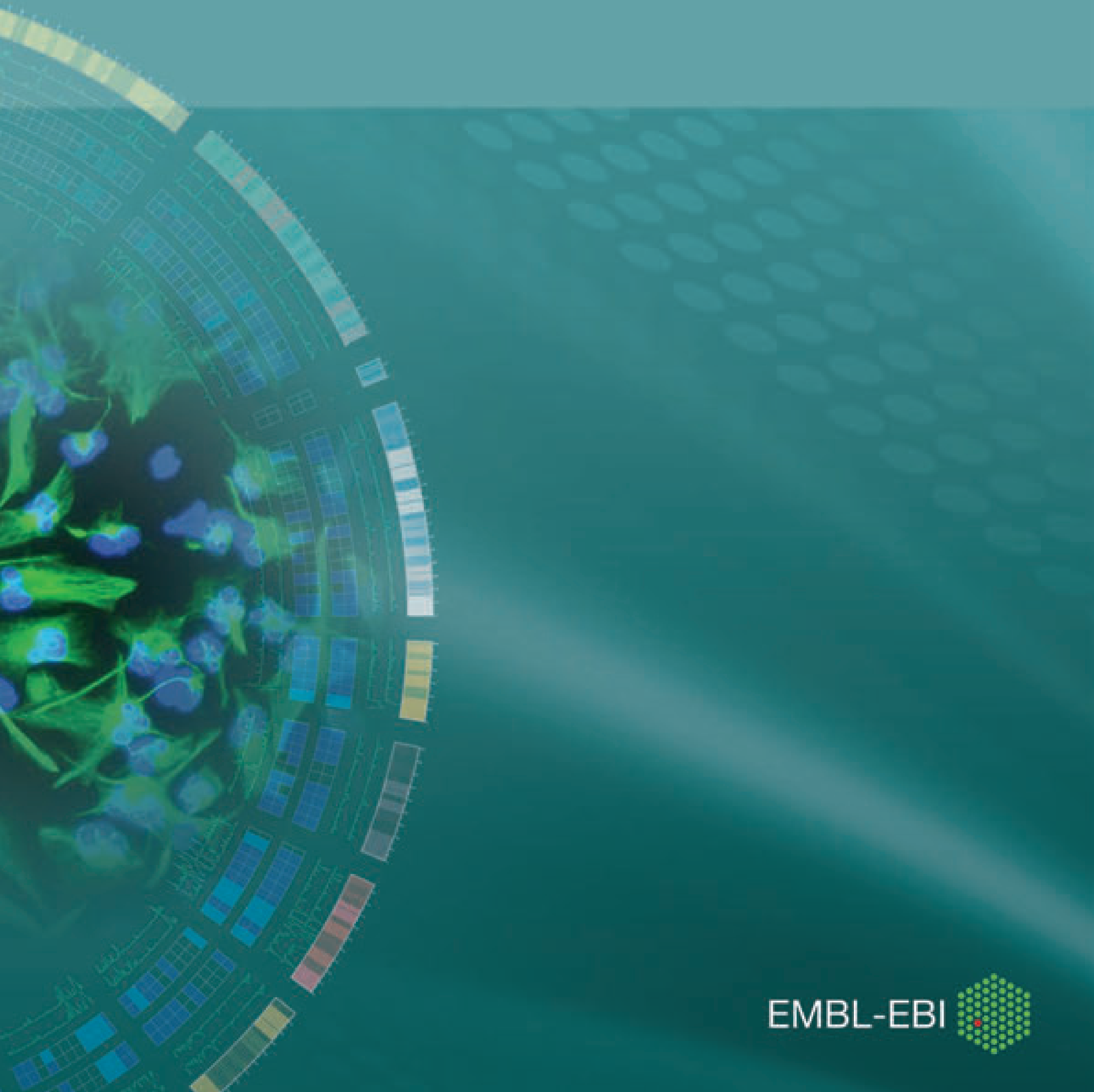
EMBL-European Bioinformatics Institute

Annual Scientific Report 2012



EMBL-EBI

This publication was produced by EMBL-EBI’s External Relations Team.
For more information about EMBL-EBI please contact: comms@ebi.ac.uk

Table of contents

The EMBL-European Bioinformatics Institute
The hub for bioinformatics in Europe
EMBL-EBI
The EMBL-Euro
The hub for bioinfor

Foreword

Welcome to the 2012 Annual Scientific Report of the EMBL-European Bioinformatics Institute. This document gives an overview of our institute's major achievements during the calendar year, reflecting progress in our mission areas of service, research, training, industry and European co-ordination.

