexical matching (OntoCat and Zooma).

MAJOR ACHIEVEMENTS

The main task of the group is the processing, annotation and curation of functional genomics data from direct submissions and by import from external databases. Archive software development has focussed on infrastructure development to support the submission, processing and integration of RNA-Seq data and tool development for MAGE-TAB based infrastructure and ontology development.

The EFO, an application ontology, is released monthly to support data queries in the GXA. EFO now has 3075 classes, is cross referenced to 25 public domain ontologies and has been expanded to add value to cell line terms where tissues, diseases and cell types have been added to both primary and immortal cell lines. We have also added experiment specific terms to support the query of experiments in the Archive by molecule and technology. We take a data driven approach to building the ontology in EFO, which is then used for text mining and query. EFO is mapped to public ontologies using a common, upper level ontology and relationships to promote interoperability with other semantic resources.

The production team provides open source software for data management and annotation, ontology building and lexical mapping. We released Annotare (Shankar et al., 2010), a data annotation tool supporting MAGE-TAB, jointly with colleagues in the US; Limpopo, an open source MAGE-TAB parser used by ArrayExpress and several other applications; MAGETabulator, a rule based spreadsheet generation system; as well as OntoCat, an ontology searching application, and Zooma, a lexical matching application, which jointly search and map terms to ontologies.

The team collaborates on EU- and NIH-funded research projects. For example, the EU funded GEN2PHEN project aims to unify human and model organism genetic variation databases towards increasingly holistic views into genotype to phenotype data, and to link this system with other biomedical knowledge sources via genome browser functionality. Together with project partners, we have produced an integrated data model and database for human and model organism phenotypes and are now working on tools for semantic integration of rodent model and human phenotypic data.

Tissue-specific annotation and query of multi-species functional genomic data is limited due to the lack of a homology-based, common, multi-species anatomy for mammalian species. We work with colleagues at MRC Harwell, University of Cambridge and the Phenoscope Project to generate a mammalian musculoskeletal system ontology based on homology statements. This involves aligning multiple species-specific anatomy ontologies, analysing their usage by functional genomics researchers and extracting evidence for homologous structures from the literature. We plan to extend the GXA to allow queries using these homology statements in the coming year.

FUTURE PLANS

In 2010–2011 we will work to improve the volume and quality of annotation for RNA-Seq data by working with data generating centres such as the Wellcome Trust Sanger Institute to automate RNA-Seq data submissions. EFO will be extended to support annotation of these data, for example for single cell sequencing studies, and also for data integration in the sample database, where we will develop new terms for cell lines and samples used in genome-wide association studies (GWAS) studies. Finally, we are working to use EFO for RDF export of data from the GXA jointly with the Rebholz-Schuhmann group at the EBI with support from the EBI Industry Programme.

SELECTED REFERENCES

Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute, *Nucleic Acids Res.* 38, D690–D698.

Malone, J., et al. (2010) Modeling sample variables with an experimental factor ontology *Bioinformatics* 26, 1112–1118.

Parkinson, H., et al. (2009) ArrayExpress update: from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* 37, D868–D872.

Shankar, R., et al. (2010) Annotare: a tool for annotating high-throughput biomedical investigations and resulting data. *Bioinformatics* 26, 2470–2471.

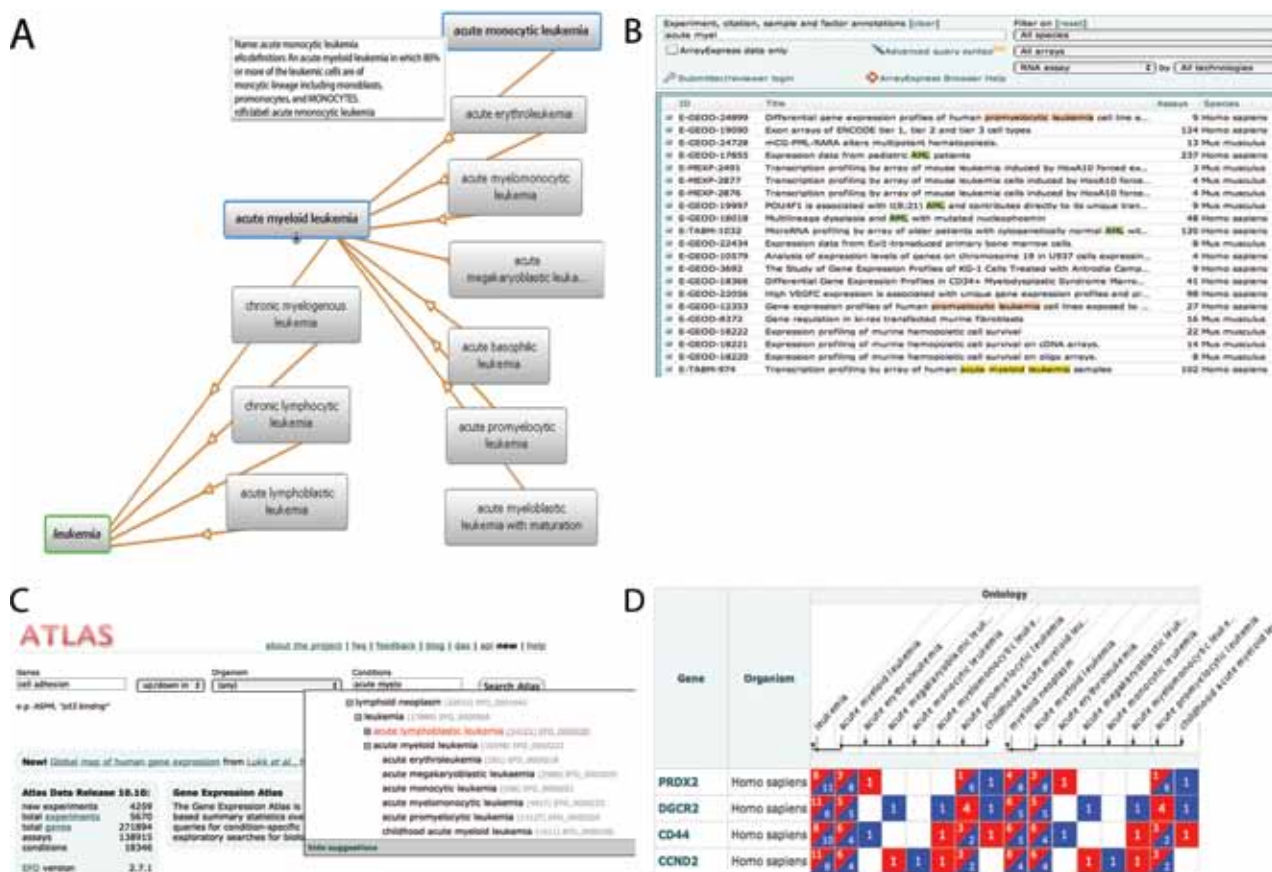


Figure 1. EFO is a data-driven application ontology that can be visualised as a node edge diagram showing terms placement and definitions in the BioPortal terminology browser (A), used to query ArrayExpress Archive Data (B) and used for query and visualisation for variables in the Gene Expression Atlas (C–D) in the heatmap view.



Peter Rice

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

Developing and Integrating Tools for Biologists

DESCRIPTION OF SERVICES

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 the EMBL-EBI through application development, training exercises and participation in international projects and standards development. Our group is 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 EMBOSS releases;
- Extended EMBOSS input formats to include SAM and BAM formats as well as versions of FASTQ format;
- Developed EDAM ontology of bioinformatics data types and methods.

MAJOR ACHIEVEMENTS

The EMBOSS development work is now fully funded by a grant from the BBSRC and has a team of three developers. During the reporting period we caught up on our backlog of maintenance and feature requests, and released two versions of EMBOSS (6.1.0 on 15 July 2009; interim 6.2.0 release on 15 January 2010). These releases complete the standardisation of EMBOSS internals for three books we are producing through Cambridge University Press. We are now free to concentrate on new developments, and have added many extensions to the current developers' version of the code. EMBOSS sequence objects now include extensive metadata derived from the richer input sources (especially the UniProt and EMBL databases), including database cross-references, ontology terms, literature references, taxons and specialised description fields from UniProt.

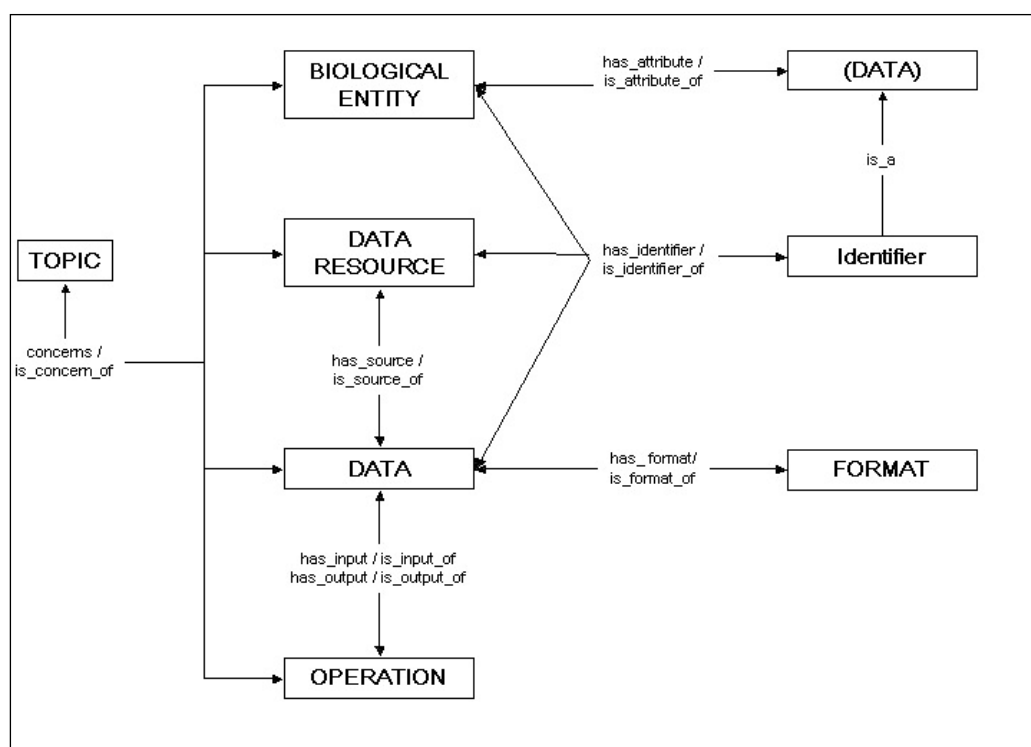


Figure. EDAM terms and relations

We extended EMBOSS input formats to include standard next-generation sequence formats, especially the SAM (sequence alignment/map) and BAM formats used by SAMtools and other packages, and the various versions of FASTQ format first used at the Sanger Institute. Together with other Open Bio Foundation projects (BioPython, BioPerl and BioRuby) we developed a standard interpretation of FASTQ format variants that can be shared across all the packages.

We split the library code for EMBOSS into five sub-libraries. This simplifies the maintenance of library code, and will allow the easier integration of code from other libraries in the future. We are particularly interested in the possibility of adding a C version of the Ensembl API.

Our maintenance tasks included a thorough review of the sequence data formats supported by EMBOSS, with special attention paid to recent changes in either the original format or its interpretation by other packages. We also removed several bottlenecks in the processing of sequence data, improving the performance of applications across the whole EMBOSS package. Source code documentation was also comprehensively reviewed, with naming standards imposed for all major library functions. This provided a clean set of sections and consistent programming interfaces for the forthcoming Developers Guide, and has significantly helped our own development efforts.

Where EMBOSS applications depend on other packages (especially 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 programme.

We recognised the need to provide semantic-level annotation of the many EMBRACE 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 is to develop a new ontology, EDAM (EMBRACE data types and methods), with 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 comprehensively over 200 EMBOSS applications 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.

FUTURE PLANS

The EMBOSS package is developing rapidly to cover new data types, new sources of data and new data access methods. In the next year we will add generic access to all public bioinformatics data resources with database types for feature annotation, bio-ontologies, taxonomic data, data resource descriptions (from our catalogue of public data resources) and general text or URL-based access to other sources of information. The EMBRACE consortium ends in July. 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.

The screenshot shows the EMBOSS web interface. At the top, there's a navigation bar with links like 'Home', 'All Databases', 'Enter Text Here', 'Reset', 'Download', and 'Site Index'. Below this is a sidebar with a tree view of categories: 'about', 'clients', 'help', 'services', 'misc', 'pfs', 'storage', 'remote', 'chustate', 'chustate2', 'distfile', 'distfile2', 'distfile3', 'distfile4', 'distfile5', 'distfile6', 'distfile7', 'distfile8', 'distfile9', 'distfile10', 'distfile11', 'distfile12', 'distfile13', 'distfile14', 'distfile15', 'distfile16', 'distfile17', 'distfile18', 'distfile19', 'distfile20', 'distfile21', 'distfile22', 'distfile23', 'distfile24', 'distfile25', 'distfile26', 'distfile27', 'distfile28', 'distfile29', 'distfile30', 'distfile31', 'distfile32', 'distfile33', 'distfile34', 'distfile35', 'distfile36', 'distfile37', 'distfile38', 'distfile39', 'distfile40', 'distfile41', 'distfile42', 'distfile43', 'distfile44', 'distfile45', 'distfile46', 'distfile47', 'distfile48', 'distfile49', 'distfile50', 'distfile51', 'distfile52', 'distfile53', 'distfile54', 'distfile55', 'distfile56', 'distfile57', 'distfile58', 'distfile59', 'distfile60', 'distfile61', 'distfile62', 'distfile63', 'distfile64', 'distfile65', 'distfile66', 'distfile67', 'distfile68', 'distfile69', 'distfile70', 'distfile71', 'distfile72', 'distfile73', 'distfile74', 'distfile75', 'distfile76', 'distfile77', 'distfile78', 'distfile79', 'distfile80', 'distfile81', 'distfile82', 'distfile83', 'distfile84', 'distfile85', 'distfile86', 'distfile87', 'distfile88', 'distfile89', 'distfile90', 'distfile91', 'distfile92', 'distfile93', 'distfile94', 'distfile95', 'distfile96', 'distfile97', 'distfile98', 'distfile99', 'distfile100'. The main content area is titled 'EMBOSS' and contains a 'Description' section, a 'Clients' section, a 'WSDL' section, and a 'Contact' section. The 'Description' section states: 'EMBOSS (European Molecular Biology Open Software Suite) is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copies with data in a variety of formats and allows transparent retrieval of sequence data from the web. Since extensive libraries are provided with the package providing a platform to allow scientists to develop and release software in the true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. For more information see: • EMBOSS homepage • BioCatalogue'. The 'Clients' section lists 'Language', 'Download', and 'Requirements' for Java, Perl, and Python. The 'WSDL' section provides a URL: 'http://www.ebi.ac.uk/Tools/web/services/EMBOSS/EMBOSS.wsdl'. The 'Contact' section provides an email address: 'services@ebi.ac.uk'.



Ugis Sarkans

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

Functional Genomics Software Development

DESCRIPTION OF SERVICES

Our team has been developing software for ArrayExpress since 2001. As of October 2010, ArrayExpress holds data from almost 500 000 microarray hybridisations and is one of the major data resources of EMBL-EBI. The software development team is building and maintaining several components of the ArrayExpress infrastructure, including data management tools for ArrayExpress Archive (the MIAME-compliant database for the data that support publications); the ArrayExpress Archive user interface; MIAMExpress (a data annotation and submission system); and an array design re-annotation system for aligning user-provided annotation to a uniform reference system. In addition, our team has participated in building the BioSamples database, a new EBI resource, since early 2010.

SUMMARY OF PROGRESS

- Built a new ArrayExpress data management infrastructure (to enter service in late 2010);
- Maintained the existing infrastructure, facilitating the 10-fold growth of ArrayExpress archive over the past four years;
- Significantly evolved the ArrayExpress Archive user interface to provide a more robust and richer service.

MAJOR ACHIEVEMENTS

During the reporting period we finished building all the major components required for the ArrayExpress infrastructure: data loading, unloading, management and export tools. Following completion of acceptance testing, this MAGE-TAB-centric infrastructure will be used in production (expected by the end of 2010).

The array design re-annotation system was significantly restructured, resulting in a better quality annotation that is propagated to the ArrayExpress Gene Expression Atlas. We continue to work together with the Atlas team to ensure that the re-annotation processes are optimised also for sequencing-based transcriptomics data, where microarrays are not used.

The ArrayExpress Archive user-interface work progressed in parallel with the back-end infrastructure developments and was released to end users independently. The search is now ontology-aware; for example, queries for 'cancer' retrieve studies annotated with 'lymphoma', and term synonyms can be used (Figure, highlight 1). Advanced query syntax is supported and can, for example, retrieve all experiments with more than 30 hybridisations. Links to the European Nucleotide Archive (ENA) are provided for sequencing-based functional genomics experiments where data is managed jointly by ArrayExpress and ENA (Figure, highlight 2). In addition, search field auto-completion has been implemented (Figure, highlight 3).

In preparation for the switch from the existing data management infrastructure based around the MAGE-ML data exchange language to the new one based on MAGE-TAB, we invested considerable effort into data conversion, cleaning and validation. Due to significant differences between abstraction levels of these two languages, more resources were required than initially planned; however, this work will ensure a higher level of data quality in the new database.

EMBL-EBI is developing a new BioSamples database, with an aim to clean up and aggregate aspects of biological sample information 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 and adapting relevant parts of the ArrayExpress software for these purposes.

We continued to refine our software development process, environment and tools. In particular, we concentrated on extensive testing, automated software environment management and improving teamwork.

FUTURE PLANS

After migration to the new software infrastructure we intend to concentrate on improving our data submission tools, taking into account the growing popularity of sequencing-based functional genomics experiments. The ArrayExpress Archive user interface will undergo further improvements, including simplifying and rationalising information presentation. In 2011 we plan to bring the BioSamples database to full production status, integrating this valuable new resource with other databases at EMBL-EBI and NCBI.

SELECTED REFERENCES

Malone, J., et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26, 1112-1118.

Kapushek, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Database issue), D690-D698.

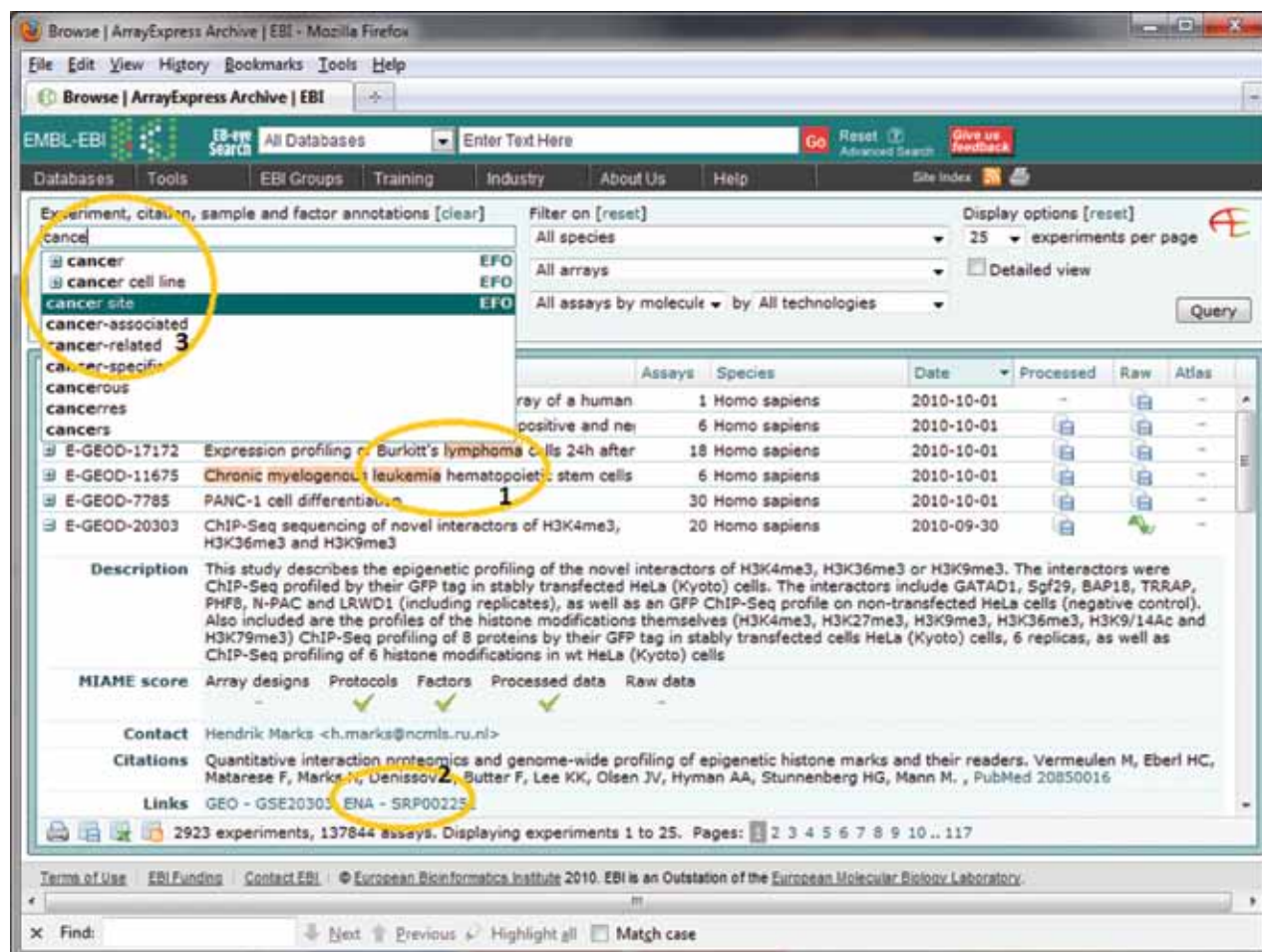


Figure. Some of the new features of the ArrayExpress Archive user interface. Highlight 1, new ontology-aware functionality; Highlight 2, data jointly managed by ArrayExpress and ENA; Highlight 3, search-field autocompletion.



Christoph Steinbeck

PhD Rheinische Friedrich-Wilhelm-Universität, Bonn, 1995. Postdoc at Tufts University, Boston, 1996-1997. Head of Research Group for Structural Chemoinformatics, Max Planck Institute of Chemical Ecology, Jena, 1997-2002. Habilitation in Organic Chemistry, Friedrich-Schiller-Universität, Jena, 2003. Head of Research Group for Molecular Informatics, Cologne University Bioinformatics Centre, Cologne, 2002-2007. Lecturer in Cheminformatics, University of Tübingen, 2007. At EMBL-EBI since 2008.

Cheminformatics and Metabolism

DESCRIPTION OF SERVICES

The Cheminformatics and Metabolism team provides the biomedical community with information on small molecules and their interplay with biological systems. Our database portfolio includes ChEBI, the 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, to determine quickly the structure of metabolites by stochastic screening of large candidate spaces and to enable the identification of molecules with desired properties. This requires algorithms for the prediction of spectroscopic and other physicochemical properties of chemical graphs based on machine learning and other statistical methods.

We are further investigating the extraction of chemical knowledge from the scientific literature by text- and graph-mining methods. This, as well as our work on chemical database technology and curation, is supported by research into chemical ontologies. Together with an international group of collaborators we have developed a number of widely known and used open-source cheminformatics software packages. The Chemistry Development Kit (CDK), which originated in our lab, is the leading open-source Java library for structural cheminformatics. Based on this, we have developed the cheminformatics workflow/pipelining system CDK-Taverna, which allows researchers to build executable data-processing workflows in a Lego™-like manner, as well as OrChem, our structure-registration and -searching system for the Oracle™-database. In collaboration with partners in Uppsala we initiated Bioclipse, an award-winning, rich client for chemo- and bioinformatics.

SUMMARY OF PROGRESS

- Released OrChem, our open-source chemical-search cartridge for Oracle™;
- Issued releases 59 to 69 of ChEBI, our ontology and database of chemical entities of biological interest;
- Issued releases 3 to 14 of Rhea and 50 to 61 of IntEnz, our enzyme resources;
- Made ChEBI fully structure-searchable based on OrChem;
- Substantially improved our chemical structure editor JChemPaint with R-groups and reactions;
- Secured a grant to establish the MetaboLights database at the EBI;
- Developed the first prototypes of the Enzyme Portal.

MAJOR ACHIEVEMENTS

The release of OrChem allows users to perform substructure, similarity and exact searching within OrChem. Recently it has included the ability to search for R-groups, enabling wild-card chemical-structure searching. It has also included descriptor calculation and exact chemical-structure searching.

In view of successful use of the ChEBI submission tool throughout the reporting period (576 submissions received from 16 individual external submitters), ChEBI has focused its efforts on improving the submission process. A bulk submission facility was developed to enhance the chemical submission experience of regular submitters by allowing programmatic provision of submitted data. As part of a collaboration with the La Jolla Institute of Allergy and Immunology in the USA, approximately 1500 entities associated with immunology were curated, involving use of a new facility for including citations in compound records. A ChEBI user workshop was organised and run, comprising both training elements and scientific talks and complemented by discussions on the future direction of ChEBI. Major improvements to Rhea/IntEnz have included the transformation of chemical compounds used in Rhea reactions from their neutral forms to their ionised states at pH 7.3 to be closer to common physiological conditions.

In 2009/2010, we were glad to host a considerable number of very talented interns, trainees and visiting scientists. Nikolas Fechner and Georg Hinselmann, University of Tübingen, worked with us in autumn 2009 on a novel method for classification of chemical compounds into ontological hierarchies, as well as on a visualisation technique for chemical spaces. Duan Lian of the East China University of Science and Technology in Shanghai, China optimised fingerprint performance in substructure search pre-screening by selecting fingerprint patterns with a novel analysis method. Leonid L. Chepelev of Carleton University in Ottawa, Canada worked on developing self-organising structure- and function-based chemical-entity hierarchies and tools for automated chemical classification in ChEBI. Laura Daniels, a student from Cranfield University in the UK, worked with our Enzyme Portal Team for her MSc thesis project, a case study about the challenges of data integration in the EBI enzyme portal. Kalai Vanii Jayaseelan of Vellore Institute of Technology (VIT) in India worked with the EBI text-mining infrastructure, developing a method for the association of small molecules with their biological targets. She has now joined our group as a bioinformatician and regular group member.

FUTURE PLANS

We have received a grant from the BBSRC to build the missing community resource for metabolomics at the EBI. The database, tentatively named MetaboLights, will be cross-species and cross-application and will cover metabolite structures and their reference spectra as well as their biological roles, locations and concentrations. The project will be fully compliant with open standards in metabolomics, including existing minimum reporting standards, or will actively contribute to their creation where these have not been developed. It will further provide the community with a repository for metabolomics experiments reference in scientific publications, matching the functionality of other 'omics' repositories at the EBI, such as the proteomics resource PRIDE. Our team is also leading an effort to integrate all enzyme-related information resources at the EBI into what we call the Enzyme Portal. The Enzyme Portal will provide unified access to resources like IntEnz and Rhea developed in our group as well as to some of the resources of the Thornton group, including their Cofactor Database, EC2PDB, the Catalytic Site Atlas and others.

SELECTED REFERENCES

- Griffin, J.L. and Steinbeck, C. (2010) So what have data standards ever done for us? The View from Metabolomics. *Genome Med.* 2, 38.
- Kuhn, T., et al. (2010) Cdk-Taverna: an open workflow environment for cheminformatics. *BMC Bioinform.* 11, 159.
- Rijnbeek, M., and Steinbeck, C. (2009) Orchem - an open source chemistry search engine for Oracle. *J. Cheminformatics* 1, 17.

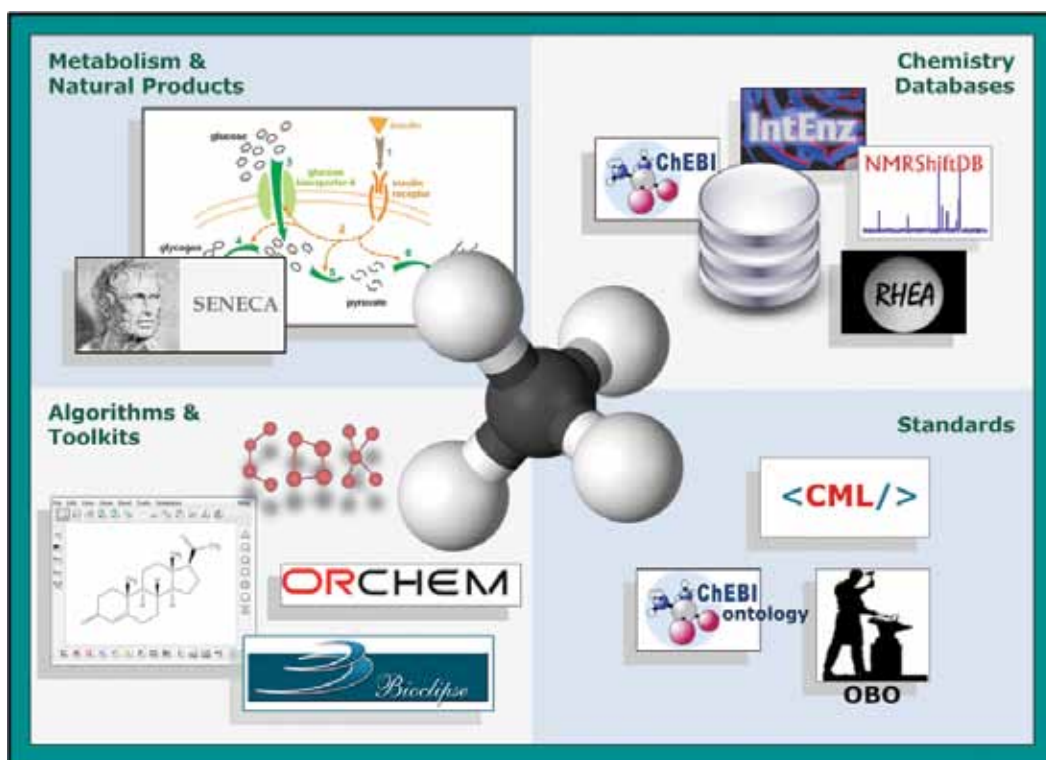


Figure. Scope of work in the cheminformatics and metabolism group.

Services Teams

PANDA

Joint Team Leader (Proteins)

Rolf Apweiler

Joint Team Leader (Nucleotides)

Ewan Birney

Team Leaders

Guy Cochrane
Paul Flicek
Henning Hermjakob
Sarah Hunter
Paul Kersey
Maria-Jesus Martin
Claire O'Donovan
John Overington
Christoph Steinbeck

Group Coordinators

Elsbeth Bruford
Paula de Matos
Glenn Proctor

Project Coordinators

Laura Clarke
Fiona Cunningham
Emily Dimmer
Xosé M. Fernandez
Javier Herrero-Sanchez
Anne Hersey
Pascal Kahlem
Ilkka Lappalainen
Daniel Lawson
Rasko Leinonen
David Lonsdale
Michele Magrane
Sandra Orchard
Manuela Pruess
Esther Schmidt
Damian Smedley
Dan Staines
Robert Vaughan
Juan Antonia Vizcaino*

Senior Scientific Database Curators

Paul Browne
Nadeem Faruque
John Stephen Garavelli
Kati Laiho
Jennifer McDowall*
Eleanor Stanley*
Matt Wright*

Scientific Database Curators

Ruth Akhtar*
Yasmin Alam-Faruque
Louisa Bellis
Patricia Bento
Ana Cerdeño-Tárraga*
Wei Mun Chan

Louise Daugherty
Margaret Duesbury*
Ruth Eberhardt
Marcus Ennis
Rebecca Foulger
Phani Garapati
Michael Gardner*
Richard Gibson
Susan Gordon
Christopher Hunter
Rachael Huntley
Jyoti Khadake
Bijay Jassal
Zara Josephs*
Steven Jupe
Duncan Legge
Gaurab Mukherjee*
Gareth Owen*
Klemens Pichler*
Ruth Seal
Harinder Sehra*
Petra ten Hoopen
David Thorneycroft*
Stephen Turner
Siew-Yit Yong

Bioinformaticians

Mark Bingley
Yuan Chen*
David Croft
Paul Derwent
Gemma Hoad
Julius Jacobsen
Michael Maguire
Craig McAnulla
Diego Poggioli

Senior Software Engineers

Ricardo Antunes
Daniel Barrell*
Richard Côté
Alexander Fedotov
Alan Horne*
Phil Jones
Samuel Kerrien
Jie Luo
John Maslen
Samuel Patient
Antony Quinn
Mark Rijnbeek
Manjula Thimma*

Software Engineers

Rafael Alcantara Martin
Bruno Aranda
Benoit Ballester
Kathryn Beal
Benoit Bely
David Binns
Lawrence Bower
Chao-Kung Chen

Ying Cheng
Matt Corbett
Ujjwal Das
Mark Davies
Bernard de Bono
Adriano Dekker
Fehmi Demiralp*
Stephen Fitzgerald
Anne Gaulton
Neil Goodgame
Leo Gordon
Matthias Haimel
Janna Hastings
Kenneth Haug
Jonathan Hinton*
Mikyung Jang
Andrew Jenkinson
Nathan Johnson
Szilveszter Juhos*
Andreas Kähäri
Damian Keefe
Stephen Keenan
Arnaud Kerhornou
Rhoda Kinsella
Gautier Koscielny
Stefan Kuhn*
Eugene Kulesha
Vasudev Kumanduri
Wudong Liu
Quan Lin*
Ian Longden
Michael Lush
Karyn Megy
Gavin O'Kelly
Rajesh Radhakrishnan
Florian Reisinger
Daniel Ríos
Tony Sawford
Guy Slater*
Richard Smith
Siamak Sobhany
Albert Vilella
Juan A. Vizcaino
Steven Wilder
Phil Wilkinson
Andy Yates
Vadim Zalunin
Holly Zheng-Bradley*

Helpdesk Officers

Jeff Almeida-King*
Bert Overduin
Michael Schuster
Giulietta Spudich

Postdocs

Patricia Bento*
Mikhail Spivakov

Group Secretaries

Shelley Goddard
Tracy Mumford

Administrative Assistant

Kerry Smith

Data Assistant

Sheila Plaister

PhD Students

Joe Foster
Andre Faure
Markus Fritz*
Alison Meynert*
Pablo Moreno
Michael Mueller*
Dace Ruklisa
Petra Schwalie
Daniel Zerbino*

Visitors

Cele Abad-Zapatero*
Jigisha Anupama*
Shelby Bidwell*
Fiona Brookes-Carter*
Leonid Chepelev*
Raymond Dalgleish*
Laura Daniels*
Kirill Degtyarenko*
Antonio Fabregat Mundo
Andre Faure*
Nikolas Fechner*
Mirko Ferraiolo
Ludvik Gomulski*
Johannes Griss
Gavin Ha*
Georg Hinselmann*
Kalai Vanii Jayaseelan*
Duan Lian*
Usha Mahadevan*
Yamile Marquez*
Thomas Maurel*
Nelson Ndegwa*
Laurence Newman*
Eric Pfeiffenberger*
Bertran Pitollat*
Emanuele Raineri*
Gustavo Salazar*
Andreas Schoenegger*
Marco Severgnini*
Stephani Sidibe*
Vipendra Singh*
Graham Taylor*
Sander Timmer*
Matthieu Visser
Jose Maria Villaveces Parda*
Christina Wass*

*Indicates part of the year only

EUROPEAN NUCLEOTIDE ARCHIVE

Team Leader
Guy Cochrane

Technical Coordinator
Rasko Leinonen

Annotation Coordinator
Bob Vaughan

Production Coordinator
Nadeem Faruque

Scientific Database Curators
Ruth Akhtar*
Ana Cerdeño-Tárraga*
Richard Gibson
Christopher Hunter*
Petra Ten Hoopen

Software Engineers
Lawrence Bower
Ying Cheng
Iain Cleland*
Fehmi Demiralp*
Neil Goodgame
Mikyung Jang
Rajesh Radhakrishnan
Siamak Sobhany
Vadim Zalunin

Bioinformaticians
Gemma Hoad
Quan Lin*

Data Assistant
Sheila Plaister

VERTEBRATE GENOMICS

Team Leader
Paul Flicek

Project Leaders
Laura Clarke (Resequencing Informatics)
Fiona Cunningham (Ensembl Variation)
Ian Dunham (Ensembl Functional Genomics)
Javier Herrero (Ensembl Compara)
Ilkka Lappalainen (Variation Archive)
Damian Smedley (Mouse Informatics)

Scientific Programmers
Jonathan Hinton
Nathan Johnson
Damian Keefe
Stephen Keenan*
Vasudev Kumanduri
Michael Maguire*
Richard Smith
Steven Wilder
Holly Zheng Bradley*

Ensembl Developers
Kathryn Beal
Yuan Chen*
Stephen Fitzgerald
Leo Gordon
Pontus Larsson*
Will McLaren
Graham Ritchie*
Daniel Sobral*
Albert Vilella

Bioinformaticians
Chao-Kung Chen
Edoardo Marcora
Phil Wilkinson

Postdoctoral Fellows
Benoit Ballester
David Thybert*

User Support Officer
Jeff Almeida-King

PhD Students
Andre Faure
Petra Schwalie

Visitors
Thomas Maurel*

Team Secretary
Kerry Smith

PROTEOMICS SERVICES

Team Leader
Henning Hermjakob

Bioinformaticians
David Croft
Attila Csordas
Marine Dumousseau
Pierre Grenon*
Gavin O'Kelly
David Ovelheiro
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

Software Engineers (including senior software engineers)
Bruno Aranda
Richard Cote
Andrew Jenkinson
Rafael Jimenez
Samuel Kerrien
Daniel Rios
Florian Reisinger
Chris Taylor
Sarala Wimalaratne

Visitors (including visiting students)
Antonio Fabregat Mundo*
Laurent Gatto*
Nelson Ndegwa Gichora
Johannes Griss*
Gustavo Salazar*
Jose Villaveces*
Matthieu Visser

INTERPRO

Team Leader
Sarah Hunter

Software Development Coordinator
Phil Jones*

Annotation Coordinator
David Lonsdale

Senior Scientific Curator
Jennifer McDowall*

Scientific Curators
Sarah Burge*
Louise Daugherty
Prudence Mutowo*
Amaia Sangrador*

Senior Software Engineers
David Binns
John Maslen
Antony Quinn
Manjula Thimma*

Bioinformaticians
Craig McAnulla
Chris Hunter*
Siew-Yit Yong

Database Production Manager
Ujjwal Das

Web Developer
Sebastien Pesseat*

ENSEMBL GENOMES

Team Leader
Paul Kersey

Coordinators
Eugene Kulesha
Daniel Lawson
Daniel Staines

Bioinformaticians
Paul Derwent
Daniel Hughes
Uma Maheswari
Karyn Megy
Michael Nuhn
Alessandro Vullo*
Derek Wilson*

Software Engineers
Matthias Haimel
Arnaud Kerhornou
Gautier Koscielny
Nick Langridge*

Visitors
Darren Waite
Valerie Wood*

Outreach
Jeff Almeida-King (shared with Paul Flicek)

UNIPROT DEVELOPMENT

Team Leader
Maria J. Martin

Project Leaders
Alexander Fedotov
Samuel Patient

Senior Software Engineers
Jie Luo

Software Engineers
Ricardo Antunes
Daniel Barrell*
Benoit Bely
Francesco Fazzini
Leyla Jael Garcia Castro*
Wudong Liu
Nikolas Pontikos
Steven Rosanoff*
Tony Sawford
Edward Turner*
Xavier Watkins
Tony Wardell*

Web Developer
Mark Bingley

Bioinformatician
Diego Pogglioli

UNIPROT CONTENT

Team Leader
Claire O'Donovan

Project Leaders
Emily Dimmer
Michele Magrane

Senior Scientific Database Curators
Paul Browne*
Wei Mun Chan
Ruth Eberhardt
Rebecca Foulger*
John Garavelli
Kati Laiho*

Scientific Database Curators
Yasmin Alam-Faruque
Michael Gardner
Rachael Huntley
Duncan Legge
Klemens Pichler
Harinder Sehra

Bioinformatician
Julius Jacobsen

ChEMBL

Team Leader
John P. Overington

Software programmers
Francis Atkinson
Patricia Bento
Jon Chambers
Mark Davies
Anna Gaulton
Kazuyoshi Ikeda
Stefan Kuhn*
Shaun McGlinchey

Curators
Ruth Akhtar
Louisa Bellis
Yvonne Light

PhD students
Felix Krueger
Ben Stauch*

Visitors
Jigisha Anupuma*
Vipendra Singh*

Coordinator
Anne Hersey

CHEMINFORMATICS AND METABOLISM

Team Leader
Christoph Steinbeck

Bioinformatician
Kalai Vanii Jayaseelan*

Coordinator
Paula de Matos

Curators (including Senior Curators)
Marcus Ennis
Zara Josephs
Gareth Owen
Steven Turner

*Indicates part of the year only

Software Engineers (including Senior Software Engineers)

Hong Cao*
Adriano Dekker
Janna Hastings
Kenneth Haugh*
Duncan Hull*
Rafael Alcantara Martin
Mark Rijnbeek

PhD Student

Pablo Moreno

Visitors (including visiting students)

Nico Adams*
Leonid Chepelev*
Nikolas Fechner*
Georg Hinselmann*
Kalai Vanii Jayaseelan*
Duan Lian*

FUNCTIONAL GENOMICS

Group Leader

Alvis Brazma

Team Leaders

Misha Kapushesky
Helen Parkinson
Ugis Sarkans

Coordinators

Maria Krestyaninova
Philippe Rocca-Serra
Susanna-Assunta Sansone

Software Developers

Mike Gostev
Eamonn Maguire

Scientists

Johan Rung
Gabriela Rustici

Bioinformaticians

Julio Fernandez Banet
Chris Taylor

PhD Students

Nils Gehlenborg
Angela Gonzalves

Personal Assistant

Lynn French

FUNCTIONAL GENOMICS ATLAS

Team Leader

Misha Kapushesky

Software Engineers

Alexey Filippov*
Olga Melnichuk
Robert Petryszak*
Nataliya Sklyar*
Andrew Tikhonov
Andrey Zorin

Research Fellow

Wanseon Lee*

Bioinformatician

Nikolay Pultsin*

FUNCTIONAL GENOMICS PRODUCTION

Team Leader

Helen Parkinson

Bioinformaticians

Tomasz Adamusiak
Anna Farne
Emma Hastings*
Ele Holloway
Natalja Kurbatova
Margus Lukk*
James Malone
Ravensara Travillian*
Eleanor Williams

Software Engineer

Tony Burdett

Visitor

Morris Swertz

FUNCTIONAL GENOMICS SOFTWARE DEVELOPMENT

Team Leader

Ugis Sarkans

Coordinator

Nikolay Kolesnikov

Software Engineers

Niran Abeygunawardena
Marco Brandizi*
Miroslaw Dylag
Ibrahim Emam
Ekaterina Pilicheva
Anjan Sharma
Nataliya Sklyar*

Database Administrator

Roby Mani

PROTEIN DATABANK IN EUROPE (PDBe)

Senior Team Leader

Gerard Kleywegt

Technical Team Leader

Tom Oldfield

Administrator

Pauline Haslam*

Curators

Matthew Conroy*
Gaurav Sahni
Sanchayita Sen
Jawahar Swaminathan
Martyn Symmons*

Software Engineers

Younes Alhroub*
Christoph Best
Harry Boutselakis*
Jose Dana*
Adel Golovin
Swanand Gore*
Aleksandras Gutmanas*
Miriam Hirshberg
Glen van Ginkel
Ingvar Lagerstedt*
Jorge Pineda Castillo
Luana Rinaldi*
Robert Slowley
Antonio Suarez-Uruena
Sameer Velankar
Wim Vranken

Database Administrators

John Melford

Postdoc

Ségolène Caboche

Visitors

Anaëlle Ailli
Laurence Newman
Gregoire Sawka

THE GENE ONTOLOGY EDITORIAL OFFICE

Team Leader

Jane Lomax

Scientific Database Curators

Jennifer Deegan
Rebecca Foulger
Midori Harris
Amelia Ireland

LITERATURE RESOURCES

Team Leader

Johanna McEntyre

Software Developers

Paula Buttery*
Norman Cobley*
Alan Horne
Jyothi Katuri*
Sharmila Pillai

Visitor

Andrew Caines*

Consultant

Peter Stoehr

TOOLS FOR BIOLOGISTS

Team Leader

Peter Rice

Senior Software Engineer

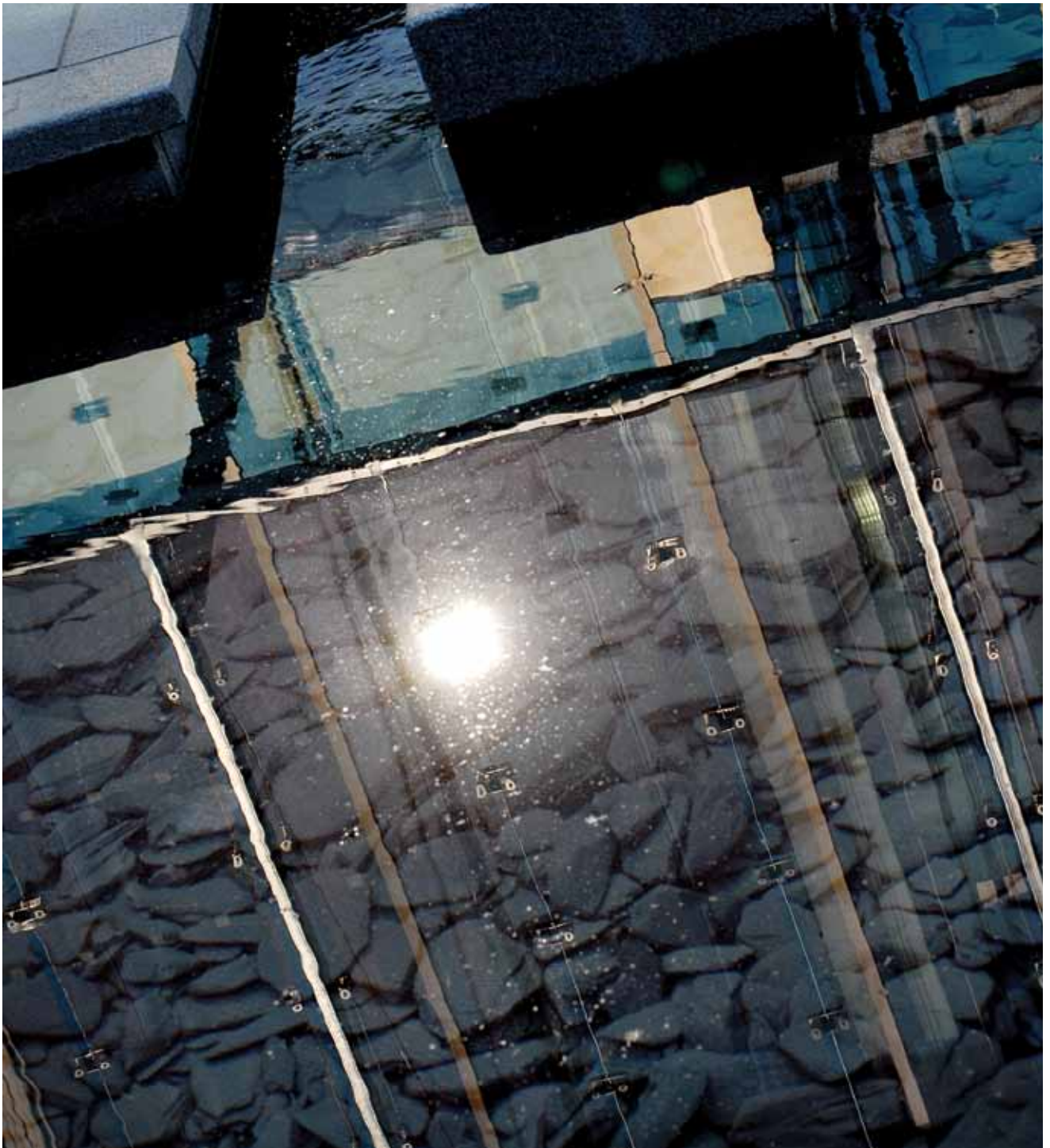
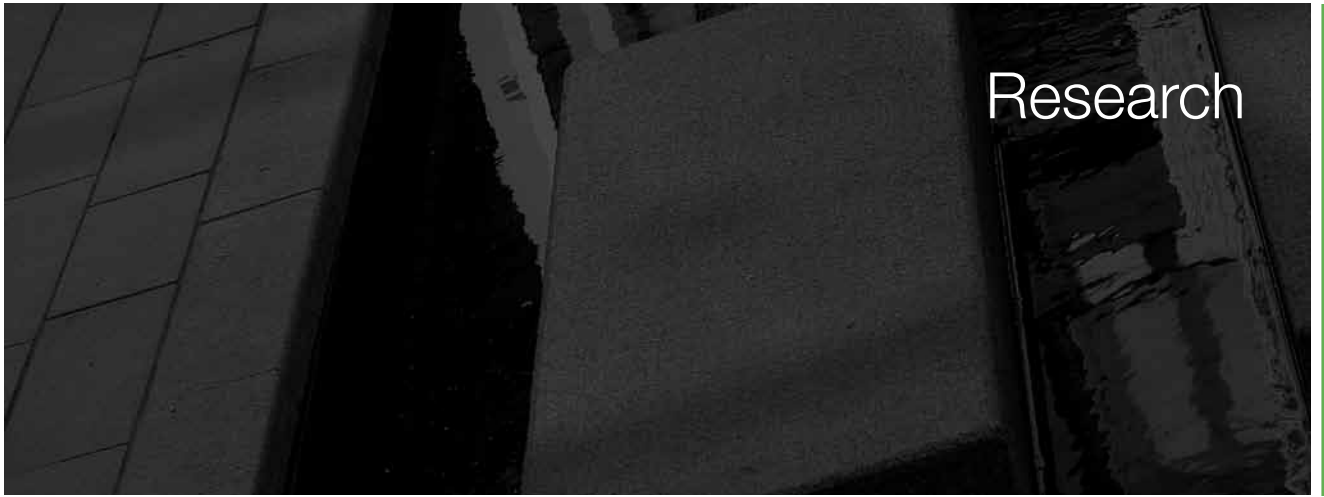
Alan Bleasby

Software Engineers

Jon Ison
Mahmut Uludag

*Indicates part of the year only







Paul Bertone

*PhD Yale University, 2005.
At EMBL–EBI since 2005.
Joint appointments in
Genome Biology and
Developmental Biology
Units.*

Pluripotency, Reprogramming and Differentiation

DESCRIPTION OF RESEARCH

We investigate the cellular and molecular processes underlying mammalian stem cell differentiation 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.

A number of molecular pathways involved in embryonic development have been elucidated, including those that influence stem cell differentiation. As a result, we know of a number of key transcriptional regulators and signalling molecules that play essential roles in manifesting nuclear potency and self-renewal capacity of embryonic and tissue-specific stem cells. Despite these efforts, only a small number of components have been identified and large-scale characterisation of cellular commitment and terminal differentiation to specific cell types remains incomplete. 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 in 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;
- Released open-source software for expression profiling and high-throughput sequencing.

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 involved in 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 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 stem 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.

Together with Steven Pollard at University College London, we have 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 which is positively correlated with patient survival rates in cases of glioblastoma multiforme.

To support these experiments we have optimised new experimental protocols for RNA sequencing that address some of the limitations present in standard approaches. We have also developed software tools for real-time expression profiling (Dvinge and Bertone 2009), and automated, genome-wide scanning of high-throughput sequencing data for binding site occupancy and transcribed sequences (Salmon-Divon et al. 2010).

FUTURE PROJECTS AND GOALS

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 REFERENCES

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.

Salmon-Divon, M., et al. (2010) PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* 11, 415.

Dvinge, H. and Bertone, P. (2009) HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics* 25, 3325-3326.

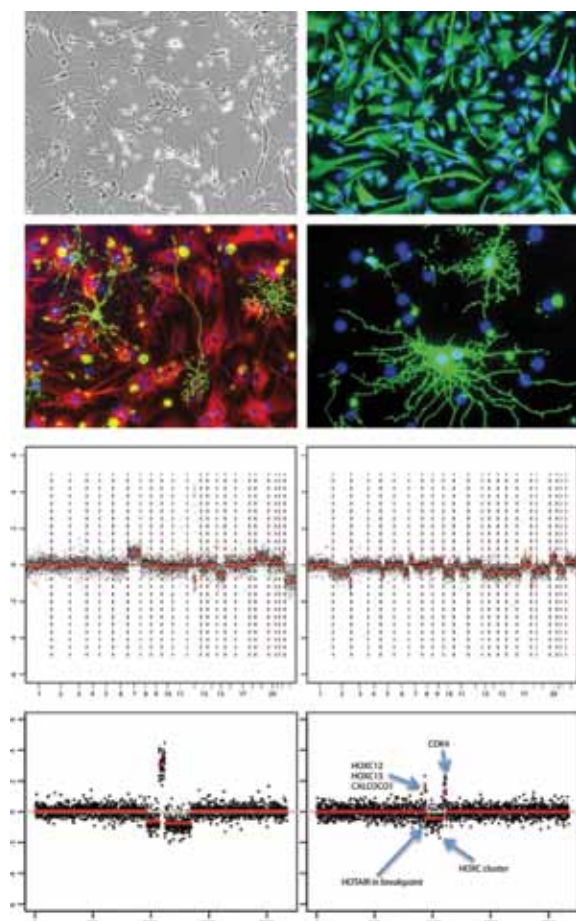


Figure. Neural cancer stem cells propagate indefinitely in culture (top) and can differentiate into the major cell types of the central nervous system, such as astrocytes and oligodendrocytes (second row). Array CGH and genome resequencing identify chromosomal abnormalities (third row) and the disruption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



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.*

Functional Genomics and Analysis of Small RNA Function

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 not annotated or possess a poorly characterised function. Our group aims to predict and describe the functions of genes, proteins and regulatory RNAs as well as their interactions in living organisms.

Regulatory RNAs have recently entered the limelight as the roles of a number of novel classes of non-coding RNAs have been uncovered. Our work is computational and involves the development of algorithms, protocols and datasets for functional genomics. Our research currently focuses on determining the functions of regulatory RNAs. We collaborate extensively with experimental laboratories on both the commissioning of experiments and analysis of experimental data. Some laboratory members take advantage of these close collaborations to gain hands-on experience in the wet lab and perform relevant experiments to support their computational projects.

SUMMARY OF PROGRESS

- Developed Sylamer, a system for finding significantly over- or under-represented sequences according to a sorted gene list.

MAJOR ACHIEVEMENTS

Stijn van Dongen and Cei Abreu-Goodger led our team's work on microRNA (miRNA) target prediction and analysis. Purely computational methods for miRNA target prediction perform well but suffer from over-prediction. For the most part, miRNA-target gene binding is detectable as a significant shift in mRNA expression of the target gene. Given an experiment where a miRNA is perturbed, we can assess the effect of that miRNA on global mRNA expression levels, query whether this is a significant and direct effect and also identify likely target genes whose expression levels have responded appropriately. We developed Sylamer, a system for finding significantly over- or under-represented sequences according to a sorted gene list.

Analysis of over-represented features in lists of genes is a powerful tool for associating function with biological effects. Instead of using a single cut-off and thus a single gene list, gene-set enrichment analysis uses all the genes, ranked according to how they change during the experiment. Sylamer rapidly assesses over- and under-representation of nucleotide 'words' of specific length in ranked gene lists. Using multiple cut-offs, it determines whether each word is more abundant at one end of the list than expected when compared to the rest, calculating significance using hyper-geometric statistics. The method takes into account compositional biases in 3'UTRs and multiple testing. The Sylamer algorithm was designed to be fast and efficient and is able to process a genome-wide dataset for hundreds of miRNA signatures in seconds. The algorithm is freely available as a stand-alone package and we are also constructing a fully featured online analysis resource to make this approach easily accessible to the community. We have constructed a web-based resource, SylArray, for automated Sylamer analysis on gene lists, which automatically assigns 3'UTR sequences to the gene list (Bartonicek et al., 2010). We have used Sylamer on a number of collaborative projects to discover miRNA effects and their target genes (Lewis et al., 2009; Rasmussen et al., 2010).

Harpreet Saini and Nenad Bartonicek led the group's work on small RNA genomics. Many miRNAs are intergenic and encoded on their own non-coding transcript, in contrast to those which lie within the introns of protein-coding genes. For most intergenic miRNAs little or nothing is known about the host transcript. It is desirable to identify the boundaries of this transcript and its regulatory features, in particular for knockout studies and prediction of miRNA transcriptional activators. We developed the miRBase::Genomics resource for the prediction and analysis of the primary transcript of miRNAs (pri-miRNA). We also worked on the prediction and analysis of other regulatory RNAs, including piwi-associated RNAs (piRNAs) and small, non-coding RNAs (sncRNAs) in bacteria. Part of this work involves prediction of the transcriptional units of common RNAs and their upstream regulatory factors. We also collaborated with Donál O'Carroll at EMBL-Monterotondo on the analysis of piRNAs in the germline of mouse. Using a combination of next generation sequencing and genetic knock-out or knock-in approaches we are trying to understand the processing and regulation of these important molecules (Figure).

Work on 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. Still, a great deal can be learned about the evolution of small RNA regulation in vertebrates. Our first approach to the problem was to identify likely orthologues and paralogues of known miRNAs across large numbers of species. The miRBase database of miRNA sequences focuses on the species from which a miRNA was first isolated and does not always attempt to find its counterparts in diverse organisms. We developed a novel mapping strategy for identifying likely miRNA loci in multiple organisms given 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. This will be used to detect evolutionary events such as deletions, duplications or transfer of miRNAs within and across species. We sought to detect those miRNAs with interesting evolutionary histories and to examine their function by virtue of their mRNA targets. 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.

FUTURE PLANS

Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. By building these integrated networks we hope to place miRNAs into a functional context that will help us better understand the function and importance of these regulatory molecules. We will also cross-compare miRNA alignments with likely target alignments to identify possible cases of correlated evolution where changes in miRNA sequence are compensated by changes to the regulatory target. We will continue our work on piRNAs, to build an accurate database of piRNA loci in animals and to explore the importance and evolution of these molecules. We are extremely interested in the evolution of regulatory RNAs and in developing phylogenetic techniques appropriate for short non-coding RNA. We will continue to build strong links with experimental laboratories working on miRNAs in different systems. Such work allows us to build better datasets with which to train and validate our computational approaches. The use of visualisation techniques to assist with the interpretation and display of complex, multi-dimensional data will continue to be an important parallel aspect of our work.

SELECTED REFERENCES

- Bartonicek, N. and Enright, A.J. (2010) SylArray: A web-server for automated detection of miRNA effects from expression data. *Bioinformatics* (in press). DOI:10.1093/bioinformatics/btq545.
- Guerra-Assuncao, J.A. and Enright, A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* 11, 133.
- Lewis MA., et al. (2009) An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat. Genet.* 41, 614-618.
- Rasmussen KD., et al. (2010) The miR-144/451 locus is required for erythroid homeostasis. *J. Exp. Med.* 207, 1351-1358.

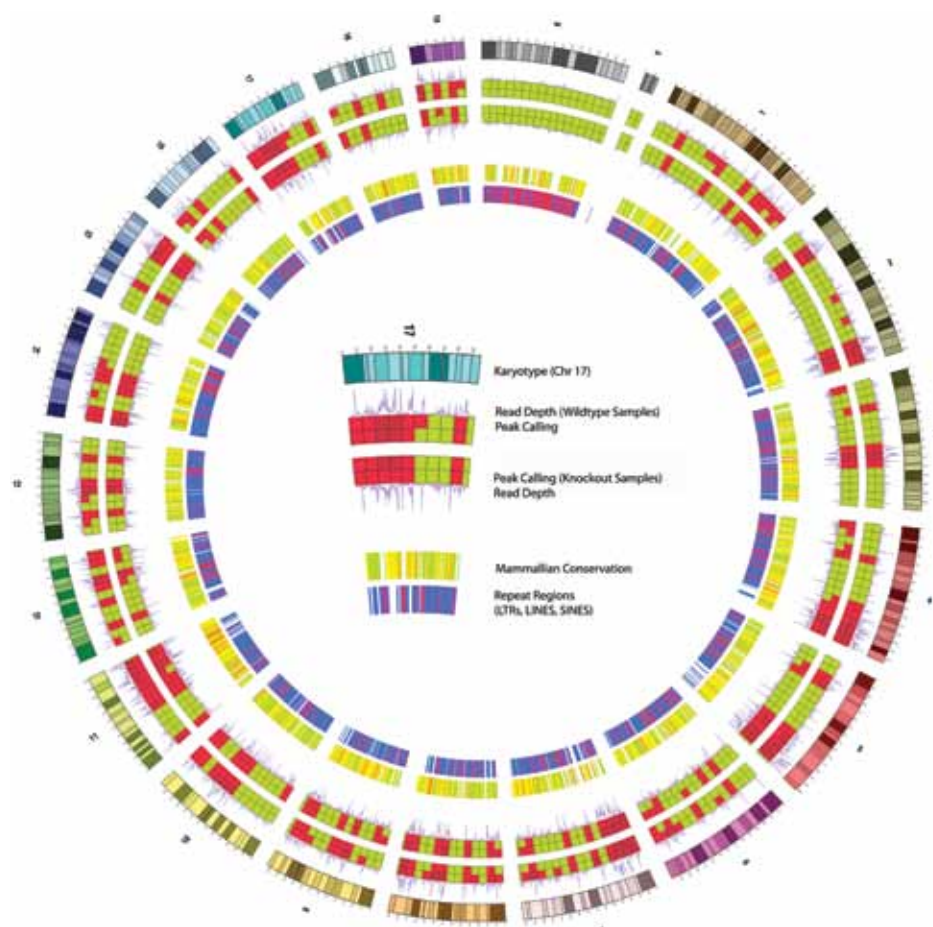


Figure. piRNA sequencing data from two wild-type and two mutant mice. The mouse karyotype is shown around the outside of the plot. Sequencing reads from both sets of samples are shown as line plots. Hotspots for piRNA biogenesis are illustrated as red squares on the middle tracks. The two inside tracks show mammalian conservation and repetitive genomic elements (innermost).