If you have visited the Genome Campus over the past year, you will have noticed a new building under construction. We were fortunate to receive funding in 2011 from the UK government's Large Facilities Capital Fund to expand our activities, and in June 2012 EMBL-EBI Director Janet Thornton broke the ground at a landmark ceremony. The rapid progress of construction means that the building is scheduled for completion in autumn 2013. The South Building will house the ELIXIR Hub as well as an Industry and Innovation Suite, which will promote the translation of biological information to applications in medicine, biotechnology and the environment.

Our funding remained stable in 2012, despite another year of austerity in many countries, and our staff numbers also remain steady. Yet we continued to handle an exponential rise in biological data and unprecedented growth in the use of our resources.

All of our efforts rely on close international collaborations. 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 allow us to build an extensive network of colleagues. We look forward to strengthening these links in the coming year and, as always, to creating new collaborations.

JThornton.

Janet Thornton, Director

Rolf Apweiler, Joint Associate Director

Ewan Birney, Joint Associate Director

EMBL-EBI 2012

It was a year of transition, as we bade a fond farewell to Associate Director Graham Cameron, and welcomed Rolf Apweiler and Ewan Birney as joint Associate Directors. It was also a year of honours, as Drs Apweiler and Birney were appointed members of EMBO and our Director, Professor Janet Thornton, was made Dame Commander of the British Empire in recognition of her contribution to bioinformatics.

Our research programme went through major changes as well, as we gained new group leaders Pedro Beltrao from the University of California, San Francisco; Oliver Stegle from the Max-Planck Institutes in Tübingen; and Sarah Teichman from the MRC Laboratory of Molecular Biology in Cambridge. In the same year, we were very sad to say farewell to three talented and experienced group leaders, who reached the end of their EMBL contracts: Nick Luscombe, now at University College London and the Cancer Research UK London Research Institute; Nicolas Le Novère, who has moved a short way to the Babraham Institute, Cambridge; and Dietrich Rebholz-Schuhmann, now at the University of Zurich's Institute of Computational Linguistics.

In 2012 our protein services saw a change of leadership, as Alex Bateman left the Wellcome Trust Sanger Institute and moved across the Genome Campus to take on the responsibilities formerly held by Rolf Apweiler. His team brings with them the Pfam, Rfam, TreeFam and MEROPS databases. A new Variation Team, led by newcomer Justin Paschall, from the National Center for Biotechnology Information (NCBI) in the US, is creating a unified interface for the European Genome-phenome Archive and the Database of Genomic Variants archive.

Perhaps 2012's largest project, involving every team at EMBL-EBI, was the redesign of the website with user experience as the driver.

Figure 1. In 2012 Professor Janet Thornton was awarded the Commander of the British Empire medal.

The new global website, to be launched in 2013, provides an easily navigable interface to EBI services and features consistent functionality which upholds the unique identity of resource brands. Many teams expended significant time and effort creating and testing new user interfaces, for example UniProt and InterPro, and their findings contributed significantly to the overall design strategy.

Research

The ENCODE project dominated scientific news worldwide in September. Co-led by Ewan Birney at EMBL-EBI and funded by the National Human Genome Research Institute in the US, the project comprised over 400 scientists in 32 labs throughout the world who produced a detailed map of genome function. ENCODE's staggering output inspired a new publishing model: upwards of 30 papers were published under open-access license in several different peer-reviewed journals, with the contents linked by topic and united for optimum exploration in a single interface, provided by *Nature* (ENCODE Project Consortium, 2012). The launch of ENCODE was all about public engagement, with a press conference featuring an exhibit at the London Science Museum and aerial acrobats re-enacting gene expression. The event captured the imagination of the world's top-tier television and print media.