Nick Goldman

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

Evolutionary Tools for Genomic Analysis

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, to describe and understand processes of evolution and to make predictions about the function of genomic sequence. The group maintains a balance between phylogenetic methodology development; the provision of these methods to other scientists via stand-alone software and web services; and applications of such techniques, focusing on comparative genomics and the bulk analysis of biological sequence data. Collaborations with sequencing consortia provide the essential state-of-the-art data and challenges to inspire and confront these new methods of sequence analysis. Intra-group collaborations between members 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. The group has traditionally been strong in examining the theoretical foundations of phylogenetic reconstruction and analysis. In 2009–2010, the group has continued to gain expertise with next-generation sequencing (NGS) data. This vast source of new data promises great gains in understanding genomes but brings with it many new challenges. Our aim is to increase our understanding of the process of evolution and to provide new tools to elucidate the function of biological molecules as they change over evolutionary timescales.

SUMMARY OF PROGRESS

- Further developed algorithms and improved accessibility of our phylogeny-aware alignment algorithms;
- Designed and implemented a general Markov chain Monte Carlo framework for phylogenetic inference with single-site variation models, allowing virtually all parameters of a classical phylogenetic model to vary along the evolutionary history (work with Samuel Blanquart, now at INRIA, Lille);
- Undertook a simulation study to understand and correct for alignment error in evolutionary sequence analysis, showing that alignment error can cause unacceptable error rates in the detection of positive selection and that although alignment filtering can significantly reduce this error rate, there is still much room for improvement;
- Made significant algorithmic improvements to our AYB base-calling method, increasing its performance while reducing memory requirements; also enhanced software quality by employing a professional programmer, Hazel Marsden, to implement AYB in a robust and user-friendly package;
- Produced simNGS, an NGS simulator based on the AYB model to produce ‘data’ from known sequence incorporating errors as though from an actual sequencing machine, and PhyloSim, an extensible, object-oriented framework written in R offering unmatched flexibility in the simulation of sequence evolution;
- Carried out a study of the constraints acting on the promoters of *Drosophila melanogaster*;
- Completed a systematic study of transcription factors (TFs) in hemiascomycetous yeasts;
- Screened the gorilla genome for genes evolving under accelerated evolution in the gorilla lineage, identifying dozens of targets including genes involved in brain development, spermatogenesis and regulation of iron homeostasis (collaboration with Nick Mundy’s group, Zoology Department, University of Cambridge).

MAJOR ACHIEVEMENTS

Our phylogeny-aware multiple sequence alignment algorithm PRANK has been shown to produce superior inferences of character homology. We developed an easy-to-use web interface, webPRANK, to generate PRANK alignments along with a web-based browser to visualise and post-process them. Our intention was for a greater range of laboratory biologists and bioinformaticians to gain access to superior data analyses without having to master all of their algorithmic intricacies or complex computer code. Building on the experience and successes of PRANK, and in response to the challenges of NGS data, we have also developed a graph-based method to represent uncertainty in reconstructed ancestral sequences. The resulting graph-alignment method, named PAGAN, is applied to phylogenetic mapping of RNA-seq data (see Figure) in the Pachinko analysis pipeline (collaboration with Albert Vilella, EBI/Compara), allowing accurate transcript assembly for non-model organisms.

With Martin Taylor (now at MRC Human Genetics Unit, Edinburgh) we assessed the general constraints acting on the promoters of *D. melanogaster*, investigating the profiles of insertions and deletions, substitution rates and base composition. The results provide new insights into transcription factor binding site (TFBS) turnover, leading to more thorough understanding of their evolutionary dynamics, which are believed to be important in the evolution of form and function. In doing this, we designed new methods for estimating neutral rates, confirmed that chromosomal context influences rates of molecular evolution in regulatory regions and developed a simple yet efficient model for selective pressure assessment in TFBSs.

Addressing the impact of gene duplication and divergence of protein-coding genes in the evolution of transcriptional regulatory networks, we completed a systematic study of TFs in hemiascomycetous yeasts. Network growth in species that underwent a whole-genome duplication (WGD) event was predominantly due to this event and was not family specific. We found TF repertoires in other species, in contrast, to have grown through a different mechanism, i.e. lineage-specific amplification of the largest families. Furthermore, we found WGD duplicates to be enriched for regulatory hubs compared to non-WGD duplicates. This underlines the adaptive potential of whole-genome duplications as well as having important implications for the study of evolution of regulatory networks in other organisms.

FUTURE PLANS

The study of genome evolution continues to inspire us with novel problems in phylogenetic methodology. The complex nature of the non-independence of sequence data due to their evolutionary relatedness continues to generate statistically challenging problems and we will continue to contribute to this theoretical field. We remain dedicated to practical applications of these methods in order to promote best practice in computational evolutionary and genomic biology, to keep in touch with the evolving needs of laboratory scientists and to continue to benefit from a supply of motivational biological questions where computational methods can help. In 2010–2011 we will continue to expand our efforts in NGS topics; it is increasingly clear that before we can gain the full benefit of inexpensive and extensive genome sequencing we will have to devise suitable sequencing strategies, understand and allow for the significant error rates of even the most advanced sequencers and take account of the evolutionary origins of the genomes we study.

SELECTED REFERENCES

Chor, B., et al. (2009) Genomic DNA K-Mer spectra: models and modalities. *Genome Biol.* 10, R108-R108.

San Mauro, D., et al. (2009) Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear rag1. *Syst. Biol.* 58, 425-438.

Talavera, D., Taylor, M.S. and Thornton, J.M. (2010) The (non)malignancy of cancerous amino acidic substitutions. *Proteins* 78, 518-529.

Yang, Z., Nielsen, R. and Goldman, N. (2009) In defense of statistical methods for detecting positive selection. *Proc. Nat. Acad. Sci. USA* 106, E95.

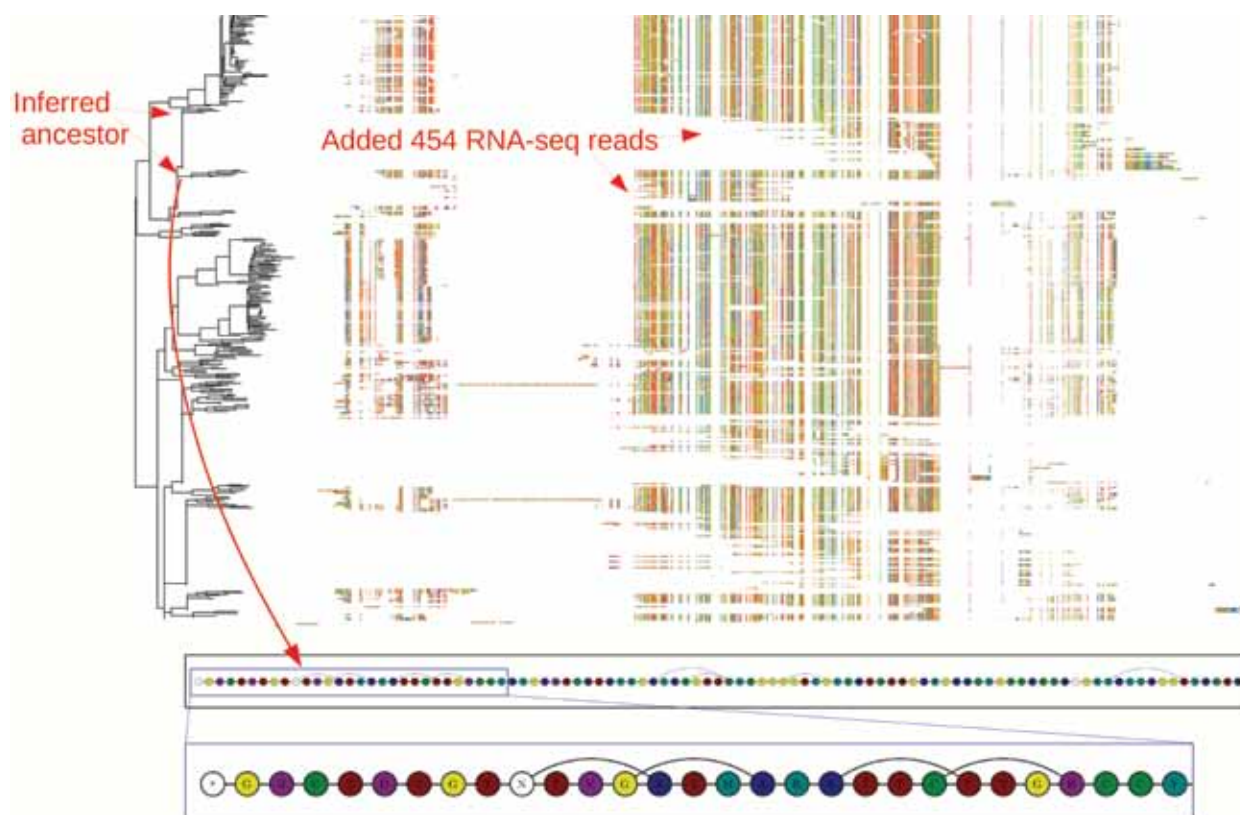


Figure. The PAGAN graph aligner can reconstruct sequence history for an existing alignment, and align NGS RNA-seq reads to reconstructed ancestors (top). The input reads can be in FASTQ format and paired-ended; homopolymer errors can be modelled and low-quality bases trimmed. Reads can be assigned to a node or a set of nodes, or the best node can be searched for either among all nodes or within a predefined subset. PAGAN is included in the Pachinko analysis pipeline (Albert Vilella, EBI/Compara) that allows the assembly of RNA transcripts from relatively low-coverage data and without a closely related genome sequence. Alternative possible ancestral sequences are recorded as graphs (bottom) that permit the transfer of both inferred alignments and uncertainty of true insertion-deletion history throughout the phylogeny.



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.*

Computational Systems Neurobiology

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, MIASE) and controlled vocabularies (e.g. SBO, the Systems Biology Ontology) to improve model semantics. In order to manage perennial cross-references, we are developing MIRIAM Resources and its associated URN scheme. Finally, the 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

- Progress on the modelling of signalling pathways involved in synaptic plasticity led to a deeper understanding of the complex equilibria and kinetic events involved in calcium signalling;
- The number of models provided by BioModels Database increased by 47%, with more than 630 models now publicly distributed;
- The use of MIRIAM URIs and the associated annotation scheme spread outside the modelling community, supported by various efforts such as the PSI consortium and Pathway Commons.

MAJOR ACHIEVEMENTS

We investigated the consequences of the fact that, for most signalling systems in vivo, the concentrations of the receptor molecules are near the values of the dissociation constants (Edelstein, S.J. et al., 2010). Specifically, we analysed cooperative signalling systems and demonstrated that increasingly strong ligand depletion progressively increases the dynamic range (i.e., the range of concentrations 'measured' by the system) and decreases the cooperativity. Therefore, an in vitro dose-response cannot be used to infer the activity of a sensing protein in vivo.

SBGN was conceived as the equivalent of circuit diagrams for biologists (Le Novère et al., 2009). Developed by a community of biochemists, modellers and computer scientists, SBGN consists of three complementary languages: process descriptions, entity relationships and activity flows. Together they enable scientists to represent networks of biochemical interactions in a standard, unambiguous way. We believe that SBGN will foster efficient and accurate representation, visualisation, storage, exchange and reuse of information on all kinds of biological knowledge, from gene regulation to metabolism and cellular signalling.

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. MIRIAM Resources will be updated in order to support more data types and to provide improved resolution services.

SELECTED REFERENCES

- Li, C., et al. (2010) BioModels Database, enhanced curated and annotated resource of published quantitative kinetic models. *BMC Systems Biol.* 4, 92.
- Tolle, D. and Le Novère, N. (2010) Brownian diffusion of AMPA receptors is Sufficient to explain fast onset of LTP. *BMC Systems Biol.* 4, 25.
- Edelstein, S.J., Stefan, M.I. and Le Novère, N. (2010). Ligand depletion in vivo modulates the dynamic range and cooperativity of signal transduction. *PLoS ONE* 5, e8449.
- Le Novère, N., et al. (2009). The Systems Biology Graphical Notation. *Nat. Biotechnol.* 27, 735-741.

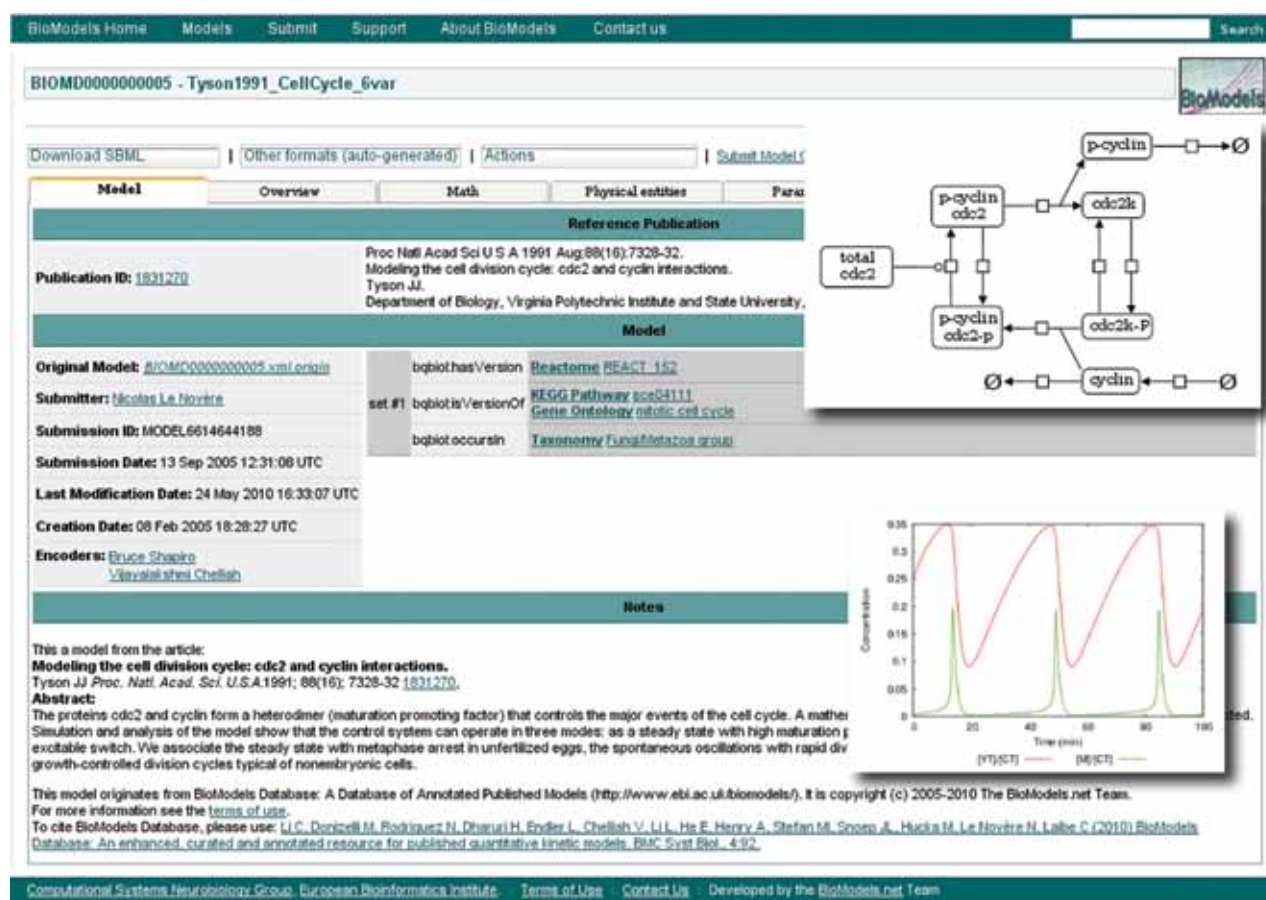


Figure. The BioModels Database: main page of a model of the cell cycle, together with the reaction graph and the reproduction of published results by our curators.



Nicholas Luscombe

*PhD University College London, 2000.
Postdoc at the Department of Molecular
Biophysics & Biochemistry, Yale
University, USA. At EMBL-EBI since
2005. Joint appointment with Gene
Expression Unit, EMBL-Heidelberg.*

Genomics and Regulatory Systems

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 a 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 has derived from molecular biological and genetic investigations. In the past decade, the availability of genome sequences and development of new laboratory techniques have – and continue to – generate information describing the function and organisation of regulatory systems on an unprecedented scale. 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 rising flood of biological data demands the application of computational methods to answer many interesting questions. The combined strength of bioinformatics and genomics gives us the ability to uncover general principles, providing global descriptions of entire systems. Research in the Luscombe Group 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

- Conducted genome-wide analysis of the repertoire, usage and cross-species conservation of transcription factors in the human genome;
- Identified and characterised nucleoporins as major, genome-wide regulators of transcription in higher eukaryotes;
- Examined how bacterial cellular systems are controlled through the combination of transcription factors, histidine kinases and small molecules;
- Developed and applied 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. 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). In 2010 we also discovered that components of the nuclear-pore complex control the expression for approximately 25% of genes by shaping the three-dimensional organisation of the genome (Vaquerizas et al., 2010). In a theoretical project, we 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, submitted).

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 have focused on transcription but this provides only a partial view of an organism's regulatory apparatus. We described additional levels of control in *E. coli* including kinases and metabolites (Seshasayee et al., 2010; Seshasayee and Luscombe, submitted). Using these data we assessed how the regulatory network intersects with the metabolic system (Seshasayee et al., 2008; Seshasayee et al., submitted).

Experimental work. Though based at a computational institute, 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 study in a prokaryote (Kharamanoglou et al., 2010) and we 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), we developed a technique, called iCLIP, to assess protein–RNA interactions at single nucleotide resolution (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.

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.

SELECTED REFERENCES

Konig, J., et al. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909-915.

Reimand, J., et al. (2010) Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res.* 38, 4768-4777.

Seshasayee, A.S.N., Fraser, G.M. and Luscombe, N.M. (2010) Comparative genomics of cyclic-di-GMP signaling in bacteria: post-translational regulation and catalytic activity. *Nucleic Acids Res.* 38, 5970-5981.

Vaquerizas, J.M., et al. (2010) Nuclear pore proteins NUP153 and Megator define transcriptionally active regions in the *Drosophila* genome. *PLoS Genet.* 6, e1000846.

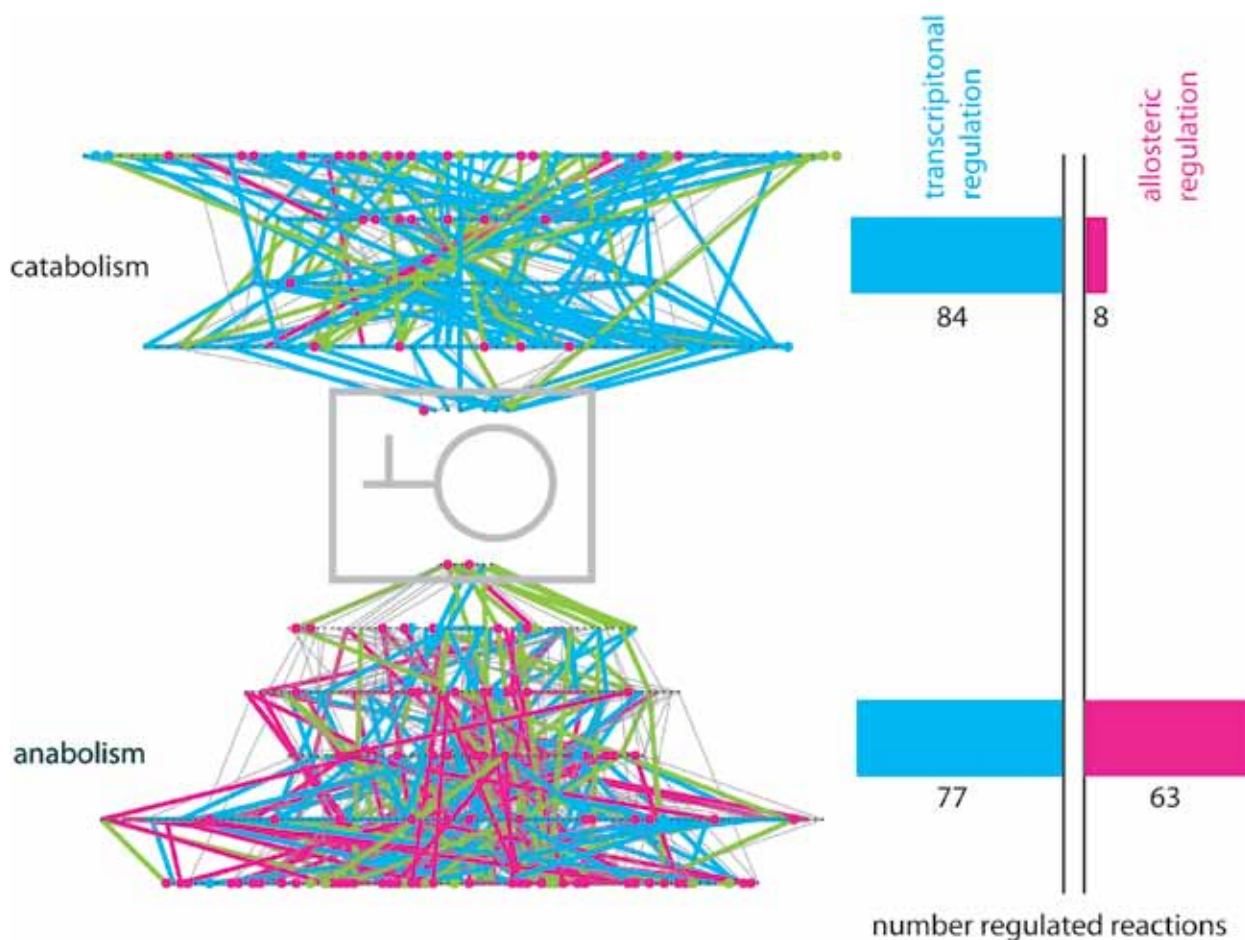


Figure. Network representation of the *E. coli* metabolic system. Nodes represent small molecules and edges depict enzymatic reactions. The reactions are coloured according to whether they are controlled transcriptionally (blue), allosterically (cyan) or by both methods (green). Allosteric feedback predominantly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



Dietrich Rebholz-Schuhmann

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

Literature Research

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. On-going 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) and all solutions contribute to the annotation and indexing of the full text scientific literature as part of the UKPMC and the SESL project;
- Finalised an ontological framework for phenotypic descriptions that offers new possibilities to represent disease phenotypes and and compare them (Hoehndorf et al., 2010);
- Developed and evaluated new solutions for the characterisation of protein splice variants from the literature and the extraction of gene regulatory events;
- Processed full text documents from major publishers – as part of the SESL project – 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;
- Launched the Journal of Biomedical Semantics (JBMS).

MAJOR ACHIEVEMENTS

Named Entity Recognition

Standardisation of the scientific literature: UKPMC, LexEBI, and CALBC

Delphine Bas, Adam Bernard, Abhishek Dixit, Senay Kafkas, Jee-Hyub Kim, Vivian Lee, Ian Lewin, Chen Li, Maria Liakata, Menaka Naraysamy, Piotr Pezik, Rohit Rexa, Shyamasri Saha, Dolf Trieschnigg, Ying Yan, Antonio Jimeno Yepes

Our research focuses on identifying named entities (e.g. genes, proteins, diseases) from the literature and linking them to an entry in a reference database. Several solutions have been provided: LexEBI (a terminological resource), IeXML (an annotation framework for documents), Whatizit (an information-extraction infrastructure) and CALBC (an evaluation infrastructure). We are generating 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. During the reporting period several bioinformatics resources were incorporated into LexEBI (e.g. the BioThesaurus), which interlinks terms across all resources according to their similarity. LexEBI supports large-scale information extraction in the biomedical domain and has been integrated with IeXML and the EBI's information-extraction infrastructure for indexing of the full body of scientific literature (UKPMC project).

By harmonising outputs from automatic text mining solutions, CALBC project partners produced SSC-I, a large-scale, annotated biomedical corpus containing annotations from four semantic groups (chemical entities and drugs; genes and proteins; diseases and disorders; and species). SSC-I has been used for the First CALBC Challenge, wherein participants were asked to annotate the corpus with their annotation solutions (Rebholz-Schuhmann, Jimeno Yepes et al., 2010). As of July 2010, SSC-I delivers

1 121 705 annotations for 100 000 Medline abstracts. The annotations are sufficiently homogeneous to be reproduced with a trained classifier (F-measure of 85%).

KNOWLEDGE DISCOVERY

Novel solutions for discourse analysis

Maria Liakata, Shyamasri Saha

We have created an annotation scheme (CoreSCs) that distinguishes the following core categories within the discourse of a scientific publication: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. We have asked chemistry experts to use this scheme to annotate a corpus of 265 full papers in physical chemistry and biochemistry. We are training and testing machine-learning algorithms for the automatic classification of sentences in papers according to this annotation scheme. An SVM classifier featuring a linear kernel and a Conditional Random Fields (CRF) classifier achieve 48% and 50% accuracy, respectively. We achieved the best accuracy for Experiment and Background.

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. For example, in DECIPHER (which uses Ensembl resources) a patient or syndrome phenotype is expressed with a specific subset of the London Dysmorphology Database (LDDb) terminology. We formalised a framework for phenotypic descriptions that makes their semantics explicit, thus providing the means to integrate phenotypic descriptions with ontologies of other domains and offering a new capability to represent disease phenotypes and perform powerful queries (Hoehndorf et al., 2010).

Improving the extraction of complex regulatory events from scientific text using ontology-based inference

Jung-Jae Kim

The extraction of complex events from biomedical text requires in-depth semantic analysis. We developed a system that deduces implicit events from explicitly expressed events using inference rules that encode domain knowledge. We evaluated the system with the inference module on different tasks and found that the inference based on domain knowledge plays a significant role in extracting complex events from text. This approach has great potential to recognise the complex concepts of biomedical ontologies (e.g. Gene Ontology) in the literature.

IT infrastructure development for information extraction

The SESL Triple Store: retrieval over large literature content

Samuel Croset, Christoph Grabmüller, Silvestras Kavaliauskas, Chen Li, Darius Sulskus

As part of the SESL project, the Pistoia Alliance (in collaboration with Nature Publishing Group, Elsevier, Oxford University Press and the Royal Society of Chemistry) produced an RDF Triple Store representation to simultaneously query the publishers' content and bioinformatics data resources using the RDF query language (SPARQL). The Triple Store delivers gene–disease associations from the scientific literature to the users through several interfaces: SPARQL queries, SOAP web services and a graphical user interface. The Triple Store contains about 14.5 million triples from the scientific literature that have been aligned with content from the Gene Expression Atlas (182 840 triples) and UniProtKb (12 552 239 triples for human). The RDF Triple Store enables simultaneous querying of the scientific literature and bioinformatics resources for evidence of gene–disease links.

SELECTED REFERENCES

Hoehndorf, R., Oellrich, A. and Rebholz-Schuhmann, D. (2010) Interoperability between phenotype and anatomy ontologies. *Bioinformatics* (in press). Published online 22 October; DOI: 10.1093/bioinformatics/btq578.

Rebholz-Schuhmann, D., E., et al. (2010) CALBC Silver Standard Corpus. *J. Bioinform. Comput. Biol.* 8, 163-179.

Rebholz-Schuhmann, D. and Nenadic, G. (2010) Biomedical Semantics: the Hub for Biomedical Research 2.0. *J. Biomed. Semantics* 1, 1.

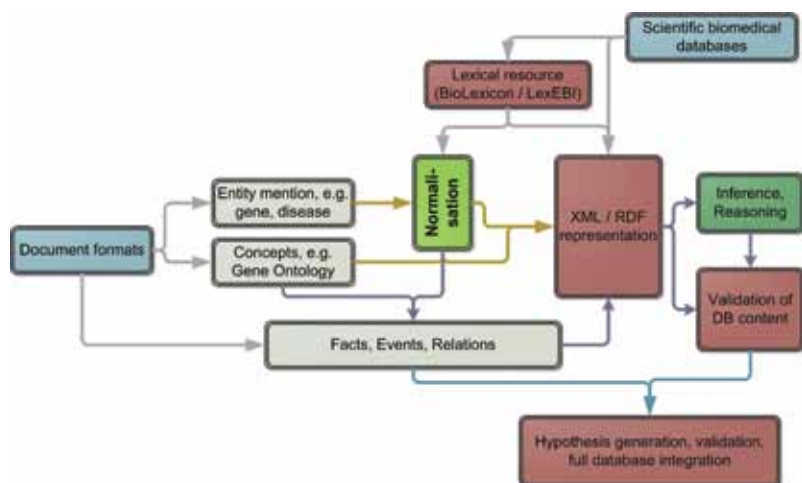


Figure: Literature analysis analyses scientific documents, identifies entities, concepts and facts (grey boxes) and normalises the entities to database entries with the support of a lexical resource (BioLexicon / LexEBI). RDF representations of the facts in combination of ontological resources supports inference and reasoning across the data content.



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.

Computational Biology of Proteins: Structure, Function and Evolution

DESCRIPTION OF RESEARCH

Our goal is to understand more about how biology works at the molecular level, with a particular focus on proteins and their three-dimensional (3D) structure and evolution. Currently we are exploring how enzymes perform catalysis and how these molecules recognise their cognate ligands. This involves gathering relevant data from the literature and developing novel software tools to characterise enzyme mechanisms and to 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). A further collaborative study on membrane proteins has focused on the structure of transmembrane pores and how this determines their different specificities. 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 gaining a deeper understanding of 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 and mice and relate these observations to effects on life span by combining information on function, context (i.e. pathways and interactions) and evolutionary relationships.

SUMMARY OF PROGRESS

- Continued analysis of enzyme mechanisms, now including metal and organic cofactors;
- Developed novel algorithms to compare enzyme reactions and to navigate through reaction (EC) space, and began using these tools to re-investigate enzyme classification;
- Completed a pipeline to characterise the evolution of enzyme families – including their sequences, specificities and mechanisms – and used it to analyse a large number of enzyme families;
- Almost completed a study of transmembrane pore geometries in all membrane proteins of known structure;
- Continued work on human mutation data, developing a pipeline to map new mutations on protein sequences and using their structures to try to understand or predict their effects on function;
- Continued to work with the functional genomics ageing data, trying to decipher key genes and their impacts on ageing.