The 1000 Genomes Project published a map of normal human variation: a major milestone for the Flicek team, which co-led the project's data co-ordination centre with the National Centre for Biotechnology Information (NCBI). This vast dataset is freely available via the Amazon Web Services Cloud (The 1000 Genomes Project Consortium, 2012).

Petra Schwalie in Paul Flicek's research group led efforts to develop a new, integrated model of the evolution of the transcription factor CTCF. Her work, carried out with Duncan Odom's group at the University of Cambridge, explains the origin of some 5000 highly conserved CTCF binding events in mammals and sheds light on how these binding sites are moved and amplified (Schmidt et al., 2012).

Another gene expression study in Nick Luscombe's lab found that an organism's most valuable assets are protected most carefully. Their work demonstrated that mutation rates in bacteria vary by more than an order of magnitude, with highly expressed genes having the lowest mutation rates (Martincorena et al., 2012).

Janet Thornton's group continues its painstaking work to understand the evolution of enzyme mechanisms. Nick Furnham, a postdoc in Janet's group, worked with Christine Orengo's group at University College London to develop a new resource called FunTree and used it to study 276 enzyme superfamilies. This allowed them to determine the extent to which enzyme functions have changed over the course of evolution–findings that have important implications for the development of new therapeutics.

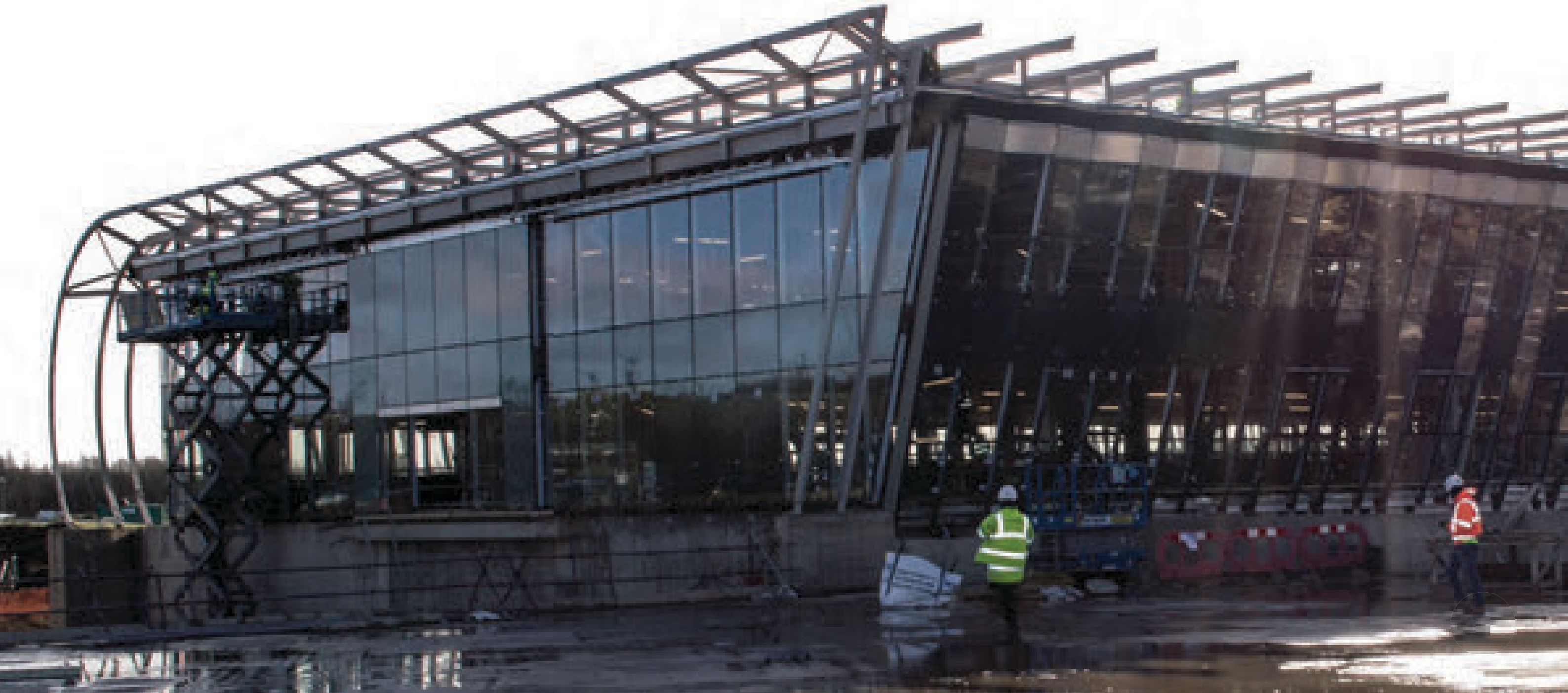
Services

After broad consultation with the service teams and support groups, we reorganised our services into 'clusters' to reflect the research communities they serve. The clusters provide a better framework to make strategic decisions whilst still enabling a deep engagement with the communities.

The new clusters are: genes, genomes and variation, led by Paul Flicek; molecular atlas, led by Alvis Brazma; proteins and protein families, led by Alex Bateman; molecular and cellular structures, led by Gerard Kleywegt; chemical biology, led by John Overington; molecular systems, also led by John Overington; and cross-domain tools and resources, unified under Rolf Apweiler.

As a result of the exercise, the ArrayExpress Archive, Expression Atlas, PRIDE and MetaboLights resources are now clustered together, in the interests of combining them into a single portal that provides access to integrated views of –omics data.

Figure 2. The new South building, which will house the ELIXIR Technical Hub, under construction on the Genome Campus.



New service developments