MAJOR ACHIEVEMENTS

During the reporting period we completed a comprehensive analysis of enzyme cofactors (Fischer et al., in press), which are critical in about one half of all enzymes, providing a database detailing their two-dimensional and 3D structures, functions and usage in different enzymes. We found that most of these molecules are constructed from nucleotide and amino-acid-type building blocks, as well as some recurring cofactor-specific chemical scaffolds. We showed that while organic cofactors are on average significantly more polar and slightly larger than other metabolites in the cell, they cover the full spectrum of physicochemical properties found in the metabolome. Furthermore, we identified intrinsic groupings among the cofactors based on their molecular properties, structures and functions (Fischer et al., in press).

Using the tools for membrane-pore analysis we developed previously, we completed a study of the specificities of aquaporins, showing that this correlates with the electrostatic potential along the channel. Specific electrostatic fingerprints were observed for the aquaporin subfamilies (according to their permeability to water and/or glycerol) and for the potassium-channel family. This simple approach gives a better understanding of specificity in this important family of transporters and explains the effects on function of some disease-related mutations (Oliva et al., in press).

To understand how caloric restriction affects longevity, in collaboration with the Functional Genomics of Ageing Consortium at UCL, we showed in mice that deletion of ribosomal S6 protein kinase 1 (S6K1), a component of the nutrient-responsive mTOR (mammalian target of rapamycin) signalling pathway, led to increased life span and resistance to age-related pathologies such as bone, immune, and motor dysfunction and loss of insulin sensitivity (Selman et al., 2009).

The last major achievement to be published this year is a new computational tool to inspect domain combinations in protein families, called ArchSchema. It provides an excellent representation of all the domain combinations for a specific domain and relates these combinations to new biological functions. For example, one of the NAD binding domains is exceptionally promiscuous and combines with many different partners, giving rise to many different enzyme functions (see Figure)(Tamuri and Laskowski, 2010).

FUTURE PROJECTS AND GOALS

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 impacts on structure, function and disease. We will seek to gain a better understanding of reaction space and its impact on pathways. This will also 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 and combining human public transcriptome data sets with results from flies, worms and mice to explore effects related to human variation and age.

SELECTED REFERENCES

- Fischer J.D., et al. (2010) The structures and physicochemical properties of organic cofactors in biocatalysis. *J. Molecular Biol.* (in press).
- Oliva, R., et al. (2010) Electrostatics of aquaporin and aquaglyceroporin channels correlates with their transport selectivity. *Proc. Nat. Acad. Sci.* 107, 4135-4140.
- Selman, C., et al. (2009) Ribosomal S6 kinase 1 signaling regulates mammalian life span. *Science* 326, 140-144.
- Tamuri, A.U. and Laskowski, R.A. (2010) ArchSchema: a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics* 26, 1260-1261.

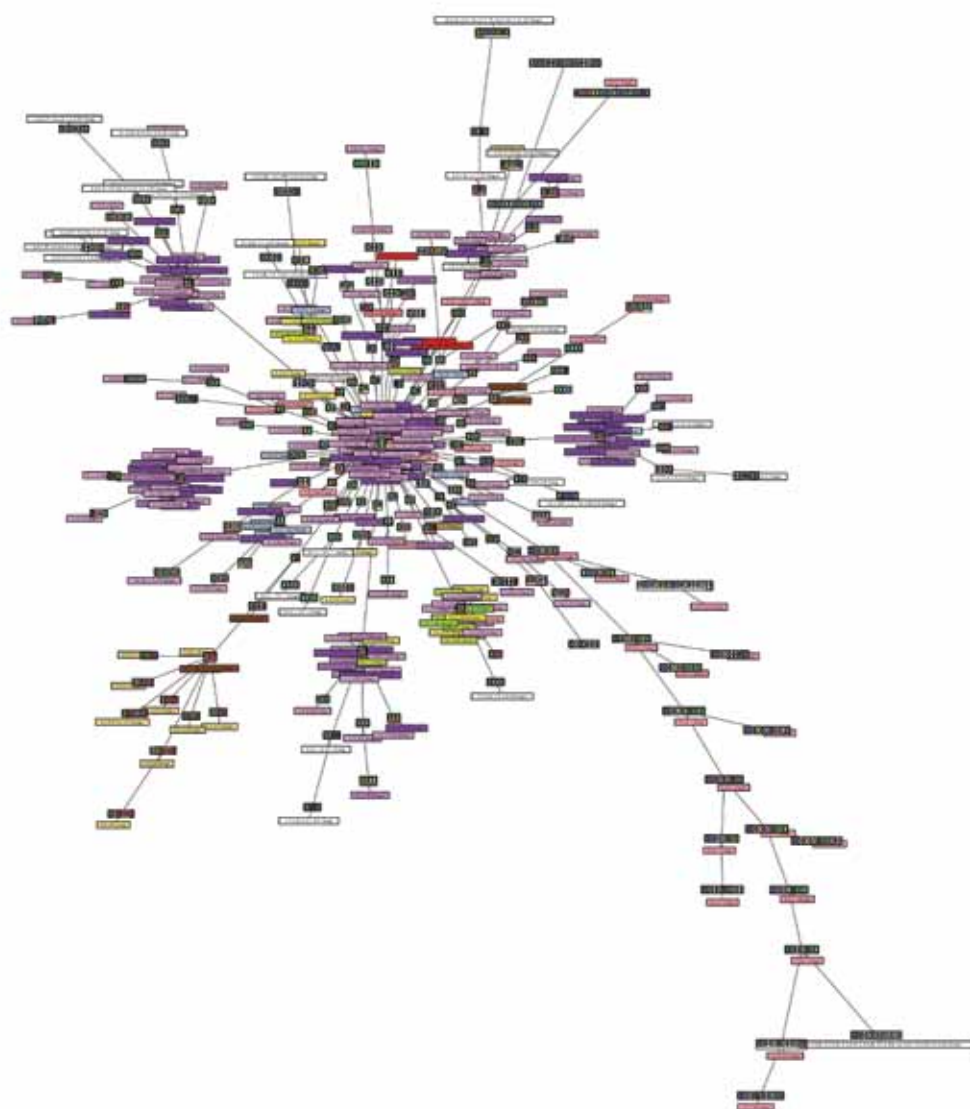


Figure. ArchSchema representation of multi-domain architectures containing a NAD(P)-binding, Rossmann-like domain. Each primary node in the graph shows a set of coloured bars, each corresponding to a structural domain, as identified by Gene3D. The NAD(P)-binding, Rossmann-like domains are shown as green bars. The central architecture of a single green domain represents all protein sequences in UniProt that contain just this domain. Linked to this are architectures having progressively more domains added. A red bar below any domain(s) indicates which parts of the proteins have 3D structural information in the Protein Data Bank (PDB). The coloured satellite nodes represent different enzymes' functions. Each colour corresponds to a different Enzyme Commission class; the deeper the hue, the more protein sequences there are. The network shows how function arises from addition of specific domains, or where a wide range of functions can be achieved by a single architecture.

The EMBL International PhD Programme at the EBI

The EMBL-EBI benefitted from the contributions of 43 PhD students between 1 July 2009 and 30 June 2010. Of these, four were formally awarded their PhDs and another seven defended their thesis. Students mentored in our four-year 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. Our strong partnerships with academic universities allow our students to obtain their degree jointly with EMBL and University of Cambridge. As a global centre of excellence for bioinformatics, EMBL-EBI offers its students a unique opportunity to participate in open and 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.



Alison Maria Meynert obtained her University of Cambridge PhD for her thesis, 'Function and evolution of regulatory elements in vertebrates', based on her work in Ewan Birney's group. The genomes of vertebrates share many non-coding sequences that are extremely well conserved among humans, mice and rats. Alison studied mammalian nonexonic ultra-conserved elements (UCEs), which have yet to be fully explained. Specifically, she explored whether the relative positioning and copy numbers of their transcription-factor binding sites are so important that mutations are simply not tolerated. Alison enhanced a framework for probabilistic modelling of transcription-factor binding by adding cooperative effects between binding sites. Using models from this framework, she found that the sequences effectively integrate variable concentrations of transcription factors into a logical on/off response. She

also demonstrated that the UCEs are likely enhancers containing multiple binding sites for transcription factors, which interact to activate their targets under specific conditions. She concluded that the arrangement of sites in the sequence is crucial to its function, hence its intolerance to mutation.



Aswin Sai Narain Seshasayee received a University of Cambridge PhD for his thesis, 'A computational study of bacterial gene regulation and adaptation on a genomic scale', which he researched in Nick Luscombe's group. Bacteria are marvellously adaptive, largely thanks to their precise and rapid regulation of protein production and activity in response to prevailing environmental signals. Using publicly available genomic and functional genomic data, Aswin conducted four comprehensive studies of various bacterial regulatory mechanisms and presented a thorough computational analysis. In one study, he examined DNA sequence, 3D topology and trans-acting proteins in bacterial transcriptional control to better understand the balance between DNA supercoiling and unwinding in transcription control.

Next, he focused on small molecules and metabolism regulation, clarifying some of the general principles guiding the deployment of two regulatory mechanisms that orchestrate genome-scale control of small-molecule metabolism in the bacterium *Escherichia coli*. A third study tackled second-messenger signalling by small molecules in controlling diverse cellular processes, including virulence. Finally, Aswin studied one- and two-component signalling mechanisms, and shed some light on how and why bacteria have evolved complex, multi-component signalling systems.



Kevin Nagel obtained his degree from University of Cambridge for his thesis, 'Automatic functional annotation of predicted active sites: combining PDB and literature mining', based on his work in Dietrich Rebholz-Schuhmann's group. The advent of functional genomics is giving rise to rapid growth in protein-structure models archived in the Protein Data Bank (PDB). Kevin developed data-mining tools to find patterns in the PDB as evolutionarily conserved configurations of protein residues, also called 'functional sites'. A result of such a mining is the identification of active sites, which is validated using the Catalytic Site Atlas. To support the biological significance of the mined result, Kevin harvested evidence of function from the literature using natural language processing techniques. He assessed the relevance of MEDLINE's abstract text and compared the extracted information on protein residues with annotations

from the on-going curation of UniProtKB. Also he examined the correlation of datasets arising from data- and text mining of PDB and MEDLINE, demonstrating that MEDLINE does not contain sufficient information to validate predicted functional sites. Kevin's novel solution impacts several aspects of protein bioinformatics research.



Katherine Lawler of Darwin College, University of Cambridge, received her joint EMBL-EBI degree based on her thesis, 'Transcriptional and post-transcriptional regulation of gene expression: computational analysis of microarray studies in fungal species', carried out in Alvis Brazma's group. DNA microarrays provide powerful means to identify changes in gene expression between different environmental conditions or developmental stages. Katherine's two computational studies of the genome-wide regulation of gene expression were based on the analysis of microarray datasets. In one, she explored the dynamics of a global gene-expression response. The regulation of mRNA abundance by both transcriptional and post-transcriptional control offers a range of possible strategies for shaping gene expression in response to a stimulus. Katherine investigated several such strategies, and tested the strength of evidence for regulated mRNA stability based on microarray time series in the fission yeast

Schizosaccharomyces pombe. She applied a dynamic model of mRNA abundance to simultaneous time series of mRNA abundance (DNA microarray) and transcription rate (RNA polymerase II ChIP-chip) datasets. This allowed her to identify candidate genes for which the gene expression response appears to be driven by a change in mRNA stability rather than by transcriptional control. Her second study combined expression analysis with transcription-associated proteins to identify genes co-expressed with putative DNA-binding transcription factors in *Fusarium graminearum*, a fungal crop pathogen. Gene clusters of co-expressed genes were identified and found to be related to secondary metabolism and *F. graminearum* pathogenicity.



PHD STUDENTS AT THE EBI, 1 JULY 2009 THROUGH 30 JUNE 2010

Joseph Foster	Chen Li	Tamara Steijger
Markus Fritz	*Lu Li	Sander Timmer
*Nils Gehlenborg	Inigo Martincorena	Diva Tommei
Angela Goncalves	Michele Mattioni	‡Daniela Weiser
Jose Afonso Guerra Martins Dos Santos Assuncao	**Alison Meynert	Ying (Yumi) Yan
Jacky Hess	Pablo Moreno	Judith Zaugg
Christine Hoyer	*Michael Mueller	Daniel Zerbino
*Garth Ilesley	**Kevin Nagel	Matthias Ziehm
Gregory Jordan	Anika Oellrich	
Myrto Areti Kostadima	Dace Ruklisa	
Felix Kruger	Petra Schwalie	* Defended thesis
**Katherine Lawler	**Aswin Sai Narain Seshasayee	** Completed thesis
	*Melanie Stefan	‡Affiliated solely with outside programme.

Research Groups

BERTONE GROUP

Group Leader
Paul Bertone

Staff Scientist
Mali Salmon-Divon

Postdoctoral Fellows
Pär Engström
Remco Loos

PhD Students
Heidi Dvinge*
Myrto Areti Kostadima
Tamara Steijger
Diva Tommei

Administrative Assistant
Kathryn Hardwick

Visitors
Yoshihiro Taguchi*

ENRIGHT GROUP

Group Leader
Anton Enright

Bioinformaticians
Stijn van Dongen

Postdocs
Cei Abreu-Goodger
Mat Davis
Harpreet Saini

PhD students
Nenad Bartonicek
Afonso Guerra-Assunção
Leonor Quintais

Visitors
Siegrun Zinca

GOLDMAN GROUP

Group Leader
Nick Goldman

Postdocs
Samuel Blaquart*
Ari Löytynoja
Tim Massingham
Emeric Sevin
Botond Sipos*
Martin Taylor*

Software Engineers
Hazel Marsden*
Nicolas Rodriguez (shared among multiple groups)

Team Administration

Tracey Andrew (shared among multiple groups)
Kathryn Hardwick (shared among multiple groups)

PhD Students
Jacky Hess
Gregory Jordan

LE NOVÈRE GROUP

Group Leader
Nicolas Le Novère

Coordinator
Camille Laibe

Software Engineers
Sarah Keating
Nicolas Rodriguez

Curators
Vijilashkimi Chelliah
Lukas Endler
Nick Juty

Postdocs
Melanie Stefan
Yang Zhan*

PhD Students
Benedetta Frida Baldi*
Christine Hoyer*
Lu Li
Michele Mattioni

Visitors
Stuart Edelstein
Catherine Lloyd*
Jean-Baptiste Pettit*
Karim Tazibt*

LUSCOMBE GROUP

Group Leader
Nick Luscombe

Staff Scientists
Annabel Todd*
Juanma Vaquerizas

Postdoctoral Fellows
Aswin Seshasayee
Kathikeyan Sivaraman*
Kathi Zarnack

PhD Students
Filipe Cadete*
Florence Cavalli
Inigo Martincorena
Judith Zaugg

REBHOLZ-SCHUHMAN GROUP

Group Leader
Dietrich Rebholz-Schuhmann

Staff Scientists
Christoph Grabmüller
Jee-Hyub Kim
Ian Lewin
Vivian Lee
Piotr Pezik*
Antonio Jimeno Yepes

Postdoctoral Fellows
Robert Hoehndorf*
Jung-Jae Kim*

Software Engineer
Silvestras Kavaliauskas*
Menaka Naraysamy

PhD Students
Adam Bernard*
Samuel Croset*
Chen Li*
Anika Oellrich
Ying Yan

Visitors
Maria Liakata*
Shyamasree Saha*
Dolf Trieschnigg*
Pinar Yildirim*

Visiting Students
Delphine Bas
Abhishek Dixit
Arun Gupta
Rohit Rexa
Darius Sulskus

THORNTON GROUP

Group Leader
Janet Thornton

Staff Scientists
Dan Andrews*
Pedro J. Ballester*
Tjaart de Beer *
Nicholas Furnham
Gemma Holliday
Roman Laskowski
Xun Li*
Irene Papatheodorou*
Marialuisa Pellegrini-Calace
Syed Asad Rahman
Daniela Wieser

PhD Students

Julia Fischer
Matthias Ziehm

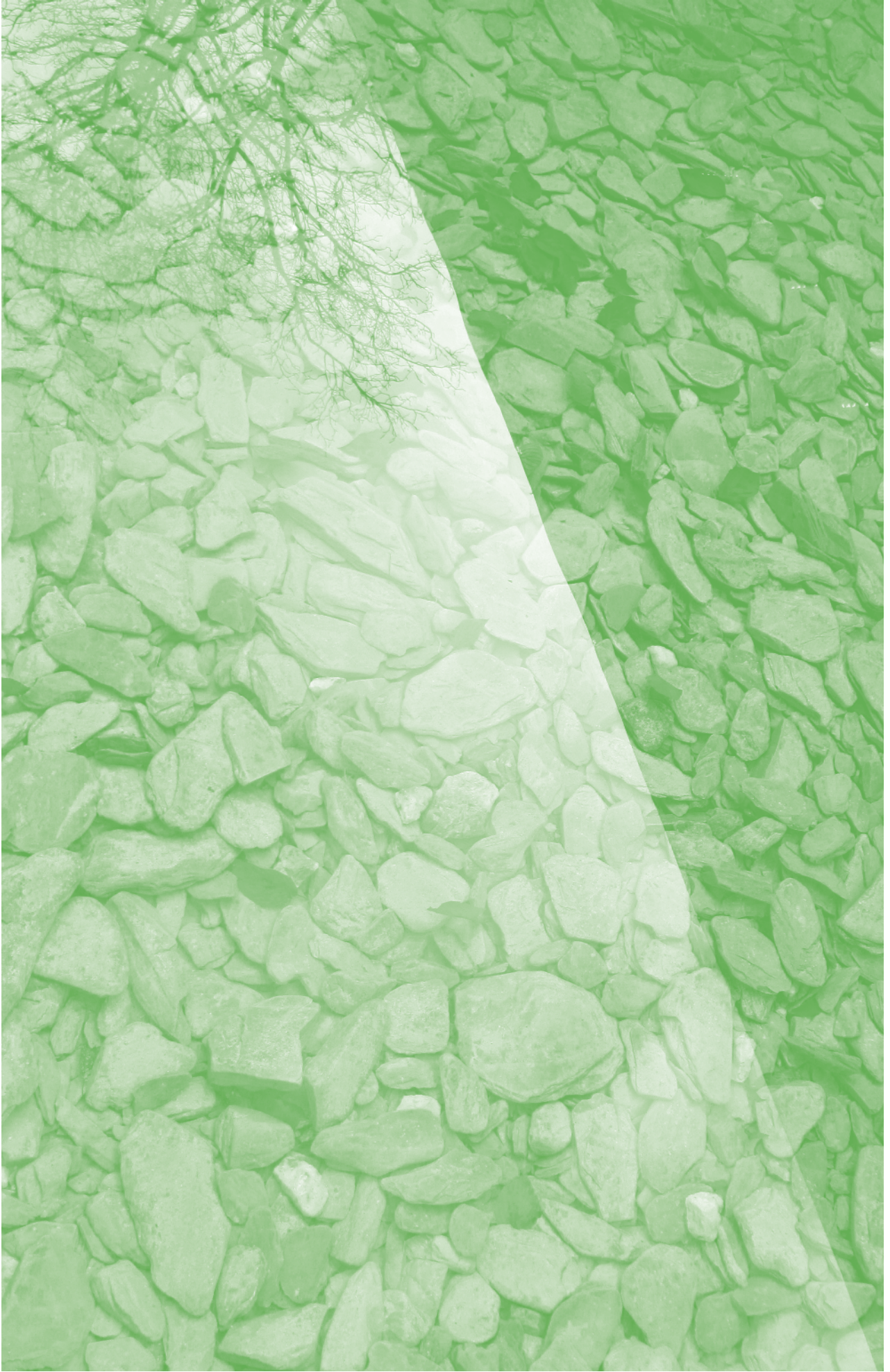
Personal Assistant to the Director
Helen Barker-Dobson

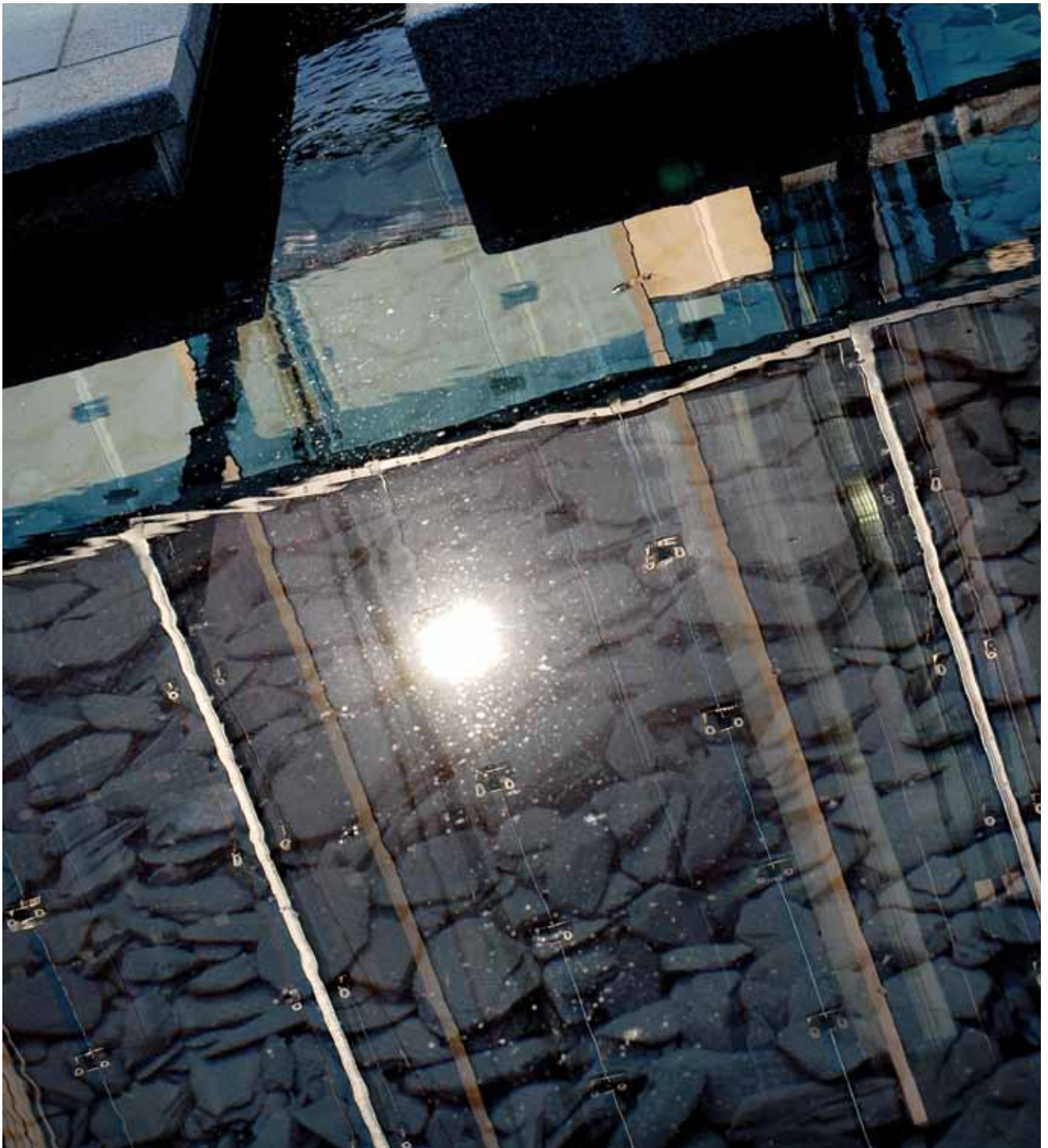
PA to the Director's Office
Stacy Schab

Visitors

Noa Berkovich
Amy Buchanan-Hughes
Angelica Datta
Hagit Eldar-Finkelman
Franz Fenninger
Saumya Kumar
Grecia Lapizco-Encinas
Sandya Tiwari
Gilleain Torrance
Mauno Vihinen
Sophie Williams

*Indicates part of the year only







Cath Brooksbank

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

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

Outreach and Training

DESCRIPTION OF ACTIVITIES

As part of the EBI's mission to serve the research community, our team engages with a wide range of people who need to know about what the EBI does. For example, our bioinformatics training programme empowers scientists at all career stages to make the most of Europe's core biological data resources and supports the wider bioinformatics training community. We engage with new user groups in academia and in industry; showcase EBI's career options; publicise the EBI's work via press campaigns and publications; engage with schools and the general public; and provide coherence to the EBI's growing and ever-diversifying range of events. We also build relationships with funders and policymakers. 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 147 EMBL-EBI personnel in running 467 events (26 demonstrations, 8 conference exhibitions, 177 courses/workshops, 57 posters, 248 presentations) and reached an audience of approximately 50 000 people in over 37 countries;
- Ran 10 hands-on training courses, serving a total of 319 trainees (average attendance, 32) – 35% had never used EBI's data resources before, most (>98%) said that they will use them in future and over 90% said they 'would recommend this training to colleagues';
- Delivered nine roadshows throughout Europe (fully funded through the EU-funded SLING Integrating Activity), serving 276 trainees (average attendance, 31) – over half had not used Europe's core data resources before, 96.5% of respondents indicated they 'will use the data resources in future' and 97.5% 'would recommend this training to colleagues';
- Delivered 14 roadshows for hosts not eligible for SLING funding (i.e. had previously hosted a roadshow or are situated outside Europe), serving 312 trainees (average attendance, 24) – over half had never used the resources before, 93.2% indicated they will use them in future and 95.5% 'would recommend this training to colleagues';
- Contributed towards a training platform for medicines research through EMTRAIN (an Innovative Medicines Initiative Education and Training Project) by helping to define quality criteria for courses; building a catalogue of continuing professional development; and collecting information on learning methodologies and tools that will enable course developers to create innovative training courses;
- Coordinated some internal training for EMBL PhD students, such as Primers for Predocs – a three-day intensive training course to familiarise the EBI's new PhD students with the data resources and tools available to them (follows on from a one-day course on EBI services in Heidelberg for all new PhD students);
- Issued 25 press releases and research highlights, resulting in coverage of EBI news in the scientific and general press;
- Hosted a successful London International Youth Science Forum, participated in the Cambridge Science Festival and supported the EBI Science and Society symposium;
- Promoted ELIXIR by creating targetted brochures for scientists and for funders – also took advantage of several opportunities to raise awareness of ELIXIR among its stakeholders.

MAJOR ACHIEVEMENTS

Our activities rely heavily on the hard work and dedication of many EBI employees, and although we do not have space to name them individually here we would like to thank all of those who have helped us to run a successful outreach and training programme this year.

Our roadshow programme, in which we train researchers by sending our trainers to their workplace, has been especially successful this year. EU funding through the SLING Integrating Activity has enabled us to train in countries with comparatively small research budgets, including several new EU Member States. Demand for this service has grown: 61 roadshows have been delivered since 2007, and by the end of the reporting period a further 13 had been scheduled for 2011 (see Figure). The roadshows attract new users to Europe's core data resources – fewer than half of our roadshow trainees have used the EBI's data resources before – and more than 95% of trainees tell us that they will continue to use our resources in the future. Our training events have a wide reach: close to 40% of hands-on training participants were UK nationals but over half travelled from outside the UK; the 45.5% who worked in the UK represented 45 different nationalities from 33 countries. Well over half (66%) of 'non-SLING' roadshow attendees hailed from outside Europe.

We made excellent progress towards developing our new eLearning portal, through an on-going project with the External Services Team. We developed a process for selecting an appropriate learning content-management system (Wright et al., 2010) and tested our top three choices to establish how well they met the requirements of trainees, trainers and systems administrators. In light of feedback from these three groups, we selected the open-source content-management system Drupal; following its installation the functionality of the portal has advanced substantially. In parallel, we developed guidelines to enable our trainers to produce consistent materials that are both pedagogically and scientifically sound, and worked with them on writing courses.

Accessing biological databases is now on the curriculum for the International Baccalaureate, highlighting the need for us to engage with teachers. In March 2010 we ran our first LearningLab for teachers in collaboration with the EMBL's European Learning Laboratory for the Life Sciences (ELLS). Through a combination of educational games, discussions and computer-based activities for direct use of biological databases in the classroom, trainers from EMBL, the Wellcome Trust Sanger Institute and UK National Centre for Biotechnology Education equipped 34 teachers with tools to bring the importance and the impact of bioinformatics into the classroom. Of these, 33 rated the event 'excellent' or 'very good'.

Our redesigned outreach web pages now clarify for visitors the EBI's mission and its key role as a coordinator of bioinformatics services in Europe, and a new press room provides a wide range of resources for the media.

FUTURE PLANS

Our users remain at the heart of our outreach mission. As bioinformatics begins to enter the classroom, public interest in personal genomics is on the rise. Over the coming year we will explore possibilities for engaging with a broader public to raise awareness of the EBI whilst continuing to serve our core users. We will continue working with schools via LearningLabs and other projects. On the training front, we will complete internal testing of the eLearning portal's functionality and launch a fully developed resource in 2011. Community engagement with other bioinformatics trainers will remain high on our agenda, as will working with ELIXIR's stakeholders to begin the construction of ELIXIR. The recruitment in 2010 of an External Relations Officer will enable us to build stronger relationships with funders and policymakers, which is vital to the ultimate success of ELIXIR.



Figure. Bioinformatics roadshows from the beginning of the roadshow programme in 2007 and including roadshows already scheduled for 2011. Blue pins are SLING-funded roadshows; red pins are non-SLING funded ones, showing that our SLING funding has enabled us to train in several new EU Member States. Inset: European roadshows.



Dominic Clark

*Industry Programme Manager
PhD in Medical Informatics, University of Wales, 1988.
Imperial Cancer Research Fund, 1987–1995.
GlaxoWellcome R&D Ltd., 1995–1999. Pharmagene,
1999–2001. Sagentia Ltd., 2001–2009. At EMBL-EBI
since 2006 (secondment 2006–2009).*



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.*

The EMBL-EBI Industry Programme

For the past 14 years the Industry Programme has been an integral part of EMBL-EBI, providing on-going and regular contact with key stakeholder groups. Established in 1996, the programme is now 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 actively support and encourage pre-competitive 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;
- Held nine workshops on topics of high importance to industrial partners;
- Supported pre-competitive projects within the context of the SESL pilot project;
- Promoted industrial involvement in ELIXIR;
- Organised an information workshop and hands-on training for small and medium-sized enterprises (SMEs);
- Coupled the Industry Programme to EMBL-EBI faculty and identified new collaborative opportunities;
- Welcomed two new partners to the programme: Astellas Pharma in Japan and Novo Nordisk in Denmark.



MAJOR ACHIEVEMENTS

During the reporting period we ran quarterly meetings in which we showcased recent developments at the EBI, reviewed progress on projects, prioritised future activities and presented our service development plans. Member companies can incorporate this information in their internal business planning processes. We also organised and ran nine training workshops on topics prioritised by the industry programme members.

A major remit of our programme is to foster pre-competitive 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 and development of new services. Once identified as being pre-competitive, a project plan is developed, including industry drivers and outcomes. Once funding is agreed, EMBLEM draws up a legal agreement and the partners define project governance and reporting procedures. During this reporting period, the Stand-alone Array Express Atlas project was completed and the Semantic Enrichment of the Scientific Literature pilot project (SESL) was initiated.

To further support pre-competitive work and the development of standards, we are active participants in IMI projects and members of the Pistoia Alliance. The IMI, funded by the EU and EFPIA, supports collaborative projects between the European pharmaceutical industry, academia, patient organisations and regulatory agencies. We are partners on the eTox (integrating bioinformatics and cheminformatics approaches for the development of expert systems allowing the in silico prediction of toxicities) project and the EMTRAIN (establishment of a network to facilitate and coordinate European, training and education relevant for stakeholders of medicines research and development) project.

The Pistoia Alliance is a not-for-profit, open foundation for data standards, ontologies and web services that aim to streamline the pharmaceutical drug-discovery workflow. EMBL-EBI is engaged in data standards-related activities and SESL. SESL focuses on integrating information about type II diabetes from the scientific literature and EMBL-EBI data resources using an information brokering architecture. The industry programme was pivotal to securing the commitment of publishing partners (Elsevier, Nature Publishing Group, Oxford University Press, the Royal Society of Chemistry), who provide the required full-text scientific literature.

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 participate in 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 a special focus on tools provided by the European Patent Office. The 2009 SME workshop was held in Munich with the co-operation of Bio-M and the European Patent Office (EPO). The event featured hands-on training, presentations, discussion and networking. Workshop topics were selected on the basis of previous interactions with SMEs and included, for example, chemogenomics, cheminformatics, proteomics, text mining, literature services and analysis, web services and patent services provided by EMBL-EBI and the EPO.

FUTURE PLANS

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 2011, the programme will also be more involved in IMI projects relating to knowledge management and the work of the Pistoia Alliance 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'.

Dates	Workshop Title
17-18 February	Open Biomedical Ontologies and controlled vocabularies: joint workshop with the OBO Foundry and the Pistoia Alliance
1-3 March	TACBAC 2010: Therapeutic Applications of Computational Biology and Chemistry
19-20 April	S4: A key component of the EBI integration strategy
17-18 May	Cheminformatics in R
21-22 June	Computational drug repositioning
22-24 September	Structural biology and PDBeMotif
11-13 October	Ontology engineering workshop
16-17 November	Toxicology ontology roadmap
30 November–2 December	Chemical structure resources



Petteri Jokinen

MSc in Computer Science, 1990, Helsinki University. At EMBL-EBI since 1996. Team Leader since 1998.

Systems and Networking

DESCRIPTION OF SERVICES/RESEARCH

The Systems and Networking team manages the EMBL-EBI's IT infrastructure. This includes compute and database servers, storage, desktop systems and networking, as well as managing our campus connection. Another important task is supporting EMBL-EBI users in their daily activities. The team works closely with all project groups maintaining and planning their specific infrastructures. The IT environment consists of 10 petabytes of disk storage and almost 9000 CPU cores. This year's growth has been enormous and subsequently so has the team's workload. In addition, we have been handling the large London Data Centre project. There has been a large quantity of equipment to install and maintain as well as an increasing number of people to support. We do our best to deploy high-grade solutions that enable the EBI scientist to work effectively without worrying about the IT solutions.

SUMMARY OF PROGRESS

- Began moving external services to London-based data centres;
- Built a 10-Gb/s private WAN that joins all of EBI's data centres;
- Consolidated the underlying storage architecture;
- Improved Oracle automation and MySQL virtualisation;
- Improved the resilience of EBI services across data centres;
- Built a new IMAP server farm that provides a more scalable email set-up;
- Developed a new approach to data storage accounting.

MAJOR ACHIEVEMENTS

We have started a significant new project, the aim of which is to move the majority of our external services to London-based data centres by the end of 2011. The move will improve the quality of the EBI services and will allow us to grow in the future. Production and research work will stay on the Hinxton campus. The project has been planned and is well on the way. We have rented space from two commercial data centres (March 2010) and equipped them with a large amount of new equipment (in total about 3600 CPU cores and 3.8PB of raw storage). These data centres operate in an active/active way, with some services running in both data centres at any given time and some being able to fail over. The first service that went live in June was the EBI FTP server.

The new data centres required significant re-working of the EBI's WAN infrastructure. We built a 10-Gb/s private WAN that joins all EBI data centres. In addition to the physical infrastructure we have implemented separate virtual networks for internal, external and management use. For internet access, we now peer directly with JANET at Harbour Exchange in London and have increased resilience of our internet access by adding a second 1-Gb/s backup link serving the London data centres. In addition, we have upgraded all of EBI's firewalls to improve security and throughput.

Once again storage doubled over the past year, from 6 000 TB to 10 000 TB (see Figure). As well as expanding our storage we are continuously trying to improve the underlying architecture. This year we have continued to consolidate to fewer storage solutions, thus making the management and use of the underlying architecture simpler for the inevitable future growth.

The Panda Team's Oracle database administrators have been moved to the Systems Team, and we have taken the responsibility of supporting these databases in addition to our existing duty of supporting MySQL – in the past we were responsible only for the underlying hardware and operating systems. We are actively improving automation of Oracle databases in order to serve our users more effectively, and optimising the hardware, operating systems and Oracle software components; these efforts will be considerably aided by the addition of two more database administrators to our team in late 2010. We have almost completed MySQL virtualisation (each MySQL instance runs on its own dedicated virtual machine), improving the reliability of individual database instances.

EBI's operation is becoming much more distributed. In order to provide resilience of EBI services across data centres it is desirable to implement Global Server Load Balancing (GSLB). This project, which is almost complete, will offer many benefits to existing services. GSLB also provides a useful infrastructure for EBI services wishing to build mirror sites elsewhere (e.g. in a cloud, at a partner/mirror site), as it permits geographically aware load balancing.

We built a new IMAP server farm, providing a more scalable email set-up and allowing us to move all existing accounts to the new servers. Because a large number of our storage servers are reaching the end of their lifetimes, we began migrating the data to new, scale-out servers. We are also finalising a new way of accounting our vast amount of storage, which will provide EBI management a better way of making budget decisions.

FUTURE PLANS

The London Data Centre project will continue into 2011. Although much of our team's work is complete, there is a lot to be done by various EBI teams before the project is complete. We will be attempting to provide unified solutions for 'safe network zones'. These will be used in two scenarios; sometimes to protect the rest of the EBI from a service (for example, a publicly open database) and sometimes to provide a solution for a service (for example, storing patient data) that needs higher security than the rest of the services. We face an increasing demand for bigger and faster MySQL databases, and are looking at appropriate, highly scalable solutions. Many EBI teams and users would like to have more Oracle database instances for testing and development. Traditionally it has been a manual process to clone databases. We will be looking how to automate this process in a way that minimises the deployment time and underlying storage requirements.

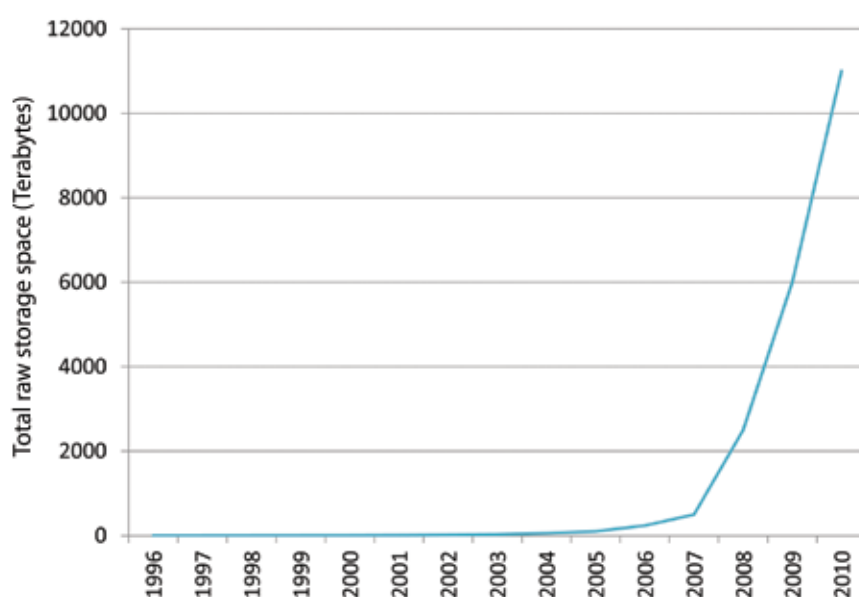


Figure. Raw storage (terabytes).



Rodrigo Lopez

*Veterinary Medicine Degree, 1984,
Oslo Veterinaer Hoyskole. Cand.
Scient. Molecular Toxicology and
Informatics, 1987, University of Oslo.
At EMBL-EBI since 1995.*

External Services

DESCRIPTION OF SERVICE

The External Services (ES) team at the EMBL-EBI focuses on three major areas: web development, web administration and service frameworks. Web development mainly involves the deployment and maintenance of core publishing frameworks for both the EBI and various EU-funded projects. Web administration concerns the provision of robust and stable service architectures for the EBI's databases and tools. Service frameworks comprise tools for generating core services (e.g. the EBI search engine) as well as applications for maintaining core analytical tools, especially in the domain of nucleotide and protein sequence analysis.

When the ES team was restructured during 2010, Technical Coordinators were appointed to each of the group's core activities to maximise efficiency. Brendan Vaughan now coordinates work on web development, Mickael Goujon on the service frameworks and Hamish McWilliam on data operations.

SUMMARY OF PROGRESS

- Engaged in the support of major migration of EBI services to new locations;
- Completed web-publishing consolidation and deployment of DRUPAL as the core EBI content-management system;
- Monitored, supported and reported on the use of EBI's services;
- Optimised web applications and web services.

MAJOR ACHIEVEMENTS

New Service Infrastructures in Hinxton and London

Moving the EBI services to London was a major achievement during 2010. EMBL-EBI has co-location agreements in place for two independent data centres in London: TeleCity Powergroup's PowerGate and Oliver Yard data centres. These will provide EMBL-EBI services with additional reliability and reinforced fail-over capacity. Also, direct connectivity to London's Harbour Exchange provides LINX connectivity to the Internet. These efforts will result in faster and more reliable services (99.98% availability) as well as improved scalability for EBI's core services.

The London move has posed two main transitional challenges: maintaining the current level of service and transferring data from Hinxton to the London data centres. This is an EBI-wide operation, involving developers and maintainers from all groups and teams. The Systems and Networking (SN) team is responsible for the physical infrastructure as well as for the main database operations; the ES team is responsible for the web architecture and day-to-day operations. At the core of the web architecture are Virtual Machine Environments, deployed in both Hinxton and London data centres. These are equivalent to private computing clouds and provide services for Oracle and MySQL database back ends as well as high-performance computing resources. This setup comprises 4000 CPU cores, physically distributed between the existing ES cluster in Hinxton and the London data centres.

The coordination of operations has been made possible by the implementation of an inventory of services (services database), which characterises each service in terms of its architecture, dependencies, running properties, developers and maintainers. At present the inventory contains 342 services, not all of which are slated to move to London. For example, submission services to the EBI databases (e.g. submissions to ENA, UniProt, ArrayExpress and PDB) and the corporate intranet are earmarked to operate from Hinxton only.

Web Development

The web publishing consolidation effort (described in the 2009 annual report), which involved the deployment of DRUPAL as the core of a new EBI-wide content management system (CMS), is complete. Currently, the main focus is importing websites from BCMS (described in the 2009 ASR) and providing training and support for staff in the day-to-day management of their web space.

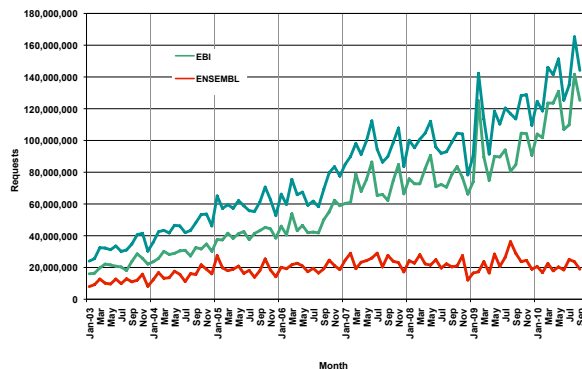


Figure 1. Web Traffic for www.ebi.ac.uk and www.ensembl.org since Jan. 2003.

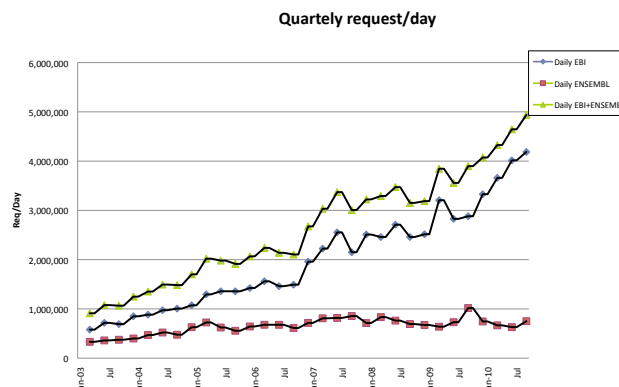


Figure 2. Quarterly requests per day for www.ebi.ac.uk and www.ensembl.org since January 2003

Because EBI is heavily involved in user-oriented services, User eXperience and Design (UXD) has become a critical component of the team's workflow. This aspect of web development focuses on the interaction between human users, machines and the contextual environments to design systems that address the user's experience. UXD methodologies are now integrated into the design process in order to meet support user needs and goals while also satisfying systems requirements and organisational objectives. Examples of where UXD is proving beneficial to the organisation include the avoidance of unnecessary features, design simplification, improving usability, streamlining the design and development effort and integrating business and marketing goals while catering to the needs of the scientific users.

The team is responsible for the main EBI web portal as well as several Wellcome Trust, BBSRC and EU- funded project websites. These include: the 1000 Genomes Project, BioCatalogue, the European Genome-phenome Archive, ELIXIR, ENA, ENFIN, SLING, IMPACT, INSDC, MIBBI and SYMBIOMatics. The ES team has also been heavily involved with the development of the new EBI e-learning portal, in collaboration with the Outreach and Training Team, as well as planning the revamp (or redesign) of a new EBI portal in the next reporting period.

Monitoring, Reporting and Supporting Use of EBI Services

Figure 1 shows web traffic volume to www.ebi.ac.uk and www.ensembl.org from 2003 to autumn 2010. Quarterly requests per day are shown in Figure 2. Monitoring and reporting on the health and activity of the EBI portals is also a responsibility of the ES team. The main components of these activities are new in-house developments, based on MySQL (EBI's web logs comprise the largest MySQL database at the EBI) that feeds into AWSTATS reports. A comprehensive set of health checks and workflows, based around the NAGIOS system monitoring package, ensure optimal accessibility for end-users.

Refactoring Web Applications and Web Services

As part of the transfer of services from Hinxton to London, development work has focused on the refactoring of web applications to ensure they make optimal use of new infrastructure and virtualisation opportunities. Examples of this work include the transition of job-dispatching services to a new, robust and fault-tolerant framework, JDISPATCHER (described in ASR 2009). For the most part, web applications are developed to run on localised single instances through the Hinxton data centre. In London, the developers and maintainers consider two tiers of services; as a result, most applications have been re-factored to take advantage of the improved load-balancing and fail-over capabilities these services offer. In this context, the responsibilities of the ES team include: education, advice on best practises, providing support in various programming languages, technical integration and advice on issue and change management.

Courses and Conferences

Team members were actively engaged in training activities during the reporting period both at the EBI and elsewhere.

June 2010	Wellcome Trust Summer School, Hinxton, UK EMBRACE Workshop for Web Service Providers in Bioinformatics, DTU, Copenhagen, DK EBI Roadshow: Prague, Czech Republic
April 2010	EBI Roadshow: University of Florence, Italy
March 2010	EBI Hands-on: Plant Bioinformatics, Hinxton, UK
February 2010	EBI Hands-on: Programmatic Access To Biological Databases (Perl), Hinxton, UK Patent Information User Group (PIUG) Meeting, Boston, USA
January 2010	Primers for Predocs, Hinxton, UK

Table. Courses and conferences attended by members of the External Services Team. Information on 2009 can be found in the 2009 Annual Report.

Support Teams

OUTREACH AND TRAINING

Team Leader

Cath Brooksbank

Research and Training

Coordinator

Nick Goldman

Training Programme Project

Leader

Vicky Schneider

Senior Scientific Training Officer

James Watson

eLearning Content Developer

Victoria Wright

Training Info & Liaison Officer

Claire Johnson

Outreach Programme Project

Leader

Louisa Wood

Scientific Outreach Officer

Katrina Pavelin

Workshop and Exhibitions

Organisers

Alison Barker

Holly Foster

Johanna Langrish*

Frank O'Donnell*

INDUSTRY PROGRAMME

Team Leader

Dominic Clark

Research Lead for

Industrial Interactions

John Overington

Administrator

Delphine Gandelin

SYSTEMS AND NETWORKING

Team Leader

Petteri Jokinen

Server and Networking

Jonathan Barker

Elizabeth Beresford

Gianluca Busiello

Dawn Johnson*

Gavin Kelman

Jenny Martin*

Manuela Menchi

Pravin Patel

Asier Roa*

Radoslaw Ryckowski

Michal Wieczorek

Desktop

William Barber

Richard Boyce

Karen Briggs

Andy Cafferkey

John Livingstone*

Systems Database

Administrators

Andy Bryant*

Mike Donnelly*

Luis Figueira*

Pieter Van Rensburg*

Software Engineer

Ville Silventoinen

Technical Administrator

Carolina Bejar

EXTERNAL SERVICES

Team Leader

Rodrigo Lopez

Software Engineers

Mickael Goujon

Weizhong Li

Hamish McWilliam

Eric Nzuobontane

Juri Paern

Silvano Squizzato

Franck Valentin

Web Developers

Asif Kibria

Thomas Laurent

Gulam Patel

Stephen Robinson

Brendan Vaughan

Peter Walker*

Francis Rowland*

Web Systems Administrators

Jenny Martin (Systems and

Networking team)

Rober Langlois

Philip Lewis*

Dietmar Sturmayer

Support and Training

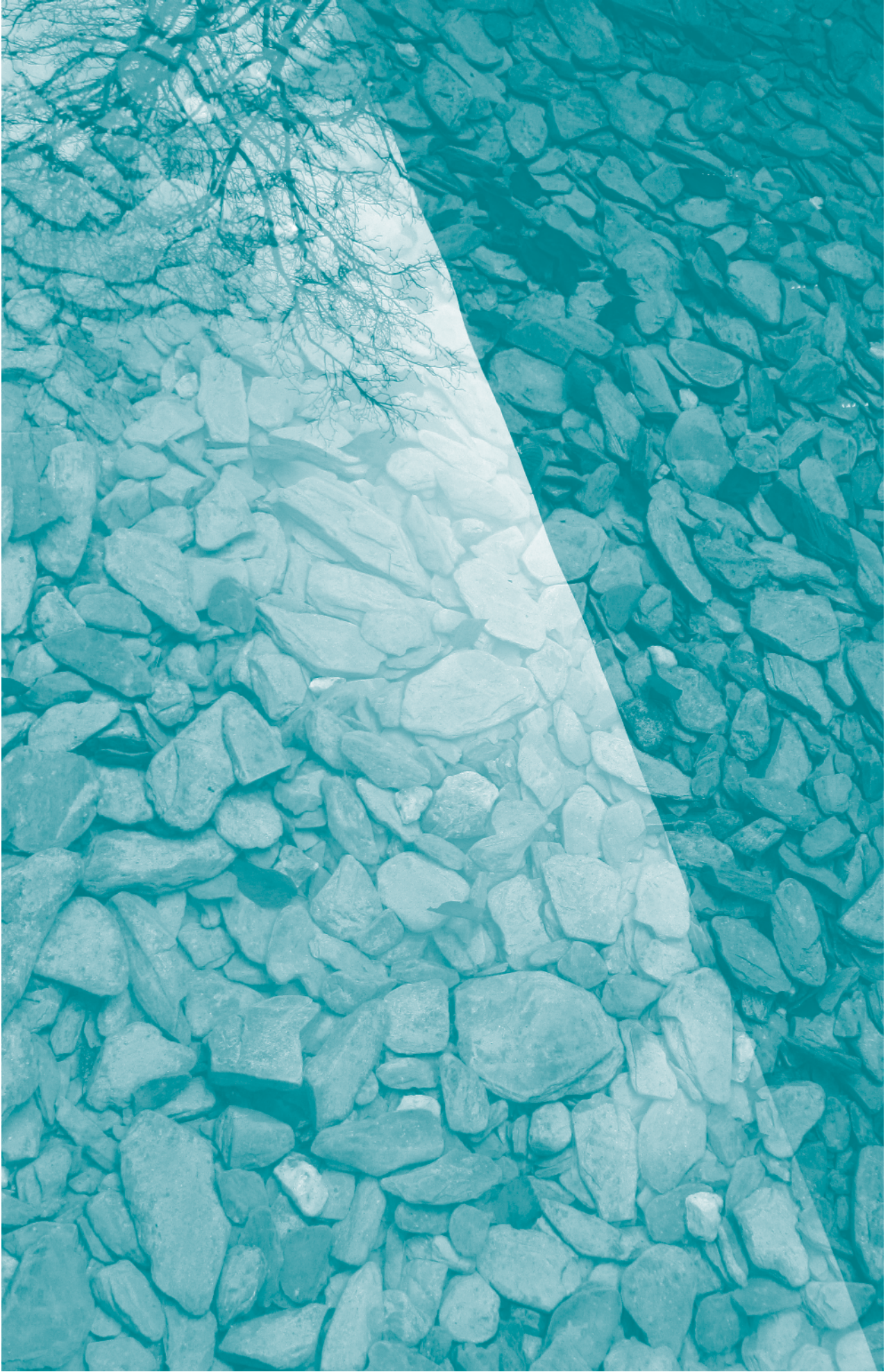
Andrew Cowley*

Jennifer McDowal*

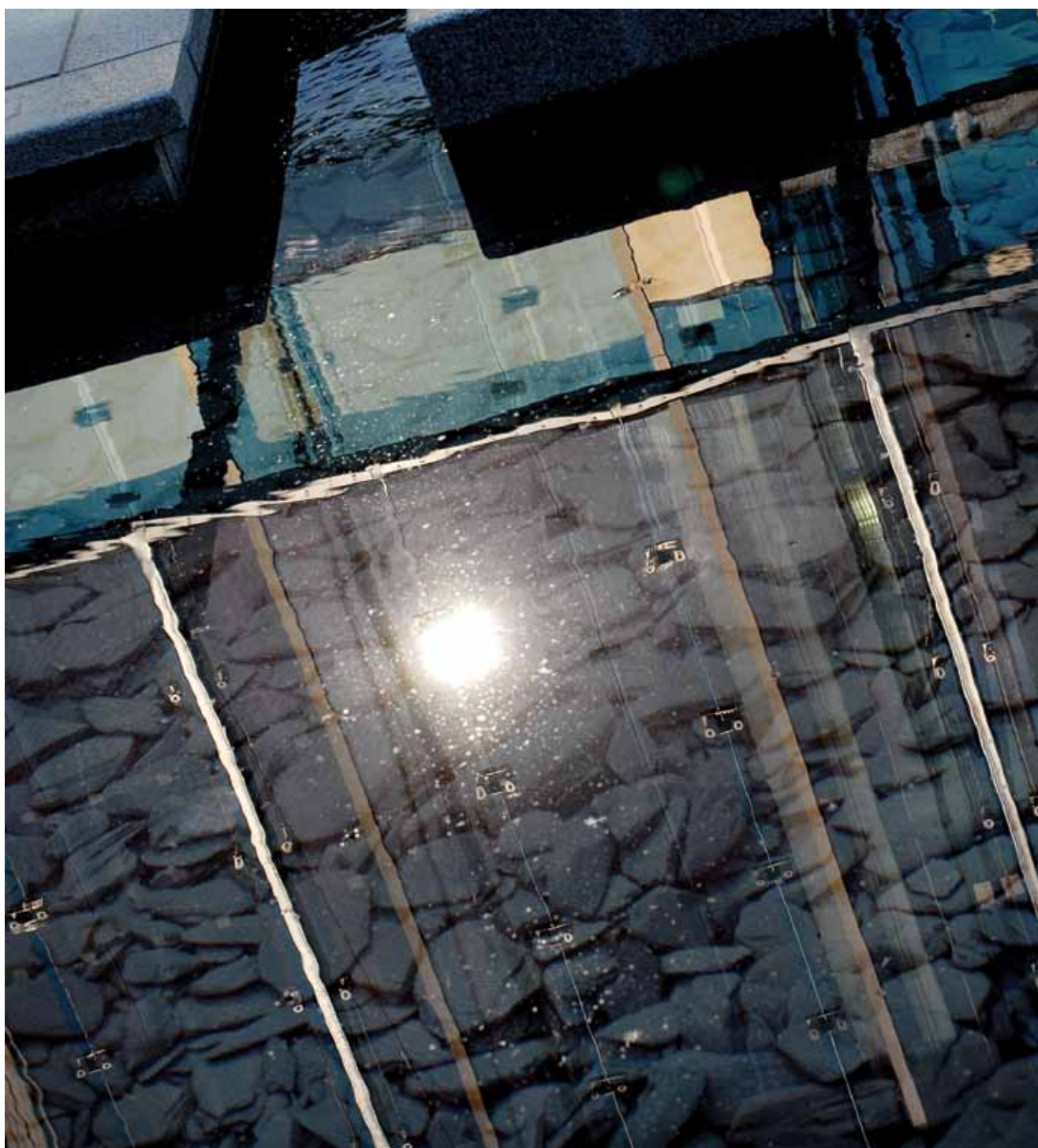
User Experience Analyst

Jenny Cham

*Indicates part of the year only



Facts and Figures



A Year in Numbers

SERVICES

The EBI hosts the major, core biomolecular resources of Europe – collecting, archiving and distributing data throughout Europe and beyond. The services continued to be well used during 2010 (see External Services, page 88). By June 2010 there were on average 4.0 million requests per day (compare to 2.9 million in 2009); this figure rises to 4.6 million when Ensembl is included (see Figure 1).

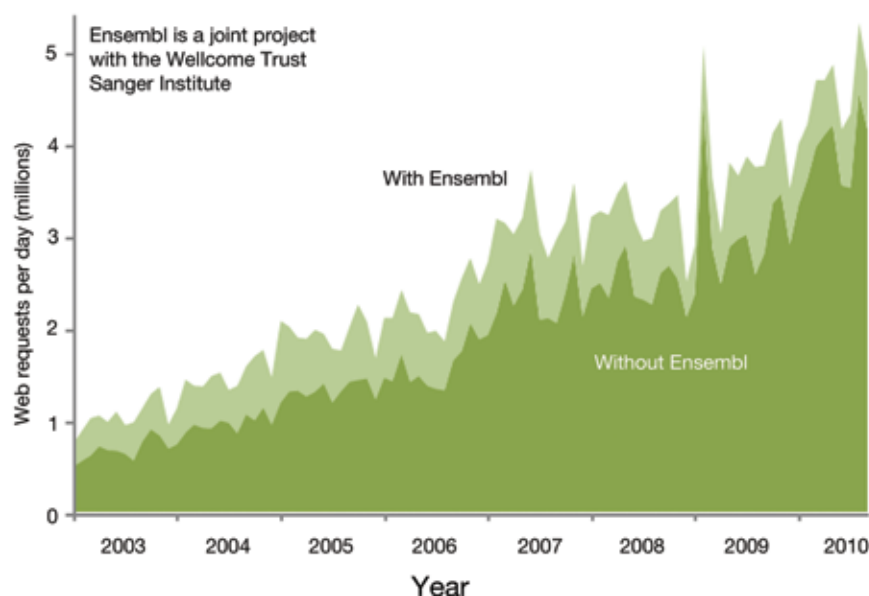


Figure 1. Web requests received by the EBI and Ensembl from January 2003 to June 2010.

From July 2009 to June 2010 all our core data resources grew significantly (Figure 2). EMBL-Bank received and processed more than 2.9×10^{11} bases (2.4×10^{10} in 2009). In total, the European Nucleotide Archive (ENA) now contains 4.31×10^{13} bases (2.27×10^{13} last year) – an accumulation rate of more than half a million bases per second. We have processed 3.3 million UniProt entries (2.8 million last year); 209 782 microarray hybridisations (60 646 last year); 6958 macromolecular structures (7174 last year) and 78 new genomes in Ensembl plus Ensembl Genomes (155 last year; 2009 was unusual as Ensembl Genomes was launched; most of these genomes were migrated from Integr8). Ensembl now holds 50 eukaryotic genomes and Ensembl Genomes holds 232 non-vertebrate genomes.

RESEARCH

EBI staff published 253 papers between July 2009 and June 2010 (compare to 218 last year); 87 of these were by the research groups (70 last year).

As of 30 June 2010 there were 39 PhD students at the EBI (see page 76). Ten new PhD students joined us during the reporting period, and four were awarded their PhD.

The research group leaders successfully applied for external support, receiving research funding to the tune of €2.8 million over the next two to five years (compare to €1.3 million last year).

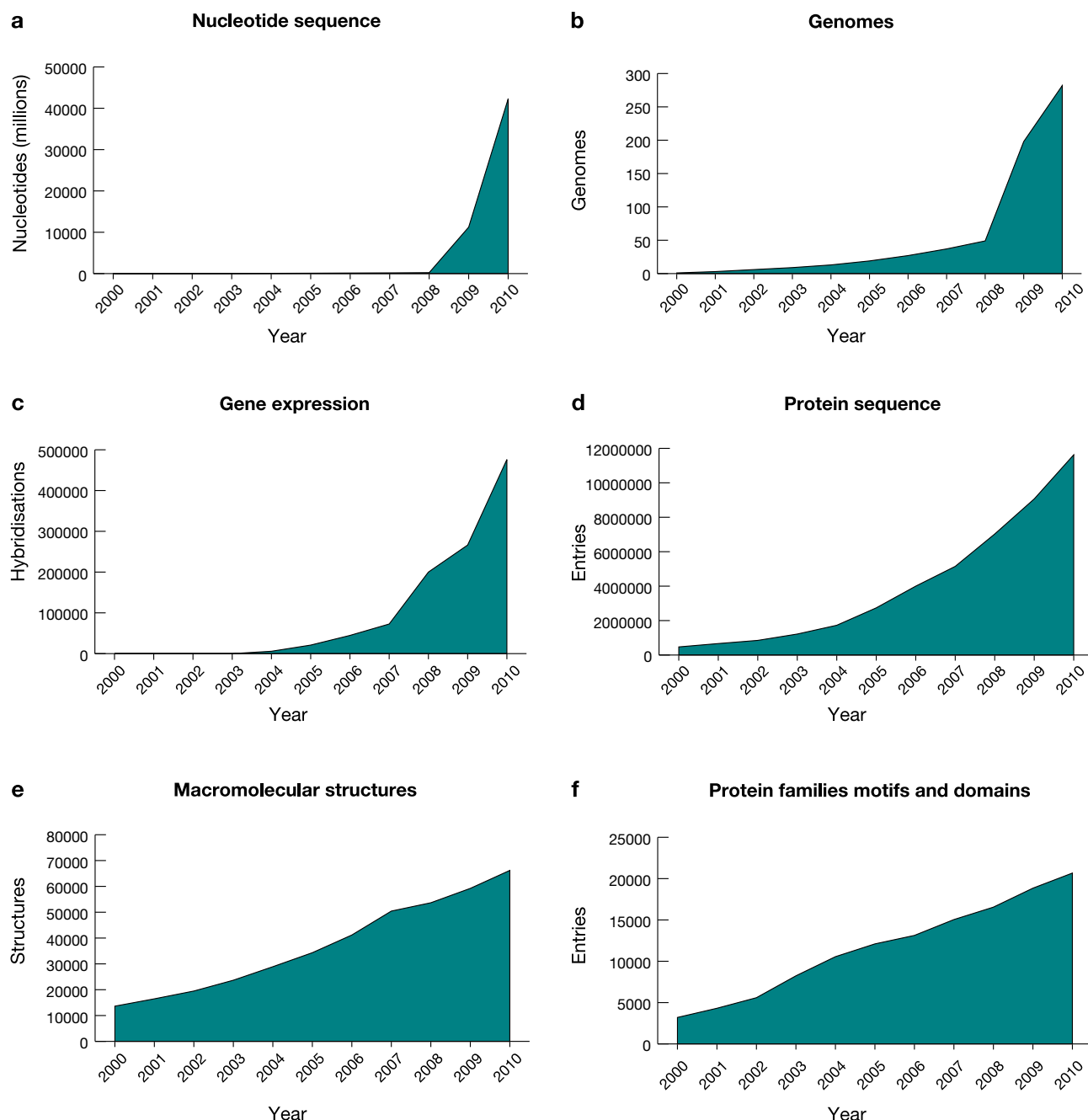


Figure 2. Growth of EMBL-EBI's core data resources, 2000–2010 (or from launch to 2010 if launch was after 2000). (a) Nucleotide sequence (bases in the European Nucleotide Archive); (b) genomes (entire genomes in Ensembl plus Ensembl Genomes combined); (c) gene expression (hybridisations in the ArrayExpress Archive); (d) protein sequence (protein sequences in UniParc); (e) macromolecular structures (structures in PDB); (f) protein families, motifs and domains (entries in InterPro).

OUTREACH, TRAINING AND INDUSTRY SUPPORT

We took part in 467 training events during the reporting period (280 last year), reaching over 50 000 participants in more than 37 countries.

The EBI hands-on training programme ran 10 courses for a total of 319 people between July 2009 and June 2010. The average number of participants was 32 (range, 20–40) per course. Delegates hailed from all over the world – over half (54.5%) came to us from outside the UK.

The training programme delivered nine roadshows under the auspices of the EU-funded SLING Integrating Action, serving 276 trainees (average attendance, 31; range, 19–43). The majority of attendees were from Europe.

A further 14 non-SLING-funded roadshows were delivered in the same period, serving 312 trainees (average attendance, 24; range, 20–30). Well over half (66%) of participants hailed from outside Europe.

Eleven workshops were run by the Industry Programme for its members, with an average of 52 delegates per event (range, 14–149).

The Industry Programme raised approximately €300 000 (£260 000) through pre-competitive projects.

STAFF

Our organisational structure (Figure 3) reflects the four parts of our mission: services, research, outreach, and training and industry support, with internal support facilitating all of these. EMBL-EBI personnel has grown by 13% (see Figure 4) from 406 in October 2009 to 461 in June 2010 (these figures exclude visitors), and retains its cosmopolitan flavour: our personnel (including long-term visitors) represents 37 countries (48 last year). During the reporting period we welcomed 28 visitors who stayed with us for longer than a month (compare to 47 visitors last year). This figure may be inexact, as we do not have a system for recording those who work with us without signing a contract.

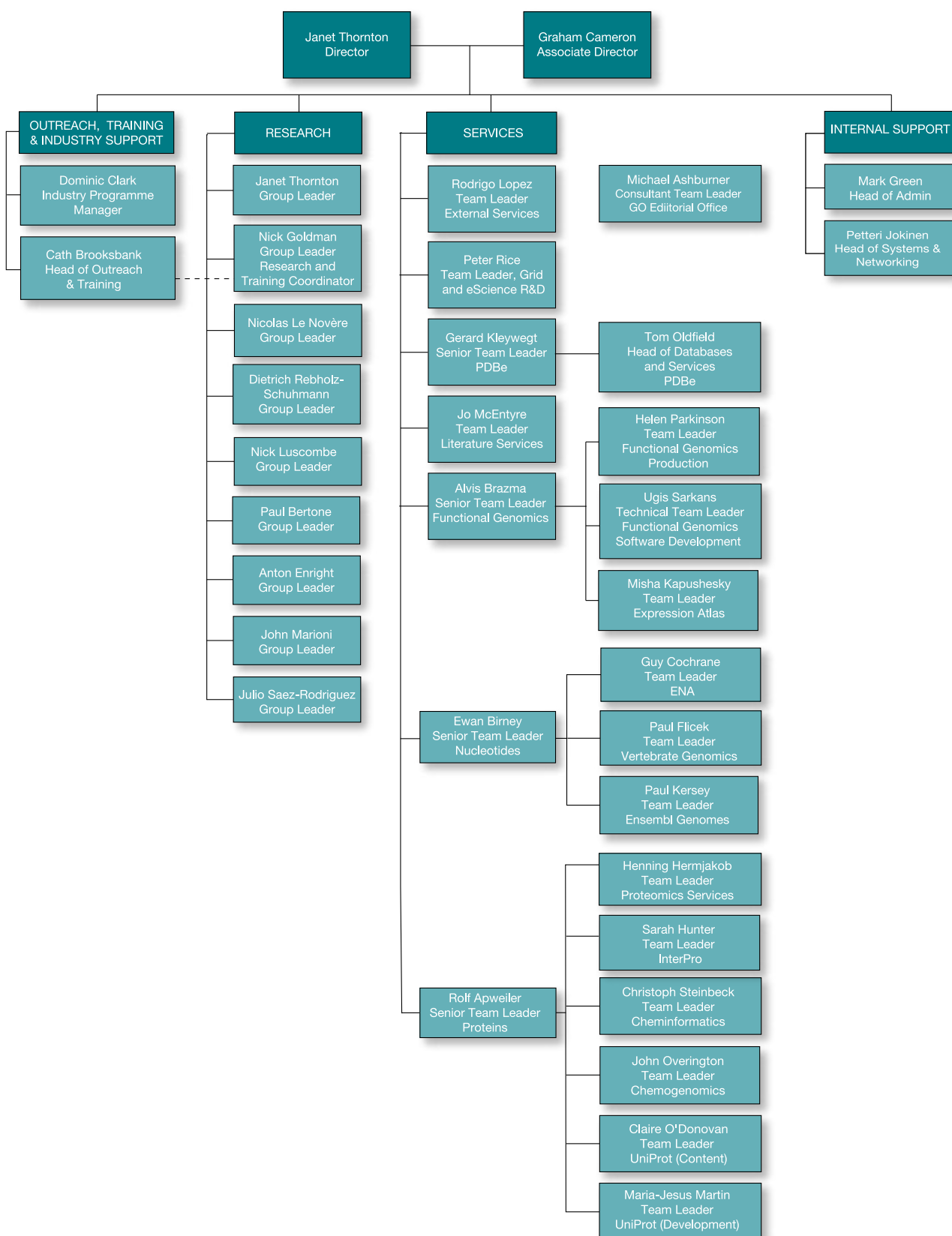


Figure 3. Organisational structure of EMBL-EBI as of June 2010.

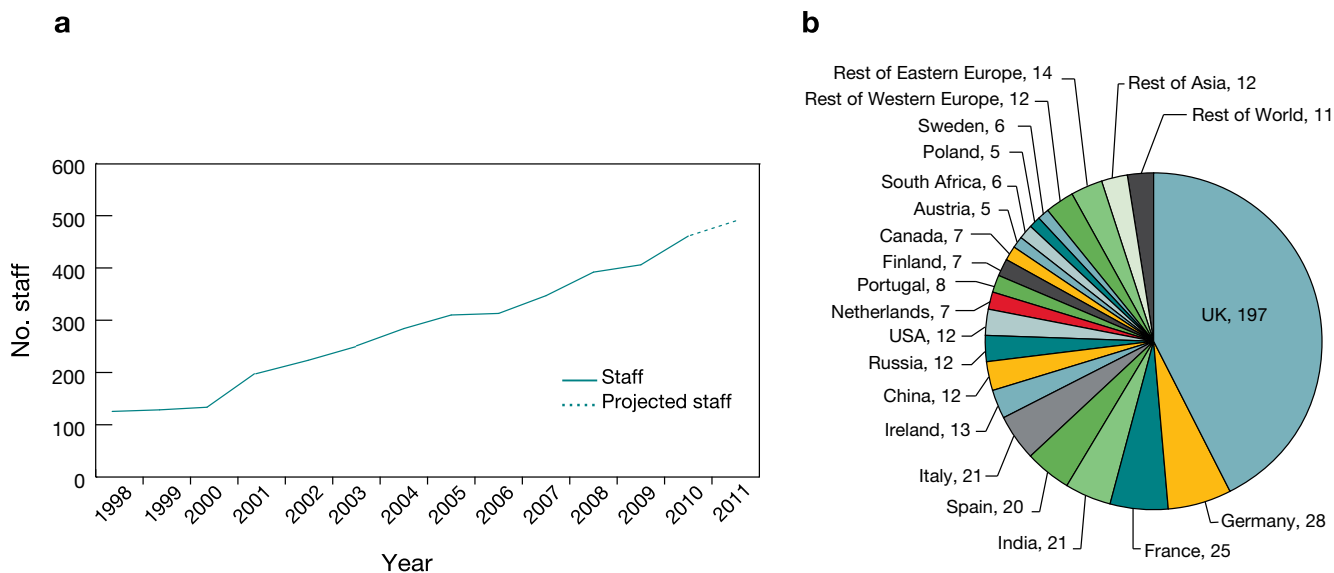


Figure 4. EMBL-EBI personnel. (a) Staff growth from 1998 to June 2010, and projected staff growth. (b) Nationalities of EMBL-EBI members of personnel as of June 2010.

COLLABORATIONS

Work at the EBI has continued to benefit from many collaborations (see Figure 5) and almost all of our resources are funded through collaborative agreements. During the reporting period, 77% of our publications involved collaborations with external colleagues from across the globe (compared with 80% last year).

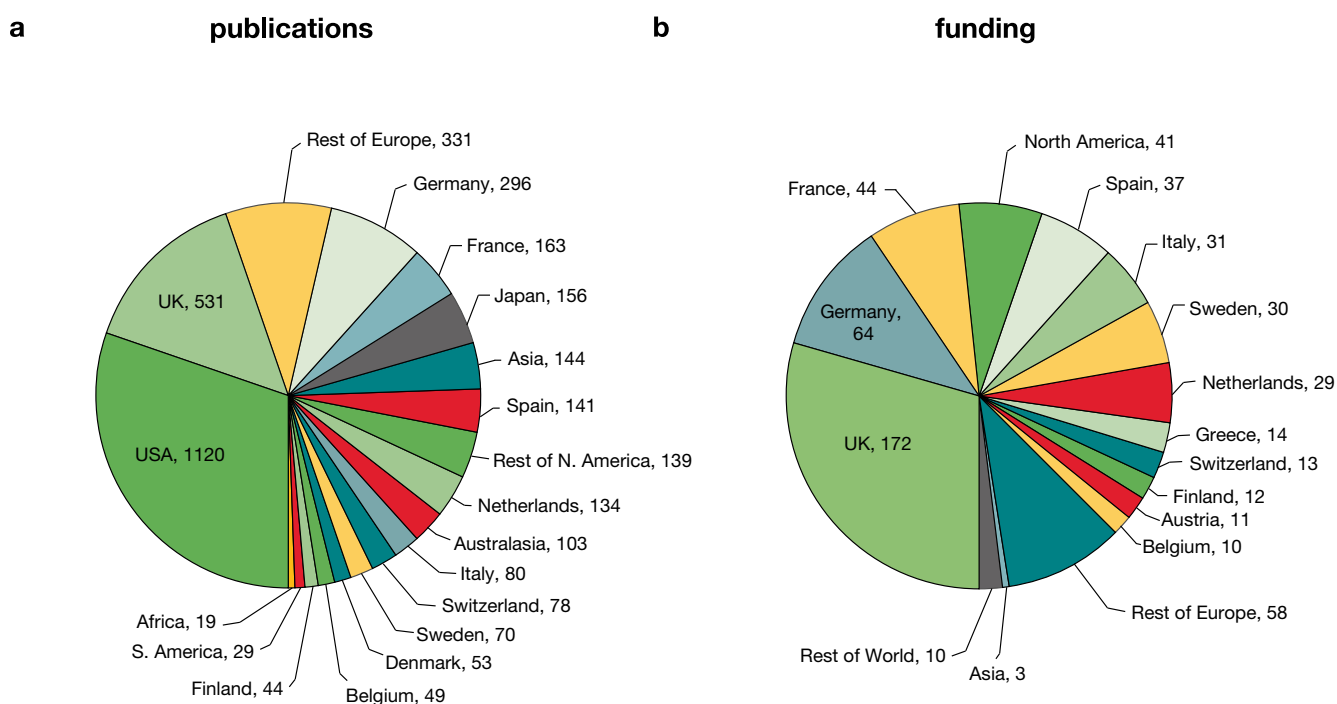


Figure 5. Collaborations as measured by (a) publications with other institutions and (b) funding shared with other institutions. Data for (a) were de-duplicated if the same institution appeared in the affiliations list of more than one paper. Data for (b) were not de-duplicated and in some cases the same institution is represented several times through different collaborations.



Figure 6. EMBL-EBI funding. (a) Growth of internal, external and total funds from 2001 to June 2010, and agreed internal funds for 2010–2011 (the end of the current EMBL indicative scheme). Asterisks indicate episodic capital funds: *Wellcome Trust, EMBL and Research Councils UK funding committed for EBI East Wing (11 M€); **Research Councils UK funding for ELIXIR hub (11.5 M€). (b) Sources of external funding for the year, as of October 2010.

FUNDING AND RESOURCE ALLOCATION

We raised €18.4 million in external funding for 2010, compared to €19 million in the last reporting period (see Figure 6a). This excludes approximately €11.5 million (£10 million) from the UK research councils towards ELIXIR's compute infrastructure.

Total internal funding to the EBI in 2010 was €22.4 million (€20 million last year), of which 58% was spent on salaries (47% last year; see Figure 7). We spent €3.0 million on computing equipment (€4.2 million last year), which represents 14% of our total internal spend (Figure 7a). Having invested heavily in new compute last year, our equipment spend from EMBL funds is considerably less this year. However, we have also spent £5 million from UK funding on compute and storage at the London Data Centre. Our storage capacity has increased from 5 PB to 10 PB but our compute power has remained stable at 9000 CPU cores.

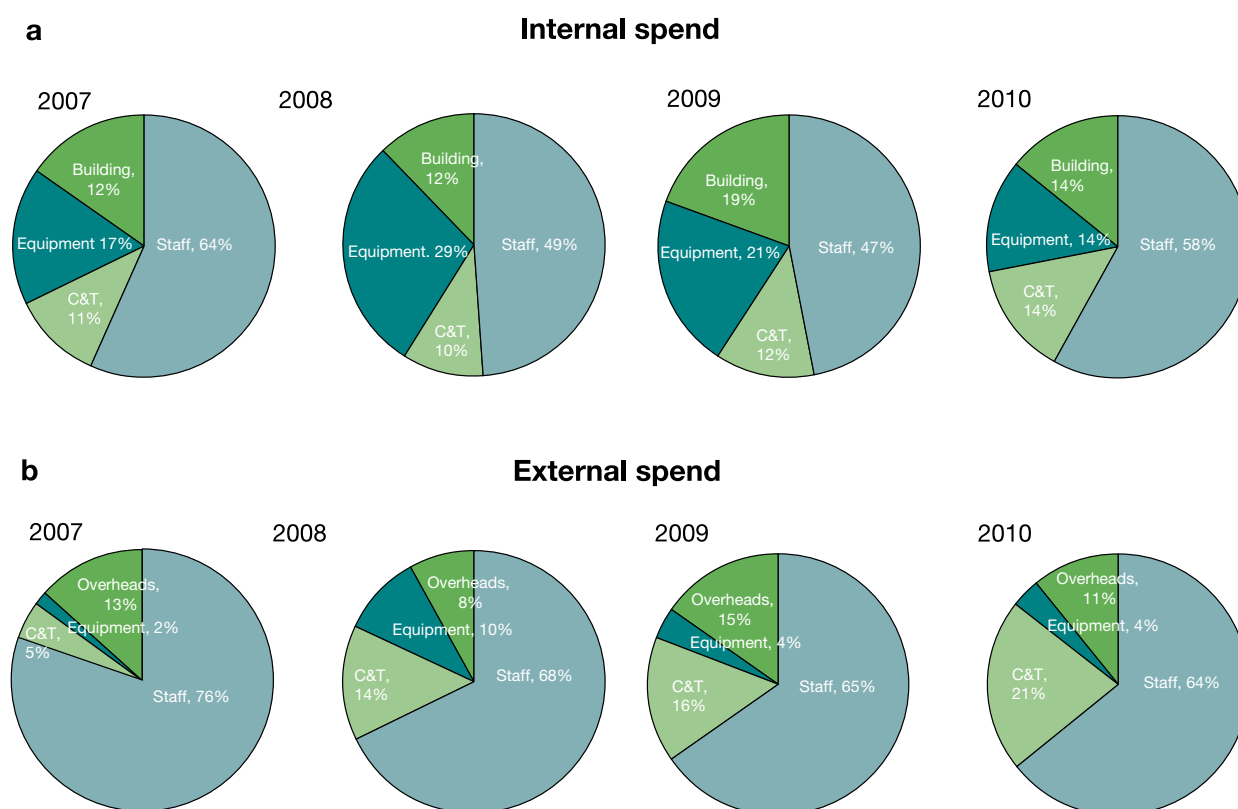


Figure 7. Breakdown of spend for 2005–2010. (a) Internal and (b) external spend. The equipment spend represented here does not include the capital funds spent on equipping the London Data Centre. 'Overheads' in the external spend represent estate costs related to externally funded staff.



Major Database Collaborations

This list shows 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

- ArrayExpress at EMBL-EBI, Hinxton, UK
- Dana Farber Cancer Institute, Boston, USA
- DDBJ Omics Archive, DNA Databank of Japan, Mishima, Japan
- Gene Expression Omnibus, NCBI, Bethesda, USA
- The Microarray Gene Expression Data Society, Stanford University, USA
- 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
- Gene Ontology at EMBL-EBI, Hinxton, UK
- 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

EMDB – Electron Microscopy Data bank (emdatbank.org)

- NCMI, Baylor College of Medicine, Houston, USA
- PDBe at EMBL-EBI, Hinxton, UK
- Research Collaboratory for Structural Bioinformatics (RSCB), USA

ENA - THE EUROPEAN NUCLEOTIDE ARCHIVE

- International Nucleotide Sequence Database Collaboration (www.insdc.org) partners
 - GenBank, NCBI, Bethesda, USA
 - Trace Archive, NCBI, Bethesda, USA
 - Sequence Read Archive, NCBI, Bethesda, USA
 - The DNA DataBank of Japan, National Institute of Genetics, Mishima, Japan
- Genomics Standards Consortium (www.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 EMBL-EBI and 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

- Central Aspergillus Data Repository, Manchester, UK
- 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 of EMBL-EBI; University of Notre Dame, South Bend, USA; Harvard University, Cambridge, USA; Institute of Molecular Biology and Biochemistry, Greece; University of New Mexico, USA; Imperial College, London, UK
- WormBase at California Institute of Technology, Pasadena, USA

THE GENE ONTOLOGY CONSORTIUM (www.geneontology.org)

- Agbase, Mississippi State University, Mississippi, USA
- The Arabidopsis Information Resource, Carnegie Institution of Washington, Stanford, USA
- Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, Berkeley, USA
- British Heart Foundation, University College London, London, UK
- Candida Genome Database, Stanford University, Stanford, USA
- DictyBase at Northwestern University, Chicago, USA
- EcoliWiki
- FlyBase at the University of Cambridge, UK
- GeneDB *S. pombe* and GeneDB for protozoa at the Wellcome Trust Sanger Institute, Hinxton, UK
- The GO Editorial Office, the Gene Ontology Annotation Project and Reactome at EMBL-EBI, 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
- Muscle TRAIT, University of Padua, Padua, Italy
- Plant-Association Microbe Gene Ontology, Virginia Polytechnic Institute and State University, Blacksburg, USA
- Rat Genome Database at the Medical College of Wisconsin, Milwaukee, USA
- Reactome at Cold Spring Harbor Laboratory, 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 – THE IMEX CONSORTIUM (imex.sourceforge.net)

- Centro Nacional de Biotecnología, Madrid, Spain
- DIP at the University of California, Los Angeles, USA
- IntAct at EMBL-EBI, Hinxton, UK
- 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
-

PDB – WORLDWIDE PROTEIN DATABANK (www.pdb.org)

- BioMagResBank, University of Wisconsin, Madison, USA
- PDBe at EMBL-EBI, Hinxton, UK
- 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
- PRIDE at EMBL-EBI, Hinxton, UK
- The Yonsei Proteome Research Center, Yonsei University, Seoul, Korea.

REACTOME

- New York University Medical Center, USA
- Ontario Institute for Cancer Research, Toronto, Ontario, Canada
- Reactome at EMBL-EBI, Hinxton, UK
- Reactome at Cold Spring Harbor Laboratory, USA

THE UNIPROT CONSORTIUM

- UniProt at EMBL-EBI, Hinxton, UK
- 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



Scientific Advisory Boards and Committees

EMBL Scientific Advisory Committee

Roberto Di Lauro, Naples, Italy (Chair)
 Sandra Schmid, Scripps Research Institute, La Jolla, USA (Vice-Chair)
 Geneviève Almouzni, Paris, France
 Siv Andersson, Uppsala, Sweden
 Konrad Basler, Zurich, Switzerland
 Barry Dickson, Vienna, Austria
 Reinhard Jahn, Göttingen, Germany
 Hiroaki Kitano, Tokyo, Japan
 Tom Muir, New York, USA
 Andrew Murray, Harvard, USA
 Helen Saibil, Birkbeck College London, UK
 Titia Sixma, Netherlands Cancer Institute, Amsterdam, the Netherlands
 Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy
 Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain

EMBL-EBI Bioinformatics Scientific Advisory Committee

Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy (Chair)
 Rob Cooke, GlaxoSmithKline Plc., UK
 Roderic Guigo, Centre de Regulació Genòmica, Barcelona, Spain
 Tim Hubbard, WT Sanger Institute, UK
 Olli Kallioniemi, VTT Medical Biotechnology, Turku, Finland
 Jonathan Knowles, The Roche Group, Basel, Switzerland
 Mathias Uhlén, Royal Institute of Technology (KTH), Stockholm, Sweden

ArrayExpress Scientific Advisory Board

Frank Holstege, University Medical Centre Utrecht, the Netherlands (Chair)
 Catherine Ball, Stanford Microarray Database, USA
 Richard Durbin, WT Sanger Institute, UK
 Roderic Guigo, Centre de Regulació Genòmica, Barcelona, Spain
 Christian Stoeckert, University of Pennsylvania, USA
 Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin, Germany

BioCatalogue Scientific Advisory Board

Jo Dicks, John Innes Centre, UK
 Robert Gill, Pronota NV, Belgium (previously at GSK)
 Steve Kemp, University of Liverpool, UK, Holger Lausen, SeekDa OG, Austria
 Terry Payne, University of Liverpool, UK
 Steve Pettifer, School of Computer Science, University of Manchester, UK
 Chris Rawlings, Rothamsted Research, UK
 Robert Stevens, School of Computer Science, University of Manchester, UK

Antoine H.C. van Kampen, Netherlands Bioinformatics Centre, the Netherlands
 Mark Wilkinson, University of British Columbia, Canada
 Anil Wipat, Newcastle upon Tyne, UK

BioModels Database Scientific Advisory Board

Jacky Snoep, Stellenbosch University, South Africa (Chair)
 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

ChEMBL/ChEBI: Chemistry Scientific Advisory Board

Steve Bryant, NIH, USA
 Edgar Jacoby, Novartis Institutes for BioMedical Research Discovery Technologies, Switzerland
 Andrew Leach, GlaxoSmithKline, UK (Chair)
 Tudor Oprea, University of New Mexico, USA
 Peter Willett, University of Sheffield, UK

EM Databank (EMDB) Advisory Committee

Joachim Frank, Columbia University, USA (Chair)
 Tao Ju, Washington University, USA
 Maryanne Martone, University of California San Diego, USA
 Andrej Sali, University of California, USA
 Paula Flicker, National Institute of General Medical Sciences, USA (Observer)
 Michael Rossmann, Purdue University, USA (Observer)

Ensembl Scientific Advisory Board

Michael Ashburner, University of Cambridge, UK (Chair)
 Søren Brunak, Technical University of Denmark, Lyngby, Denmark (co-Chair)
 Detlev Arendt, EMBL, Heidelberg, Germany
 Stephan Beck, UCL, London, UK
 Allan Bradley, WT Sanger Institute, UK (Observer)
 Michele Clamp, Broad Institute, USA
 Michael Eisen, University of California and Howard Hughes Medical Institute, Berkeley, USA
 Jim Kent, University of California Santa Cruz, USA
 Mark McCarthy, University of Oxford, UK
 Chris Ponting, University of Oxford, UK
 Nick Walton, University of Cambridge, UK
 Deepak Singh, Amazon Web Services, USA
 Johan den Dunnen, Leiden University Medical Centre, the Netherlands
 Michael Stratton, WT Sanger Institute, UK (Observer)

Ensembl Genomes Scientific Advisory Board

Julian Parkhill, WT Sanger Institute, UK (Chair)
 Detlev Arendt, EMBL, Heidelberg, Germany
 Mike Bevan, John Innes Centre, Norwich, UK
 Michele Clamp, Broad Institute, USA
 Michael Eisen, University of California and Howard Hughes Medical Institute, Berkeley, USA
 Steve Oliver, University of Cambridge, UK
 Jane Rogers, BBSRC Genome Analysis Centre, UK
 Doreen Ware, CSHL, USA

European Nucleotide Archive

Tim Hubbard, WT Sanger Institute, UK (Chair)
 Mark Blaxter, University of Edinburgh, UK
 Antoine Danchin, CNRS, Institut Pasteur, Paris, France
 Roderic Guigo, Centre de Regulació Genòmica, Barcelona, Spain

Jim Ostell, NCBI, USA
 Babis Savakis, University of Crete and IMBB-FORTH, Heraklion, Greece
 Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin, Germany
 Jean Weissenbach, Génoscope, Evry, France
 Patrick Wincker, Génoscope, Evry, France

European Nucleotide Archive (INSDC International Advisory Committee, European Members)

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

Gene Ontology Scientific Advisory Board

Lawrence Hunter, University of Colorado Health Sciences Center, Aurora, USA (Chair)
 David Botstein, Lewis-Sigler Institute, Princeton University, USA
 Philip Bourne, University of California, San Diego, USA
 Richard Scheuermann, University of Texas Southwestern Medical Center, Dallas, USA
 Michael Schroeder, Technische Universität Dresden, Germany
 Barry Smith, SUNY Buffalo, USA
 Simon Tavaré, University of Southern California, Los Angeles, USA and University of Cambridge, UK
 Michael Tyers, Scottish Universities Life Sciences Alliance, Edinburgh, UK

InterPro/Pfam Scientific Advisory Board

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

PDBe Scientific Advisory Board

Keith Wilson, University of York, UK (Chair)
 Udo Heinemann, Max Delbrück Centre for Molecular Medicine, Berlin, Germany
 Ernest Laue, University of Cambridge, UK
 Tomas Lundqvist, AstraZeneca Research and Development, Mölndal, Sweden
 Andrea Mattevi, University of Pavia, Italy
 Randy Read, University of Cambridge, UK
 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

Reactome Scientific Advisory Board

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

UniProt Scientific Advisory Board

Michael Ashburner, University of Cambridge, UK (Chair)
 Helen Berman, Rutgers University, USA
 Judith Blake, The Jackson Laboratory, USA
 Takashi Gojobori, National Institute of Genetics, Mishima, Japan
 Manuel Peitsch, Philip Morris International, Bern, Switzerland
 David Searls, Consultant, Philadelphia, USA
 Gunnar von Heijne, Stockholm University, Sweden

world wide Protein Data Bank (wwPDBAC) Advisory Committee

Stephen K. Burley, Eli Lilly, USA (Chair)
 Andreas Engel, University of Basel, Switzerland
 Masatsune Kainosho, Tokyo Metropolitan University, Japan
 Genji Kurisu, Institute for Protein Research, Osaka University, Japan
 Guy Montelione, Rutgers University, USA
 Randy J. Read, University of Cambridge, UK
 Michael G. Rossmann, Purdue University, USA
 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)
 Wah Chiu, Baylor College of Medicine, USA (Ex Officio)
 Manju Bansal, Education Research Network, India (Associate Member)
 Jianpeng Ding, Shanghai Institutes for Biological Sciences, China (Associate Member)

ELIXIR Committees

ELIXIR benefits from the deep involvement of a wide range of stakeholders representing universities, research institutes and industry in many countries throughout Europe. Because ELIXIR is still in its preparatory phase, its Scientific Advisory Board is yet to be finalised. However, we would like to acknowledge the commitment of its many committee members here.

ELIXIR Communities

Bengt Persson, Linköping University, Sweden (Chair)
 Nick Goldman, EMBL-EBI (co-Chair)
 Ron Appel, Swiss Institute of Bioinformatics, Switzerland
 Michael Ashburner, University of Cambridge, UK
 Carsten Carlberg, University of Luxembourg, Luxembourg
 Bernard de Bono, EMBL-EBI (representing Malta)
 Antoine de Daruvar, UB2 and representing MENESR, France
 Jan Gorodkin, University of Copenhagen, Denmark
 Elmars Grens, University of Latvia, Latvia
 Sampsa Hautaniemi, University of Helsinki, Finland
 Des Higgins, University College Dublin, Ireland
 Inge Jonassen, University of Bergen, Norway
 Lubos Klucar, Institute of Molecular Biology, Slovak Academy of Sciences, Slovakia
 Sophia Kossida, Biomedical Research Foundation, Academy of Athens, Greece
 Julian Parkhill, WT Sanger Institute, UK
 José Pereira-Leal, Instituto Gulbenkian de Ciencia, Portugal
 Andrei-Jose Petrescu, Institute of Biochemistry, Romania
 Vasilis Promponas, University of Cyprus, Cyprus
 Björgvin Richardsson, deCODE Genetics, Iceland
 Leszek Rychlewski, BioInfoBank Institute, Poland
 Dietmar Schomburg, Technische Universität Braunschweig, Germany
 Zlatko Trajanoski, University of Graz, Austria
 Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy
 Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain
 Antoine van Kampen, Netherlands Bioinformatics Centre, the Netherlands
 Yves van de Peer, University of Ghent–VIB Research, Belgium
 Ceslovas Venclovas, Institute of Biotechnology, Lithuania

Jaak Vilo, University of Tartu, Estonia
 Jiri Vohradsky, Institute of Microbiology, Czech Republic
 Blaz Zupan, University of Ljubljana, Slovenia

ELIXIR Data Resources

Graham Cameron, EMBL-EBI (Chair)
 Janet Thornton, EMBL-EBI (co-Chair)
 Gianni Cesarini, University of Rome, Italy
 Rob Cooke, GlaxoSmithKline Plc., UK
 Dawn Field, Centre for Ecology and Hydrology, UK
 Des Higgins, University College Dublin, Ireland
 Hans-Werner Mewes, National Research Centre for Health and Environment (GSF), Neuherberg, Germany
 Torsten Schwede, Swiss Institute of Bioinformatics, Switzerland
 Christopher Southan, Chris DS Consulting Ltd., UK
 Jean Weissenbach, Génomoscope, Evry, France

ELIXIR Financial

Alf Game, BBSRC (Chair)
 Dominic Clark, EMBL-EBI
 Paula Clements, Medical Research Council, UK
 Jean-Louis Coatrieux, INSERM, France
 Olivier Collin, Institut National de Recherche en Informatique et en Automatique (INRIA), France
 Sarah Collinge, Natural Environment Research Council, UK
 Jasper Diderich, Netherlands Organisation for Scientific Research, the Netherlands
 Michael Dunn, Wellcome Trust, UK
 Vera Herkommer, EMBL, Heidelberg, Germany
 Frank Laplace, Federal Ministry of Education and Research (BMBF), Germany
 Giuseppe Martini, Consiglio Nazionale delle Ricerche, Italy
 Hans-Werner Mewes, National Research Centre for Health and Environment (GSF), Neuherberg, Germany
 Julian Parkhill, WT Sanger Institute, UK
 Nikolai Raffler, Deutsche Forschungsgemeinschaft (DFG), Germany
 Marta Sabec-Paradiz, Ministry of Higher Education Science and Technology, Slovenia
 Silke Schumacher, EMBL, Heidelberg, Germany

ELIXIR Industry

Mark Forster, Syngenta, UK (Chair)
 Dominic Clark, EMBL-EBI (co-Chair)
 Ian Dix, AstraZeneca, UK
 Ian Harrow, Pfizer, UK
 Björgvin Richardsson, deCODE Genetics, Iceland

ELIXIR Interdisciplinary

Mark Forster, Syngenta, UK (Chair)
 Rolf Apweiler, EMBL-EBI (co-Chair)
 Endre Barta, Agricultural Biotechnology Centre, Hungary
 Erik Bongcam-Rudloff, Linnaeus Centre for Bioinformatics, Sweden
 Christine Gaspin, French National Institute for Agricultural Research (INRA), France
 Steve Goff, University of Arizona, USA
 Dietmar Schomburg, Technische Universität Braunschweig, Germany
 Christoph Steinbeck, EMBL-EBI

ELIXIR Interoperability

Amos Bairoch, Swiss Institute of Bioinformatics, Switzerland (Chair)
 Michael Ashburner, University of Cambridge, UK (co-Chair)
 Vincent Breton, CNRS-IN2P3, France
 Graham Cameron, EMBL-EBI
 Susanna-Assunta Sansone, EMBL-EBI
 Johan van der Lei, Erasmus Medical Centre, the Netherlands
 Gert Vriend, Radboud University Nijmegen Medical Centre (RUNMC), the Netherlands

ELIXIR Literature

Alfonso Valencia, Structural Computational Biology Group, CNIO, Madrid, Spain (Chair)
 Dietrich Rebholz-Schuhmann, EMBL-EBI (co-Chair)
 Matthew Cockerill, BioMed Central, London, UK
 Carole Goble, University of Manchester, UK
 Udo Hahn, Jena University, Germany
 Timo Hannay, Nature Publishing Group, London, UK
 Lawrence Hunter, University of Colorado Health Sciences Center, Aurora, USA
 Robert Kiley, Wellcome Trust, UK
 Manuel Peitsch, Novartis Institutes for Biomedical Research, Cambridge, USA
 Peter Stoehr, EMBL-EBI
 Jun'ichi Tsujii, NacTeM, UK

ELIXIR Medical

Johan van der Lei, Erasmus Medical Centre, the Netherlands (Chair)
 Alvis Brazma, EMBL-EBI (co-Chair)
 Erik Bongcam-Rudloff, Linnaeus Centre for Bioinformatics, Sweden
 Jean-Louis Coatrieux, INSERM, France
 Nathalie Costet, INSERM, France
 Paul Flicek, EMBL-EBI
 Antoine Geissbuhler, University of Geneva, Switzerland
 Jane Grimson, University of Dublin, Ireland
 Maria Krestyaninova, EMBL-EBI
 Jan-Eric Litton, Karolinska Institutet, Sweden
 Mark McCarthy, University of Oxford, UK
 Helmut Meyer, Ruhr-University Bochum, Germany
 Philippe Rocca-Serra, EMBL-EBI
 Ben van Ommen, Netherlands Organization for Applied Scientific Research (TNO), the Netherlands
 Gert-Jan van Ommen, Leiden University Medical Centre, the Netherlands
 Eero Vuorio, University of Turku, Finland

ELIXIR Organisational and Legal

Silke Schumacher, EMBL, Heidelberg, Germany (Chair)
 Nigel Watts, Medical Research Council, UK (co-Chair)
 Alf Game, BBSRC, UK
 Mark Green, EMBL-EBI
 Vera Herkommer, EMBL, Heidelberg, Germany
 Andrew Lyall, EMBL-EBI
 Janet Thornton, EMBL-EBI

ELIXIR Compute Infrastructure

Tim Hubbard, WT Sanger Institute, UK (Chair)
 Ewan Birney, EMBL-EBI (co-Chair)
 Phil Butcher, WT Sanger Institute, UK
 Mark Green, EMBL-EBI
 Petteri Jokinen, EMBL-EBI
 Andrew Lyall, EMBL-EBI
 Janet Thornton, EMBL-EBI

ELIXIR Steering Committee

Janet Thornton, EMBL-EBI (Chair)
 Andrew Lyall, EMBL-EBI (coChair)
 Ron Appel, Swiss Institute of Bioinformatics, Switzerland
 Søren Brunak, Technical University of Denmark, Lyngby, Denmark
 Graham Cameron, EMBL-EBI
 Dominic Clark, EMBL-EBI
 Antoine de Daruvar, UB2 and representing MENESR, France
 Mark Forster, Syngenta, UK

Alf Game, BBSRC, UK

Tim Hubbard, WT Sanger Institute, UK

Mark Lathrop, Centre National de Genotypage, Evry, France

Hans-Werner Mewes, National Research Centre for Health and Environment (GSF), Neuherberg, Germany

Bengt Persson, Linköping University, Sweden

Silke Schumacher, EMBL, Heidelberg, Germany

Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy

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

Johan van der Lei, Erasmus, the Netherlands

ELIXIR Supercomputing

Sarah Hunter, EMBL-EBI (Chair)

Josep Luis Gelpi, Barcelona Supercomputing Centre, Spain

Marko Myllynen, CSC - Scientific Computing Ltd., Espoo, Finland

Tommi Nyrönen, CSC - Scientific Computing Ltd., Espoo, Finland

Modesto Orozco, Barcelona Supercomputing Centre, Spain

Antony Quinn, EMBL-EBI

David Torrents, Barcelona Supercomputing Centre, Spain

ELIXIR Systems Biology

Nicolas le Novère, EMBL-EBI (Chair)

Nick Juty, EMBL-EBI

Renate Kania, EML Research (SABIO-RK), Germany

Camille Laibe, EMBL-EBI

Lennart Martens, Ghent University, Belgium

Sandra Orchard, EMBL-EBI

Babette Regierer, University of Potsdam - FORSYS, Germany

Susanna-Assunta Sansone, EMBL-EBI

Ugis Sarkans, EMBL-EBI

Esther Schmidt, German Cancer Research Centre (DKFZ), Germany

Ida Schomburg, Technische Universität Braunschweig, Germany

Erik Sonnhammer, Stockholm University, Sweden

Christopher Southan, Chris DS Consulting Ltd., UK

Neil Swainston, University of Manchester, UK

Andrei Zinovyev, Institut Curie, Paris, France

ELIXIR Tools

Søren Brunak, Technical University of Denmark Lyngby, Denmark (Chair)

Henning Hermjakob, EMBL-EBI (co-Chair)

Christophe Blanchet, Institute for Protein Biology and Chemistry, France

Jan Christian Bryne, Bergen Centre for Computational Science, Norway

Ib Groth Clausen, AstraZeneca Research and Development, Sweden

Joaquin Dopazo, Centro de Investigacion Principe Felipe, Valencia, Spain

Adam Godzik, Sanford-Burnham Medical Research Institute, California, USA

Paul Gordon, University of Calgary, Canada

Rodrigo Gouveia-Oliveira, Technical University of Denmark Lyngby, Denmark

Rodrigo Lopez, EMBL-EBI

Francis Ouellette, Ontario Institute for Cancer Research, Canada

Niels Tolstrup, Exiqon, Denmark

ELIXIR Training

Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy (Chair)

Cath Brooksbank, EMBL-EBI (co-Chair)

Terri Attwood, University of Manchester, UK

Thomas Blicher, Technical University of Denmark Lyngby, Denmark

Janusz Bujnicki, International Institute of Molecular and Cell Biology, Warsaw, Poland

Pedro Fernandes, Instituto Gulbenkian de Ciência, Oerias, Portugal

Matthias Haury, EMBL, Heidelberg, Germany

Heta Kero, CSC - Scientific Computing Ltd., Espoo, Finland
 Hans-Werner Mewes, National Research Centre for Health and Environment (GSF), Neuherberg, Germany
 Tommi Nyronen, CSC - Scientific Computing Ltd., Espoo, Finland
 Vicky Schneider, EMBL-EBI
 Kari Tuononen, University of Helsinki, Finland
 Jaak Vilo, University of Tartu, Estonia

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, Université 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, WT Sanger Institute, UK
 Christopher Southan, Chris DS Consulting Ltd., UK (Observer)
 John Walshaw, BBSRC John Innes Centre in Norwich, UK
 Mathew Woodward, Medimmune (a subsidiary of AstraZeneca), USA

EMBRACE Scientific Advisory Board

Rita Casadio, University of Bologna, Italy
 Kay Hoffman, Miltenyi Biotec, Germany
 Mathew Woodward, 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

International Nucleotide Sequence Database Collaboration (INSDC) Scientific Advisory Board

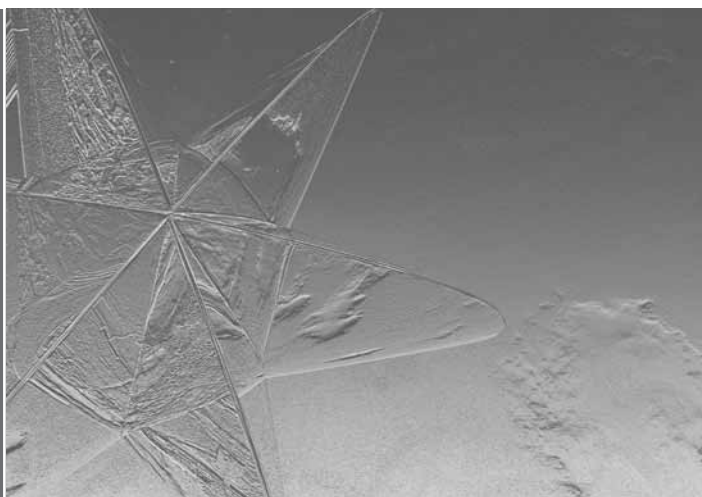
Antoine Danchin, CNRS, Institut Pasteur, Paris, France
 Babis Savakis, University of Crete and IMBB-FORTH, Heraklion, Greece
 Jean Weissenbach, Génoscope, Evry, France

SYBARIS Scientific Advisory Board

Jane Kaye, University of Oxford, UK
 Mihai G. Netea, Radboud University Nijmegen Medical Centre (RUNMC), the Netherlands
 Ken Smith, University of Cambridge, UK
 Ioannis Xenarios, Swiss Institute of Bioinformatics, Switzerland

VBO Tool for Bridging Vertebrate Anatomy Ontologies, Biotechnology and Biological Sciences Research Council, Bioinformatics and Biological Resources fund 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



External Seminar Speakers

The EMBL-EBI has hosted a weekly series of seminars in systems biology 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. These seminars, organised by EIPOD Fellow Mikhail Spivakov and Nicolas Le Novère, has focussed on system-wide approaches to biology and addressed both computational and wet-lab issues. The list below shows speakers who have kindly presented for this particular series; however, EMBL-EBI has also hosted a very large number of internal and external ad-hoc seminars that are not listed here.

EXTERNAL SEMINAR SPEAKERS

Date	Speaker	Title
14 July 2009	Felix Naef	Studying circadian clocks using comparative genomics and modelling
28 July 2009	David Thybert	Detecting functional potentialities in prokaryotic genomes: application to the ROS/RNS detoxification sub-system
31 July 2009	Sheldon McKay	Challenges in comparative genome browsing
4 September 2009	Jackie Han	Inferring molecular networks
7 September 2009	Ramit Mehr	The immunomics of lymphocyte repertoires
22 September 2009	Balazs Papp	Systems biology of robustness and genetic interactions in yeast
29 September 2009	Mikkel Schierup	Population genomics of ancestral species using hidden Markov models
13 October 2009	Andrew Firth	Finding the 'hidden' genes: lessons from viruses
20 October 2009	Uwe Sauer	Constructing directed protein interaction networks for activated EGF/Erk signalling
28 October 2009	Denis Thieffry	Logical modelling of cell fate specification



Date	Speaker	Title
3 November 2009	Victor Kunin	Accurate estimation of microbial community using pyrotags AND The Open Journal: social network, journal club and peer-reviewed journal with automated editors and production
10 November 2009	Moritz Kreysing	Mammalian photoreceptor nuclei adapt to vision
24 November 2009	Michael Watson	Evolution of mammalian transcriptional control
15 December 2009	Tracey Bray	From structure to function in enzymes
26 January 2010	Jernej Ule	iCLIP RNA maps elucidate TIA1 and TIAL1 as master regulators of RNA splicing
2 February 2010	Aoife McLysaght	Recent de novo origin of human protein-coding genes
19 February 2010	Sydney Brenner	Reading the human genome
2 March 2010	Lamia Zaghloul	Relating domains of strand composition asymmetry to the large-scale organisation of the human genome
13 April 2010	Mark Isalan	Shuffled gene networks and liar paradoxes
20 April 2010	Yves Lussier	Quantized phenotypes emerge as multiscale mechanisms underpinning complex traits from mining genetic narratives, ontologies, and RNA sequences
27 April 2010	Manuel Irimia	Origin and evolution of NOVA splicing networks in vertebrates and other metazoans
18 May 2010	Johan Elf	Probing intracellular kinetics at the level of single molecules: transcription factor search kinetics and stringent response in <i>E. coli</i>
25 May 2010	Jürgen Wehland	Perturbation of the host cytoskeleton by bacterial pathogens
1 June 2010	Marius Ueffing	Targeting protein networks on a quantitative scale: affinity-based approaches to studying molecular interactions
8 June 2010	Victor Sourjik	Robustness of signalling in bacterial chemotaxis
15 June 2010	Rolf Backofen	Non-coding RNAs: how to find them and how to find targets
22 June 2010	Joerg Stelling	Systems analysis of cellular regulation under uncertainty



Publications

SERVICES

APWEILER TEAM, JULY – DECEMBER 2009

Apweiler, R. (2009) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38 (Database issue), D142-D148.

Apweiler, R., et al. (2009) Approaching clinical proteomics: current state and future fields of application in cellular proteomics. *Cytometry Part A* 75, 816-832.

Aranda, B., et al. (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38 (Database issue), D525-D531.

Binns, D.E. et al. (2009) QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* 25, 3045-3046.

Eisenacher, M., et al. (2009) Proteomics Data Collection - 5th ProDaC Workshop 4 March 2009, Kolymari, Crete, Greece. *Proteomics* 9, 3626-3629.

Eisenacher, M., et al. (2009) Getting a grip on proteomics data - Proteomics Data Collection (ProDaC). *Proteomics* 9, 3928-3933.

Furnham, N., et al. (2009) Missing in action: enzyme functional annotations in biological databases. *Nature Chemical Biology* 5, 521-525.

Ilsey, G.R., Luscombe, N.M. and Apweiler, R. (2009) Know your limits: assumptions, constraints and interpretation in systems biology. *Biochim. Biophys. Acta* 1794, 1280-1287.

Schlüter, H., et al. Finding one's way in proteomics: a protein species nomenclature. *Chem. Cent. J.* 3, 11.

Toronto International Data Release Workshop, et al. (2009) Prepublication Data Sharing. *Nature* 461, 168-170.

The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, Supplement 1 (2009), D169-D174.

APWEILER TEAM, JANUARY – JUNE 2010

Alam-Faruque, Y., et al. (2010) The renal gene ontology annotation initiative. *Organogenesis* 6, 71-75.

Alterovitz, G., et al. (2010) Ontology Engineering. *Nat. Biotech.* 28, 128-130.

Bruford, E.A. (2010) Highlights of the 'gene nomenclature across species' meeting. *Hum. Genomics* 4, 213-217.

Cochrane G. et al. (2010) Public data resources as a foundation for a worldwide metagenomics data infrastructure. In: *Metagenomics. Theory, Methods and Applications*, Marco, D., Ed. Norfolk, UK: Caister Academic Press, pp 79-105.

Fernández-Suárez, X.M. and Schuster, M.K. (2010) Using the Ensembl genome server to browse genomic sequence data. *Curr. Protoc. Bioinformatics*, June, Chapter 1, Unit1.15.

Freitas, A.A., Wieser, D.C. and Apweiler, R. (2010) On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 7, 172-182.

Hinz, U. and The UniProt Consortium. (2010) From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledge Base. *Cell. Mol. Life Sci.* 67, 1049-1064.

Sedransk, N., et al. (2010) Make research data public? Not always so simple: a dialogue for statisticians and science editors. *Statistical Sci.* 25, 41-50.

Ye, K., et al. (2010) Mining unique-*m* substrings from genomes. *J. Proteomics Bioinform.* 3, 99-100.

BIRNEY TEAM, JULY – DECEMBER 2009

Atanur, S.S., et al. (2009) The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. *Genome Res.* 20, no. 6 (2010): 791-803.

Durinck, S., et al. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184-1191.

Flicek, P., et al. (2009) Ensembl's 10th year. *Nucleic Acids Res.* 38 (Database issue), D557-562.

Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* 6 (Suppl 11), S6-S12.

Flicek, P. and Birney, E. (2009) Visualising the epigenome. In: *Epigenomics*. Greally, J.M. and Martienssen, R.C. and Ferguson-Smith, A.C., Eds. Dordrecht: Springer Netherlands, pp. 55-66.

Gabalón, T., et al. (2009) Joining forces in the quest for orthologs. *Genome Biol.* 10, 403.

Kahlem, P. and Newfeld, S.J. (2009) Informatics approaches to understanding Tgfr² pathway regulation. *Development* 136, 3729-3740.

Kahlem, P., et al. (2009) ENFIN - a European network for integrative systems biology. *C. R. Biol.* 332, 1050-1058.

Kersey, P.J., et al. (2009) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.* 38 (Database issue), D563-D569.

Leinonen, R., et al. (2009) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.* 38 (Database issue), D39-D45.

Pruitt, K.D., et al. (2009) The Consensus Coding Sequence (CCDS) Project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* 19, 1316-1323.

Zerbino, D.R., et al. (2009) Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One* 4, e8407.

BIRNEY TEAM, JANUARY – JUNE 2010

Bruford, E.A. (2010) Highlights of the 'gene nomenclature across species' meeting. *Hum. Genomics* 4, 213-217.

Chen, Y., et al. (2009) Ensembl Variation Resources. *BMC Genomics* 11, 293.

Daelemans, C., et al. (2010) High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *BMC Genet.* 19, 25.

Fernández-Suárez, X.M. and Schuster, M.K. (2010) Using the Ensembl genome server to browse genomic sequence data. *Curr. Protoc. Bioinformatics*, June, Chapter 1, Unit1.15.

Green, R. E., et al. (2010) A draft sequence of the Neandertal genome. *Science* 328, 710-722.

Hoffman, M.M. and Birney, E. (2010) An effective model for natural selection in promoters. *Genome Res.* 20, 685-692.

Hudson, T. J., et al. (2010) International network of cancer genome projects. *Nature* 464, 993-998.

Jassal, B., et al. (2010) The systematic annotation of the three main GPCR families in Reactome. Database (Oxford) 2010, baq018.

Jung, M., et al. (2010) A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS One* 5, e10709.

McDaniell, R., et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328, 235-239.

Pleasant, E.D., et al. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463, 184-190.

Reuveni, E., Birney, E. and Gross, C.T. (2010) The consequence of natural selection on genetic variation in the mouse. *Genomics* 95, 196-202.

Rios, D., et al. (2010) A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics* 11, 238.

Spudich, G. and Fernández-Suárez, X.M. (2010) Touring Ensembl: a practical guide to genome browsing. *BMC Genomics* 11, 295.

Toronto International Data Release Workshop. (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome Res.* 20, 249-256.

BRAZMA TEAM, JULY – DECEMBER 2009

Aebersold, R., et al. (2009) Report on EU-USA workshop: How systems biology can advance cancer research (27 October 2008). *Mol. Oncol.* 3, 9-17.

Alianni, S., et al. (2009) The fission yeast homeodomain protein Yox1p binds to MBF and confines MBF-dependent cell-cycle transcription to G1-S via negative feedback. *PLoS Genet.* 5, 1-12.

Field, D., et al. (2009) Omics data sharing. *Science* 326, 234-236.

Gehlenborg, N. and Brazma, A. (2009) Visualization of large microarray experiments with space maps. *BMC Bioinformatics* 10, 7.

Gehlenborg, N., et al. (2009) The Prion Disease Database: A comprehensive transcriptome resource for systems biology research in prion diseases. *Database (Oxford)*. 2009, bap011.

Harttig, U., et al. (2009) Owner controlled data exchange in nutrigenomic collaborations: The NuGO information network. *Genes Nutr.* 4, 113-122.

Kapushesky, M., et al. (2009) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38, D690-D98.

Krestyaninova, M., et al. (2009) A system for information management in biomedical studies - SIMBioMS. *Bioinformatics* 25, 2768-2769.

Rung, J., et al. (2009) Genetic variant near *irs1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat. Genet.* 41, 1110-1115.

BRAZMA TEAM, JANUARY – JUNE 2010

Antonov, A.V., et al. (2010) R spider: A network-based analysis of gene lists by combining signalling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res.* 38 (Suppl.), W78-W83

Brinkman, R., et al. (2010) Modeling biomedical experimental processes with OBI. *J. Biomed. Semantics* 1, S7.

Gehlenborg, N., et al. (2010) Visualisation of omics data for systems biology. *Nat. Methods* 7, S56-S68.

Hoogland, C., et al. (2010) Guidelines for reporting the use of gel image informatics in proteomics. *Nat. Biotechnol.* 28, 655-656.

Hudson, T.J., et al. (2010) International network of cancer genome projects. *Nature* 464, 993-998.

Lukk, M., et al. (2010) A global map of human gene expression. *Nat. Biotechnol.* 28, 322-324.

Ma, L.J., et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *fusarium*. *Nature* 464, 367-373.

Malone, J., et al. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics* 26, 1112-1118.

McKinney, E.F., et al. (2010) A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nat. Med.* 16, 586-591.

O'Donoghue, S.I., et al. (2010) Visualizing biological data - now and in the future. *Nat. Methods* 7, S2-S4.

Peltonen, J., et al. (2010) An information retrieval perspective on visualization of gene expression data with ontological annotation. In: *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2178-2181.

Rocca-Serra, P., et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26, 2354-2356.

Swertz, M.A., et al. (2010) XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biol.* 11, R27.

COCHRANE TEAM, JULY 2009 – JUNE 2010

Cochrane, G.R. and Galperin, M.Y. (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res.* 38 (Database issue), D1-D4.

Leinonen, R., et al. (2009) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.* 38, D39-D45.

Shumway, M., et al. (2009) Archiving next generation sequencing data. *Nucleic Acids Res.* 38 (Database issue), D870-D871.

Cochrane G., Martin M.J. and Apweiler, R. (2010) Public data resources as a foundation for a worldwide metagenomics data infrastructure. In: *Metagenomics. Theory, Methods and Applications*, Marco, D., Ed. Norfolk, UK: Caister Academic Press, pp 79-105.

FLICEK TEAM, JULY – DECEMBER 2009

Chandras, C., et al. (2009) Models for financial sustainability of biological databases and resources. *Database (Oxford)* 2009, bap017.

Fernández-Suárez, X.M. and Birney, E. (2009) Ensembl Genome Browser. In: *Vogel and Motulsky's human genetics: Problems and approaches*. Speicher, M.R., et al., Eds. Heidelberg: Springer, pp. 923.

Flicek, P., et al. (2009) Ensembl's 10th year. *Nucleic Acids Res.* 38 (Database issue), D557-562.

Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* 6 (Suppl 11), S6-S12.

Flicek, P. and Birney, E. (2009) Visualising the epigenome. In: *Epigenomics*. Greally, J.M. and Martienssen, R.C and Ferguson-Smith, A.C., Eds. Dordrecht: Springer Netherlands, pp. 55-66.

Flicek, P. (2010) Journal club. A computational geneticist looks at mechanisms of chromosomal evolution. *Nature* 463, 713.

Gabaldon, T., et al. (2009) Joining forces in the quest for orthologs. *Genome Biol.* 10, 403.

Haider, S., et al. (2009) BioMart Central Portal--unified access to biological data. *Nucleic Acids Res.* 37 (Web server issue), W23-W27.

Kersey, P. J., et al. (2009) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.* 38 (Database issue), D563-D569.

Krestyaninova, M., et al. (2009) A System for Information Management in BioMedical Studies--SIMBioMS. *Bioinformatics* 25, 2768-2769.

Schofield, P.N., et al. (2009) Post-publication sharing of data and tools. *Nature* 461, 171-173.

Xue, Y., et al. (2009) Population differentiation as an indicator of recent positive selection in humans: An empirical evaluation. *Genetics* 183, 1065-1077.

FLICEK TEAM, JANUARY – JUNE 2010

Atanur, S.S., et al. (2010) The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome Res.* 20, 791-803.

Ballester, B., et al. (2010) Consistent annotation of gene expression arrays. *BMC Genomics* 11, 294.

Chen, Y., et al. (2010) Ensembl variation resources. *BMC Genomics* 11, 293.

Daelemans, C., et al. (2010) High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *BMC Genetics* 11, 25.

Dalgleish, R., et al. (2010) Locus reference genomic sequences: An improved basis for describing human DNA variants. *Genome Med.* 2, 24.

Fernández-Suárez, X.M. and Schuster, M.K. (2010) Using the Ensembl genome server to browse genomic sequence data. *Curr. Protoc. Bioinformatics* 2007 Chapter 1, Unit 1.15.

Gheldof, N., et al. (2010) Cell-type-specific long-range looping interactions identify

distant regulatory elements of the CFTR gene. *Nucleic Acids Res.* 38, 4325-36.

Gruenberger, M., et al. (2010) Towards the integration of mouse databases - definition and implementation of solutions to two use-cases in mouse functional genomics. *BMC Res. Notes* 3, 16.

International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature* 464, 993-998.

McDaniell, R., et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328, 235-239.

McLaren, W., et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* 26, 2069-2070.

Peric-Hupkes, D., et al. (2010) Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* 38, 603-613.

Rios, D., et al. (2010) A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics* 11, 238.

Schmidt, D., et al. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* 20, 578-588.

Schmidt, D., et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036-1040.

Severin, J., et al. (2010) eHive: An artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics* 11, 240.

Smedley, D., et al. (2010) Finding and sharing: New approaches to registries of databases and services for the biomedical sciences. *Database (Oxford)* 2010, baq014.

Spudich, G. and Fernández-Suárez, X.M. (2010) Touring Ensembl: A practical guide to genome browsing. *BMC Genomics* 11, 295.

Sudbery, I., et al. (2010) Systematic analysis of off-target effects in an RNAi screen reveals microRNAs affecting sensitivity to TRAIL-induced apoptosis. *BMC Genomics* 11, 175.

Swertz, M.A., et al. (2010) XGAP: A uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biol.* 11, R27.

Warren, W.C., et al. (2010) The genome of a songbird. *Nature* 464, 757-762.

Wilkinson, P., et al. (2010) Emma-mouse mutant resources for the international scientific community. *Nucleic Acids Res.* 38 (Database issue), D570-576.

Ye, K., et al. (2010) Mining unique-*m* substrings from genomes. *J. Proteomics Bioinform.* 3, 99-100.

Zouberakis, M., et al. (2010) Mouse Resource Browser--a database of mouse databases. *Database (Oxford)* 2010, baq010.

HERMJAKOB TEAM, JULY – DECEMBER 2009

Aranda, B., et al. (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38 (Database issue), D525-D531.

Barsnes, H., et al. (2009) OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics* 9, 3772-3774.

Colaert, N., et al. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* 6, 786-787.

Côté, R., et al. (2010) The Ontology Lookup service: bigger and better. *Nucleic Acids Res.* 38 (Suppl.), W155-W160

Eisenacher, M., et al. (2009) Proteomics Data Collection - 5th ProDaC Workshop: 4 March 2009, Kolymbari, Crete, Greece. *Proteomics* 9, 3626-3629.

Eisenacher, M., et al. (2009) Getting a grip on proteomics data - Proteomics Data Collection (ProDaC). *Proteomics* 9, 3928-3933.

Kahlem, P., et al. (2009) ENFIN - a European network for integrative systems biology. *C. R. Biol.* 332, 1050-1058.

Kathiresan, T., et al. (2009) A protein interaction network for the large conductance Ca(2+)-activated K(+) channel in the mouse cochlea. *Mol. Cell. Proteomics* 8, 1972-87.

Kim, Y.H., et al. (2009) Toward a Successful Clinical Neuroproteomics The 11th HUPO Brain Proteome Project Workshop 3 March, 2009, Kolymbari, Greece. *Proteomics Clin. Appl.* 3, 1012-1016.

Mehta, A. and Orchard, S. (2009) Nucleoside diphosphate kinase (NDPK, NM23, AWD): recent regulatory advances in endocytosis, metastasis, psoriasis, insulin release, fetal erythroid lineage and heart failure; translational medicine exemplified. *Mol. Cell. Biochem.* 329, 1-13.

Montecchi-Palazzi, L., et al. (2009) The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* 9, 5112-5119.

Orchard, S. (2009) Ending the "publish and vanish" culture: how the data standardization process will assist in data harvesting. *J. Proteome Res.* 8, 3219.

Orchard, S. (2009) Data deposition as an integral part of the publication process. *J. Proteomics Bioinform.* 2, 334-335.

Orchard, S., et al. (2009) 2nd joint publication - psi workshop 24th april 2009, turku finland. *Proteomics* 9, 4426-4428.

Orchard, S., et al. (2009) Annual spring meeting of the proteomics standards initiative. *Proteomics* 9, 4429-4432.

Reisinger, F. and Martens, L. (2009) Database on Demand - an online tool for the custom generation of FASTA-formatted sequence databases. *Proteomics* 9, 4421-4424.

Salwinski, L., et al. (2009) Recurated protein interaction datasets. *Nat. Methods* 6, 860-861.

Vizcaino, J.A., et al. (2009) The proteomics identifications database: 2010 update. *Nucleic Acids Res.* 38, D736-D742.

Vizcaino, J.A., et al. (2009) A guide to the proteomics identifications database proteomics data repository. *Proteomics* 9, 4276-4283.

HERMJAKOB TEAM, JANUARY – JUNE 2010

Antonov, A.V., et al. (2010) R spider: A network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res.* 38 (Suppl.), W78-W83.

Barsnes, H., et al. (2010) OLS dialog: An open-source front end to the ontology lookup service. *BMC Bioinformatics* 11, 34.

Bourbeillon, J., et al. (2010) Minimum information about a protein affinity reagent (MIAPAR). *Nat. Biotechnol.* 28, 650-653.

Côté, R.G., et al. (2010) jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* 10, 1332-1335.

Gloriam, D.E., et al. (2010) A community standard format for the representation of protein affinity reagents. *Mol. Cell. Proteomics* 9, 1-10.

Helsens, K., et al. (2010) ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics* 10, 1261-1264.

Hunter, P., et al. (2010) A vision and strategy for the virtual physiological human in 2010 and beyond. *Philos. Transact. A. Math. Phys. Eng. Sci.* 368, 2595-614.

Jones, P. and Martens, L. (2010) Using the PRIDE proteomics identifications database for knowledge discovery and data analysis. *Methods Mol. Biol.* 604, 297-307.

Lee, K., et al. (2010) Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *Plant Cell.* 22, 997-1005.

Muth, T., et al. (2010) jTraX: A free, platform independent tool for isobaric tag quantitation at the protein level. *Proteomics* 10, 1223-1225.

Orchard, S., et al. (2010) Implementing data standards: A report on the HUPO-PSI workshop. *Proteomics* 10, 1895-1898.

Orchard, S., et al. (2010) The publication and database deposition of molecular interaction data. *Curr. Protoc. Protein Sci.* Chapter 25, Unit 25.3.

Orchard, S. and Kerrien, S. (2010) Molecular interactions and data standardisation. *Methods Mol Biol.* 604, 309-318.

Perreau, V.M., et al. (2010) A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease. *Proteomics* 10, 2377-2395.

Vizcaino, J.A., et al. (2010) PRIDE: Data submission and analysis. *Curr. Protoc. Protein Sci.* Chapter 25, Unit 25.4.

HUNTER TEAM, JULY 2009 – JUNE 2010

Jones, P. and Martens, L. (2010) Using the PRIDE proteomics identifications database for knowledge discovery and data analysis. *Methods Mol. Biol.* 604, 297-307.

Schneider, M.V., et al. (2010) Bioinformatics training: a review of challenges, actions and support requirements. *Brief. Bioinform.* 11, 544-551.

KAPUSHESKY TEAM, JULY 2009 – JUNE 2010

Goncalves, A., et al. (2010) ArrayExpressHTS: distributed computing for RNA-seq data processing and quality assessment. *Bioinformatics* (submitted).

Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Database issue), D690-698.

Lukk, M., et al. (2010) A global map of human gene expression. *Nat. Biotechnol.* 28, 322-324.

Malone, J., et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26, 1112-1118.

KERSEY TEAM, JULY 2009 – JUNE 2010

Flicek, P., et al. (2009) Ensembl's 10th year. *Nucleic Acids Res.* 38 (Database issue), D557-562.

Kersey, P.J., et al. (2009) Ensembl genomes: Extending ensembl across the taxonomic space. *Nucleic Acids Res.* 38 (Database issue), D563-D569.

Kirkness, E.F., et al. (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Nat. Acad. Sci. USA* 107, 12168-12173.

KLEYWEGT TEAM, JULY 2009 – JUNE 2010

Berman, H.M., et al. (2010) Safeguarding the integrity of protein archive. *Nature* 463, 425.

De Simone, A., et al. (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *J. Am. Chem. Soc.* 131, 16332-16333.

Doreleijers, J.F., et al. (2009) The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. *J. Biomol. NMR* 45, 389-396.

Dougherty, M.T., et al. (2009) Unifying biological image formats with HDF5. *Communications of the Association for Computing Machinery (ACM)* 52, 42-47.

Fogh, R.H., et al. (2010) MEMOPS: Data modelling and automatic code generation. *J. Integr. Bioinform.* 7, 123.

Gorbalenya, A.E., et al. (2010) Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antiviral Res.* 87, 95-110.

Krissinel, E. (2010) Crystal contacts as nature's docking solutions. *J. Comput. Chem.* 31, 133-143.

Rosato, A., et al. (2009) CASD-NMR: Critical assessment of automated structure determination by NMR. *Nat. Methods* 6, 625-626.

Velankar, S., et al. (2010) PDBE: Protein Data Bank in Europe. *Nucleic Acids Res.* 38 (Database issue), D308-D317.

LOMAX TEAM, JULY 2009 – JUNE 2010

Alterovitz, G., et al. (2010) Ontology engineering. *Nat. Biotechnol.* 28, 128-130.

Consortium, G.O. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38, D331-D335.

Giglio, M.G., et al. (2009) Applying the Gene Ontology in microbial annotation. *Trends Microbiol.* 17, 262-268.

Mungall, C.J., et al. (2010) Cross-product extensions of the Gene Ontology. *J. Biomed. Inform.* (in press). Published online 16 February; DOI: 10.1016/j.jbi.2010.02.002.

MARTIN TEAM, JULY 2009 – JUNE 2010

The UniProt Consortium. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38 (Database issue), D142-D148.

Jain, E., et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10, 136.

The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38 (Database issue), 331-D335.

Cochrane, G., Martin, M.J. and Apweiler, R. (2010) Public data resources as a foundation for a worldwide metagenomics data infrastructure. In: *Metagenomics: theory, methods, and applications*. Chapter 5. Marco, D., Ed. Norwich, UK: Caister Academic Press, 212 pp.

O'DONOVAN TEAM, JULY 2009 – JUNE 2010

The UniProt Consortium. (2010) .The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38 (Database issue), D1142-D1148.

Binns, D., et al. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 25, 3045-3046.

Alam-Faruque, Y., et al. (2010). The Renal Gene Ontology Annotation Initiative. *Organogenesis* 6, 71-75.

The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38 (Database issue), D331-D335.

OVERINGTON TEAM, JULY 2009 – JUNE 2010

Berriman, M., et al. (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature* 460, 352-358.

Gaulton, A. and Overington, J.P. (2010) Role of open chemical data in aiding drug discovery and design. *Future Medicinal Chemistry* 2, 903-907.

Harland, L. and Gaulton, A. (2009) Drug target central. *Expert Opin. Drug Discov.* 4, 857-872.

PARKINSON TEAM, JULY 2009 – JUNE 2010

Brinkman, R., et al. (2010) Modeling biomedical experimental processes with OBI. *J. Biomed. Semantics* 1, S7.

Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Database issue), D690-D698.

Kauffmann, A., et al. (2009) Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* 25, 2092-2094.

Lukk, M., et al. (2010) A global map of human gene expression. *Nat. Biotechnol.* 28, 322-324.

Malone, J., et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26, 1112-1118.

Parkinson, H., et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the Atlas of Gene Expression. *Nucleic Acids Res.* 37 (Database issue), D868-D872.

Shankar, R., et al. (2010) Annotare - a tool for annotating high-throughput biomedical investigations and resulting data. *Bioinformatics* 26, 2470-2471.

Swertz, M.A., et al. (2010) XGAP: A uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biol.* 11, R27.

RICE TEAM, JULY 2009 – JUNE 2010

Cock, P. et al. (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767-1771.

Pettifer, S., et al. (2010) The EMBRACE Web Service Collection. *Nucleic Acids Res.* 38 (Suppl. 2), W683-W688.

SARKANS TEAM, JULY 2009 – JUNE 2010

Malone, J., et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26, 1112-1118.

Kapushesky, M., et al. (2010) Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* 38 (Database issue), D690-D698.

RESEARCH

BERTONE GROUP, JULY 2009 – JUNE 2010

Dvinge, H. and Bertone, P. (2009) HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R.. *Bioinformatics* 25, 3325-3326.

Fredman, D., Engstrom, P.G. and Lenhard, B. (2009) Web-based tools and approaches to study long-range gene regulation in Metazoa. *Brief. Funct. Genomic. Proteomic.* 8, 231-242.

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.

Loos, R., Manea, F. and Mitran, V. (2010) Small universal accepting hybrid networks of evolutionary processors. *Acta Informatica* 47, 133-146.

Ragvin, A., et al. (2010) Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Nat. Acad. Sci. U.S.A.* 107, 775-780.

Salmon-Divon, M., et al. (2010) PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* 11, 415.

ENRIGHT GROUP, JULY 2009 – JUNE 2010

- Fabani, M.M., et al. (2010) Efficient inhibition of miR-155 function in vivo by peptide nucleic acids. *Nucleic Acids Res.* 38, 4466-4675.
- Guerra-Assunção, J.A., and Enright, J.A. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* 11, 133.
- Manakov, S.A., Grant, S.G.N. and Enright, A.J. (2009) Reciprocal regulation of microRNA and mRNA profiles in neuronal development and synapse formation. *BMC Genomics* 10, 419.
- Palmer, R. D., et al. (2010) Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets. *Cancer Res.* 70, 2911-2923.
- Piipari, M., T.A. Down, H. Saini, A. Enright, and T.J.P. Hubbard. iMotifs: an integrated sequence motif visualization and analysis environment. *Bioinformatics* 26, 843-844.
- Sudbery, I., et al. (2010) Systematic analysis of off-target effects in an RNAi screen reveals microRNAs affecting sensitivity to TRAIL-induced apoptosis. *BMC Genomics* 11, 175.

GOLDMAN GROUP, JULY 2009 – JUNE 2010

- Alon, N., et al. (2010) Approximate maximum parsimony and ancestral maximum likelihood. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 7, 183-187.
- Chor, B., et al. (2009) Genomic DNA K-Mer spectra: models and modalities. *Genome Biol.* 10, R108-R108.
- Flicek, P., et al. (2009) Ensembl's 10th year. *Nucleic Acids Res.* 38 (Database issue), D557-D562.
- Pardi, F., Guillemot, S. and Gascuel, O. (2010) Robustness of phylogenetic inference based on minimum evolution. *Bull. Math. Biol.* 72, 1820-1839.
- San Mauro, D., et al. (2009) Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear rag1. *Syst. Biol.* 58, 425-438.
- Talavera, D., Taylor, M.S. and Thornton, J.M. (2010) The (non)malignancy of cancerous amino acid substitutions. *Proteins* 78, 518-529.
- Washietl, S. (2010) Sequence and structure analysis of noncoding RNAs. *Methods Mol. Biol.* 609, 285-306.
- Washietl, S. and Hofacker, I.L. (2010) Nucleic acid sequence and structure databases. *Methods Mol. Biol.* 609, 3-15.
- Yang, Z., Nielsen, R. and Goldman, N. (2009) In defense of statistical methods for detecting positive selection. *Proc. Nat. Acad. Sci. USA* 106, E95.

LE NOVÈRE GROUP, JULY 2009 – JUNE 2010

- Baldi, B.F., et al. (2010) Schizophrenic: forever young? *Genome Med.* 2, 32.
- Chelliah, V., et al. (2009) Data integration and semantic enrichment of systems biology models and simulations. *Lecture Notes in Computer Science* 5647, 5-15.
- Edelstein, S.J. and Changeux, J.P. (2010) Relationships between structural dynamics and functional kinetics in oligomeric membrane receptors. *Biophys. J.* 98, 2045-2052.
- Edelstein, S.J., et al. (2010) Ligand depletion in vivo modulates the dynamic range and cooperativity of signal transduction. *PLoS ONE* 5, e8449.
- Le Novère, N., et al. (2009) The systems biology graphical notation. *Nat. Biotechnol.* 27, 735-741.
- Li, C., et al. (2010) BioModels.net Web Services, a free and integrated toolkit for computational modelling software. *Brief. Bioinform.* 11, 270-277.
- Li, C., et al. (2010) BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* 4, 92.
- Stefan, M.I., et al. (2009) Computing phenomenologic Adair-Klotz constants from microscopic MWC parameters. *BMC Syst. Biol.* 3, 68.
- Tolle, D.P. and Le Novère, N. (2010) Brownian diffusion of AMPA receptors is sufficient to explain fast onset of LTP. *BMC Syst. Biol.* 4, 25.
- Tolle, D.P. and Le Novère, N. (2010) Meredys, a multi-compartment reaction-diffusion simulator using multistate realistic molecular complexes. *BMC Syst. Biol.* 4, 24.

LUSCOMBE GROUP, JULY 2009 – JUNE 2010

- Ilsey, G.R., et al. (2009) Know your limits: assumptions, constraints and interpretation in systems biology. *Biochim. Biophys. Acta* 1794, 1280-1287.
- Jolma, A., et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 20, 861-873.
- Konig, J., et al. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* 17, 909-915.
- Kotte, O., et al. (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol. Syst. Biol.* 6, 355.
- Raja, S.J., et al. (2010) The nonspecific lethal complex is a transcriptional regulator in *Drosophila*. *Mol. Cell* 38, 827-841.
- Reimand, J., et al. (2010) Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res.* 38, 4768-4777.
- Sepulcre, J., et al. (2010) The organization of local and distant functional connectivity in the human brain. *PLoS Comput. Biol.* 6, e1000808.
- Seshasayee, A.S., et al. (2010) Comparative genomics of cyclic-di-GMP signalling in bacteria: post-translational regulation and catalytic activity. *Nucleic Acids Res.* 38, 5970-5981.
- Vaquerizas, J.M., et al. (2010) Nuclear pore proteins nup153 and megator define transcriptionally active regions in the *Drosophila* genome. *PLoS Genet.* 6, e1000846.

REBHOLZ-SCHUHMANN GROUP, JULY – DECEMBER 2009

- Baker, C.J.O. and Rebholz-Schuhmann, D. (2009) Between proteins and phenotypes: Annotation and interpretation of mutations. *BMC Bioinformatics* 8 (Suppl.), 11.
- Grego, T., et al. (2009) Identification of chemical entities in patent documents. *Lecture Notes Comp. Sci.* 5518, 942-949.
- Jimeno-Yepes, A., et al. (2009) Terminological cleansing for improved information retrieval based on ontological terms. In: *ESAIR '09 Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM Digital Library. DOI: 10.1145/1506250.1506253.
- Jimeno-Yepes, A., et al. (2009) Exploitation of ontological resources for scientific literature analysis: searching genes and related diseases. *Conf Proc. IEEE Eng. Med. Biol. Soc.* 2009, 7073-7078.
- Jimeno-Yepes, A., et al. (2009) Ontology refinement for improved information retrieval. *Information Processing and Management* 46, 426-435.
- Jimeno-Yepes, A., et al. (2009) Reuse of terminological resources for efficient ontological engineering in Life Sciences. *BMC Bioinformatics* 10 (Suppl. 10), S4.
- Nagel, K., et al. (2009) Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb. *BMC Bioinformatics* 10 (Suppl. 8), S4.
- Pezik, P., et al. (2009) Using biomedical terminological resources for information retrieval. In: *Information Retrieval in Biomedicine*. Prince, V. and Roche, M., Eds. Hershey, USA: IGI Global Publishing, pp. 58-77.
- Rinaldi, F. and Rebholz-Schuhmann, D. (2009) Support tools for literature-based information access in molecular biology. In: *IUCS '09 Proceedings of the 3rd International Universal Communication Symposium*. ACM Digital Library. DOI: 10.1145/1667780.1667836.
- Waagmeester, A., et al. (2009) Pathway enrichment based on text mining and its validation on carotenoid and vitamin A metabolism. *OMICS* 13, 367-379.
- REBHOLZ-SCHUHMANN GROUP, JANUARY – JUNE 2010**
- Hoehndorf, R., Oellrich, A. and Rebholz-Schuhmann, D. (2010) Interoperability between phenotype and anatomy ontologies. *Bioinformatics* (in press). DOI: 10.1093/bioinformatics/btq578.
- Rebholz-Schuhmann, D., et al. (2010) Wrestling with biomedical research results: language resources and literature analysis. *J. Bioinform. Comput. Biol.* 8, 129-30.
- Rebholz-Schuhmann, D., et al. (2010) Measuring prediction capacity of individual verbs for the identification of protein interactions. *J. Biomed. Inform.* 43, 200-207.
- Rebholz-Schuhmann, D., et al. (2010) Papermaker: validation of biomedical scientific publications. *Bioinformatics* 26, 982-984.
- Rebholz-Schuhmann, D. and Nenadic, G. (2010) Biomedical semantics: the Hub for Biomedical Research 2.0. *J. Biomed. Semantics* 1, 1.

Rebholz-Schuhmann, D., et al. (2010) CALBC silver standard corpus. *J. Bioinform. Comput. Biol.* 8, 163-79.

THORNTON GROUP, JULY – DECEMBER 2009

Andreini, C., et al. (2009) Metal-MACiE: A database of metals involved in biological catalysis. *Bioinformatics* 25, 2088-2089.

Brooksbank, C., et al. (2009) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* 38 (Database issue), D17-D25.

Capra, J.A., et al. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* 5, e1000585.

Cuff, A., et al. (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure* 17, 1051-1062.

Furnham, N., et al. (2009) Missing in action: enzyme functional annotations in biological databases. *Nat. Chem. Biol.* 5, 521-525.

Herrera, J.L., et al. (2009) Toll-like receptor stimulation differentially regulates vasoactive intestinal peptide type 2 receptor in macrophages. *J. Cell. Mol. Med.* 13, 3209-3217.

Holliday, G.L., et al. (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.* 390, 560-577.

Kahraman, A., et al. (2009) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins* 78, 1120-1136.

Laskowski, R. (2009) Integrated servers for structure-informed function prediction. In: *From protein structure to function with bioinformatics*. Rigden, D.J., Ed. New York/Heidelberg: Springer, pp. 251-272.

Laskowski, R. (2009) Inferring protein function from structure. *Structural bioinformatics, 2nd Edition*. Gu, J. and Bourne, P.E., Eds. Hoboken: Wiley-Blackwell, pp. 341-375.

Laskowski, R. (2009) Protein structure databases. In: *Data mining techniques for the life sciences. (Methods in molecular biology, Vol. 609)* Carugo, O. and Eisenhaber, F., Eds. Clifton, USA: Springer-Humana Press, pp. 59-82.

Laskowski, R.A., et al. (2009) The fine details of evolution. *Biochem. Soc. Trans.* 37, 723-726.

Macchiarulo, A., et al. (2009) Mapping human metabolic pathways in the small molecule chemical space. *J. Chem. Inf. Model.* 49, 2272-2289.

Oliva, R., et al. (2009) Porelogo: A new tool to analyse, visualize and compare channels in transmembrane proteins. *Bioinformatics* 25, 3183-3184.

Pellegrini-Calace, M., et al. (2009) Porewalker: A novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput. Biol.* 5, e1000440.

Rahman, S.A., et al. (2009) Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.* 1, 12.

Selman, C., et al. (2009) Ribosomal protein s6 kinase 1 signaling regulates mammalian life span. *Science* 326, 140-144.

Smith, L.J., et al. (2010) Heme proteins-diversity in structural characteristics, function, and folding. *Proteins* 78, 2349-2368.

Soranzo, N., et al. (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the haemgen consortium. *Nat. Genet.* 41, 1182-1192.

Talavera, D., et al. (2009) Alternative splicing of transcription factors' genes: beyond the increase of proteome diversity. *Comp. Funct. Genomics* 2009, 905894.

Toronto International Data Release Workshop. (2009) Prepublication data sharing. *Nature* 461, 168-170.

Wieser, D. and Niranjan, M. (2009) Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. *In Silico Biol.* 9, 89-103.

THORNTON GROUP, JANUARY – JUNE 2010

Bashton, M. and Thornton, J.M. (2010) Domain-ligand mapping for enzymes. *J. Mol. Recognit.* 23, 194-208.

Chan, A.W.E., et al. (2010) Chemical fragments that hydrogen bond to Asp, Glu, Arg, and His side chains in protein binding sites. *J. Med. Chem.* 53, 3086-3094.

Grönke, S., et al. (2010) Molecular evolution and functional characterization of *Drosophila* insulin-like peptides. *PLoS Genet.* 6, e1000857.

O'Donoghue, S.I., et al. (2010) Visualization of macromolecular structures. *Nat. Methods* 7, S42-S55.

Oliva, R., et al. (2010) Electrostatics of aquaporin and aquaglyceroporin channels correlates with their transport selectivity. *Proc. Natl Acad. Sci. U.S.A.* 107, 4135-4140.

Slack, C., et al. (2010) Regulation of lifespan, metabolism, and stress responses by the *Drosophila* SH2B protein, Lnk. *PLoS Genet.* 6, e1000881.

Talavera, D., et al. (2010) The (non)malignancy of cancerous amino acid substitutions. *Proteins* 78, 518-529.

Tamuri, A.U. and Laskowski, R.A. (2010) ArchSchema: A tool for interactive graphing of related Pfam domain architectures. *Bioinformatics* 26, 1260-1261.

SUPPORT

BROOKSBANK TEAM, JULY 2009 – JUNE 2010

Brooksbank, C., Cameron, G. and Thornton, J. (2010) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* 38, D17-D25.

Schneider, M.V., et al. (2010) Bioinformatics training: a review of challenges, actions and support requirements. *Brief. Bioinform.* (in press). DOI: 10.1093/bib/bbq021.

Wright, V.A., et al. (2010) Bioinformatics training: selecting an appropriate learning content management system – an example from the European Bioinformatics Institute. *Brief. Bioinform.* (in press). DOI: 10.1093/bib/bbq023.

LOPEZ TEAM, JULY 2009 – JUNE 2010

Bhagat, J., et al. (2010) Biocatologue: A universal catalogue of web services for the life sciences. *Nucleic Acids Res.* 38, W689-W694.

Goujon, M., et al. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* W695-W699.

Leinonen, R., et al. (2009) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.* 38, D39-D45.

Li, W., et al. (2009) ExprAlign--the identification of ESTs in non-model species by alignment of cDNA microarray expression profiles. *BMC Genomics* 10, 560.

Li, W., et al. (2009) Non-redundant patent sequence databases with value-added annotations at two levels. *Nucleic Acids Res.* 38, D52-D56.

Robinson, J., et al. (2010) IPD--the Immuno Polymorphism Database. *Nucleic Acids Res.* 38, D863-D869.



Index

- A**
- ageing 74, 75
 - ArchSchema 75, 119
 - ArrayExpress 5, 20, 21, 24, 25, 32, 47, 48, 49, 52, 53, 72, 88, 95, 100, 104, 117
- B**
- bacteria 2, 34, 35, 64, 70, 71, 76, 118
 - BBSRC 2, 9, 50, 55, 89, 105, 108, 109, 110, 111
 - BioModels 68, 69, 105, 118
 - BioSample Database 5, 20, 21, 22
- C**
- CALBC 8, 10, 72, 73, 111, 119
 - cancer 4, 27, 52, 62, 63, 115, 116
 - ChEBI 6, 14, 16, 17, 39, 54, 55, 100, 105
 - ChEMBL 5, 6, 14, 17, 46, 47, 57, 100, 105
 - ChIP-seq 10, 18, 62, 63
 - CiteXplore 42, 43
 - CluSTr 30, 31
 - Collaborations 66, 97
 - comparative genomics 26, 34, 66, 112
 - conservation 7, 27, 65, 70, 119
- D**
- DAS 14, 17, 28, 29, 31, 40, 41
 - Data Centre 2, 9, 86, 87, 98
 - DDMoRe 10
 - DGVa 14, 26
 - drug-discovery 46, 85
 - DRUPAL 88
- E**
- EIPOD 17, 18, 47, 112
 - ELIXIR 3, 8, 9, 10, 82, 83, 84, 85, 89, 98, 107, 108, 109, 110
 - EMBOSS 50, 51, 111
 - EMBRACE 10, 50, 51, 89, 111, 117
 - EMDB 36, 37, 100, 105
 - Ensembl 2, 4, 5, 14, 15, 16, 17, 18, 24, 25, 26, 27, 30, 33, 34, 35, 40, 41, 44, 47, 50, 57, 65, 73, 94, 95, 101, 105, 114, 115, 116, 117, 118
 - Ensembl Genomes 2, 5, 14, 15, 16, 18, 24, 34, 35, 94, 95, 101, 105, 114, 115
 - entity recognition 72
 - enzyme 14, 31, 54, 55, 70, 74, 75, 114, 119
 - European Commission 3, 9, 75
 - EU-OPENSREEN 47
 - ELIXIR 3, 8, 9, 10, 82, 83, 84, 85, 89, 98, 107, 108, 109, 110
 - EMBOSS 50, 51, 111
 - EMBRACE 10, 50, 51, 89, 111, 117
 - European Research Area 8
 - FP7 9, 28, 33
 - GeN2PhEN 48
 - SLING 28, 82, 83, 89, 95
 - SESL 72, 73, 84, 85
 - EU-funded projects 3, 9, 26, 28, 35, 47, 82, 83, 85, 94,
 - European genome–phenome Archive 21
 - European Nucleotide Archive (ENA) 5, 14, 15, 16, 21, 24, 25, 30, 32, 33, 42, 48, 52, 94, 95, 105, 106, 114, 115, 119
 - evolution 2, 7, 8, 10, 27, 65, 66, 67, 70, 74, 75, 116, 117, 119
 - Experimental Factor Ontology 21, 32, 33, 48, 53, 117
- F**
- FP7 9, 28, 33
 - Functional genomics 2, 5, 20, 21, 22, 23, 32, 33, 48, 49, 52, 53, 57, 74, 97, 98
 - funding 97, 98
- G**
- GeN2PhEN 48
 - Gene Expression Atlas 2, 5, 20, 21, 22, 32, 33, 48, 49, 52, 53, 73, 115, 117
 - genome-wide association 22, 48
 - global server load balancing 87
 - GO 6, 16, 30, 38, 39, 41, 43, 44, 45, 101
 - GOA 16, 30, 40, 41, 44, 45
 - grid technology 50
 - growth of core data resources 95
- H**
- HUGO 14, 15
 - human mutation 74
- I**
- Industry 2, 4, 9, 84, 85, 95, 108
 - infrastructure 2, 3, 4, 5, 8, 9, 10, 20, 21, 25, 26, 32, 34, 40, 41, 46, 47, 48, 52, 53, 55, 69, 72, 73, 82, 84, 85, 86, 87, 88, 89, 98, 117
 - Innovative Medicines Initiative (IMI) 9, 10, 47, 82, 84, 85
 - IntAct 7, 16, 17, 28, 29, 101, 114, 116
 - InterPro 14, 16, 30, 31, 40, 42, 45, 95, 102, 106

- intuitive 24, 31, 36, 37, 41
- J**
- Journal of Biomedical Semantics 72
- L**
- literature 6, 8, 16, 26, 28, 29, 37, 42, 46, 50, 54, 72, 73, 74, 76, 85, 116, 118
- M**
- MAGE-TAB 48, 52
- map of gene expression 7
- MetaboLights 54, 55
- metagenomics 30, 41, 114, 115, 117
- microarray 5, 7, 10, 32, 52, 63, 77, 94, 115, 117, 119
- microme 34, 35
- MicroRNAs 7
- MIRIAM 68, 69
- miRNAs 7, 64, 65
- MySQL 86, 87, 88, 89
- N**
- nationalities 97
- next-generation sequencing 4, 7, 10, 22, 63, 66, 117
- non-coding RNAs 62, 64
- O**
- ontology 6, 10, 16, 20, 32, 33, 38, 39, 40, 41, 45, 48, 49, 50, 51, 52, 53, 54, 72, 73, 85, 114, 115, 116
- open source 5, 6, 14, 20, 28, 32, 50, 54, 62, 83, 85, 116
- Oracle 6, 55, 86, 87, 88
- OrChem 6, 14, 54
- organisation chart 96
- P**
- patents 42
- PDBe 2, 6, 36, 37, 42, 47, 58, 88, 95, 100, 102, 106, 117
- personnel 7, 56, 57, 58, 82, 90, 96, 97
- pharmaceutical 8, 9, 32, 47, 72, 85
- PhD student 17, 22, 76, 77
- phenotype 18, 20, 26, 72, 73, 113, 118
- phylogeny 8, 10, 66, 67
- Pistoia Alliance 73, 84, 85
- pore 7, 10, 70, 71, 74, 118
- PRANK 8, 66
- pre-competitive research 2, 84, 85
- press 7, 8, 10, 29, 39, 46, 65, 73, 74, 75, 82, 83, 117, 118, 119
- PRIDE 5, 16, 17, 28, 29, 40, 55, 102, 116
- proteomics 14, 16, 17, 28, 29, 114, 116
- PubMed 6, 41, 42, 43
- R**
- R Cloud 5, 20, 32, 33
- RefSeq 40, 41, 44
- roadshows 2, 31, 36, 37, 82, 83, 95
- S**
- SBGN 8, 68
- Science and Society 76, 82
- search 5, 6, 14, 24, 25, 31, 33, 37, 40, 41, 42, 43, 48, 52, 54, 55, 68, 88, 113, 116
- SESL 72, 73, 84, 85
- signal transduction 68, 69, 118
- SLING 28, 82, 83, 89, 95
- small molecules 6, 17, 46, 47, 54, 55, 70, 71, 74, 76
- SMEs 84, 85
- speakers 112
- stand-alone software 5, 20, 32, 33, 64, 66
- standards 2, 7, 10, 26, 28, 29, 41, 44, 47, 50, 55, 68, 85, 115, 116
- stem cell 62, 63
- storage 2, 35, 68, 86, 87, 98
- Sylamer 64
- T**
- text mining 42, 55
- 1000 Genomes Project 2, 5, 14, 26, 27, 89
- training 2, 4, 9, 10, 14, 15, 16, 20, 22, 24, 31, 36, 37, 46, 50, 54, 73, 76, 82, 83, 84, 85, 88, 89, 95, 96, 116, 119
- transcription 7, 10, 17, 22, 27, 32, 38, 39, 62, 66, 67, 70, 71, 76, 77, 113, 115, 116, 118, 119
- U**
- UK PubMed Central 6, 43
- UniParc 30, 31, 40, 41, 44, 95
- UniProt 7, 14, 16, 18, 24, 29, 30, 31, 40, 41, 42, 44, 45, 47, 50, 72, 75, 88, 94, 102, 107, 114, 117
- User experience 6, 89, 90
- W**
- Web requests 94
- wwPDB 36, 37
- Y**
- yeast 22, 33, 35, 67, 77, 112, 115