- Understanding the basis of crop diseases was the driver behind the launch of PhytoPath, a new portal for plant pathogen data, while the Kersey team's involvement in the transPLANT project gave rise to a new integrative portal for plant genomics data.
- The EBI Metagenomics resource was officially launched by the Hunter and Cochrane teams, enabling users to submit, archive and analyse genomic information from environments containing many species.
- The Enzyme Portal, launched in 2012, draws together data that previously resided in ten different databases. Its development and design were based entirely on user demand and feedback (Pavelin et al., 2012).
- MetaboLights, another resource launched by the Steinbeck team, gives a home to metabolomics experiments and derived information.
- The neglected disease community received a boost from ChEMBL, as the Overington team collaborated with the Medicines for Malaria Venture to provide one-stop access to MalariaBox and other open-access data, including new, high-value malaria and tuberculosis datasets.
- UK PubMed Central re-launched as Europe PubMed Central late in the year, as three European funders, including the ERC, joined forces with 18 British funders. Europe PMC, led by EMBL-EBI, is now our single literature brand. This is expected to increase visibility and exploration of this cross-cutting resource, and to focus our efforts in this domain.

Training

- Our user-training programme continued to develop alongside our service offerings and in 2012 actively involved more than 150 members of EMBL staff in 232 events, reaching an audience of over 9770 people in 35 countries on six continents. The Train online portal saw a rapidly rising number of users in its first year, and our hands-on programme offered new courses on –omics-based data analysis for plant biologists and a combined experimental/computational metagenomics course, amongst many others. The Cochrane and Brazma teams significantly expanded their training to cope with demand, both from large-scale sequencing centres and research groups incorporating next-generation sequencing into their portfolio of methodologies.

European co-ordination

ELIXIR, the pan-European research infrastructure for biological information, was the focus of our public affairs activities as it entered the last year of its Preparatory Phase. Five member states signed the ELIXIR Memorandum of Understanding in 2012, bringing the total number of signatories to 15. EMBL-EBI lent close support to the efforts of member states in the drafting of the ELIXIR Consortium Agreement (ECA), which will serve as the legal basis for: ELIXIR; the ELIXIR Hub interim operating budgets for 2012 and 2013; and the representation of ELIXIR within the EU institutions. A founding Director was recruited, and in 2013 Dr Niklas Blomberg of AstraZeneca R&D Mölndal, Sweden will take up his post.

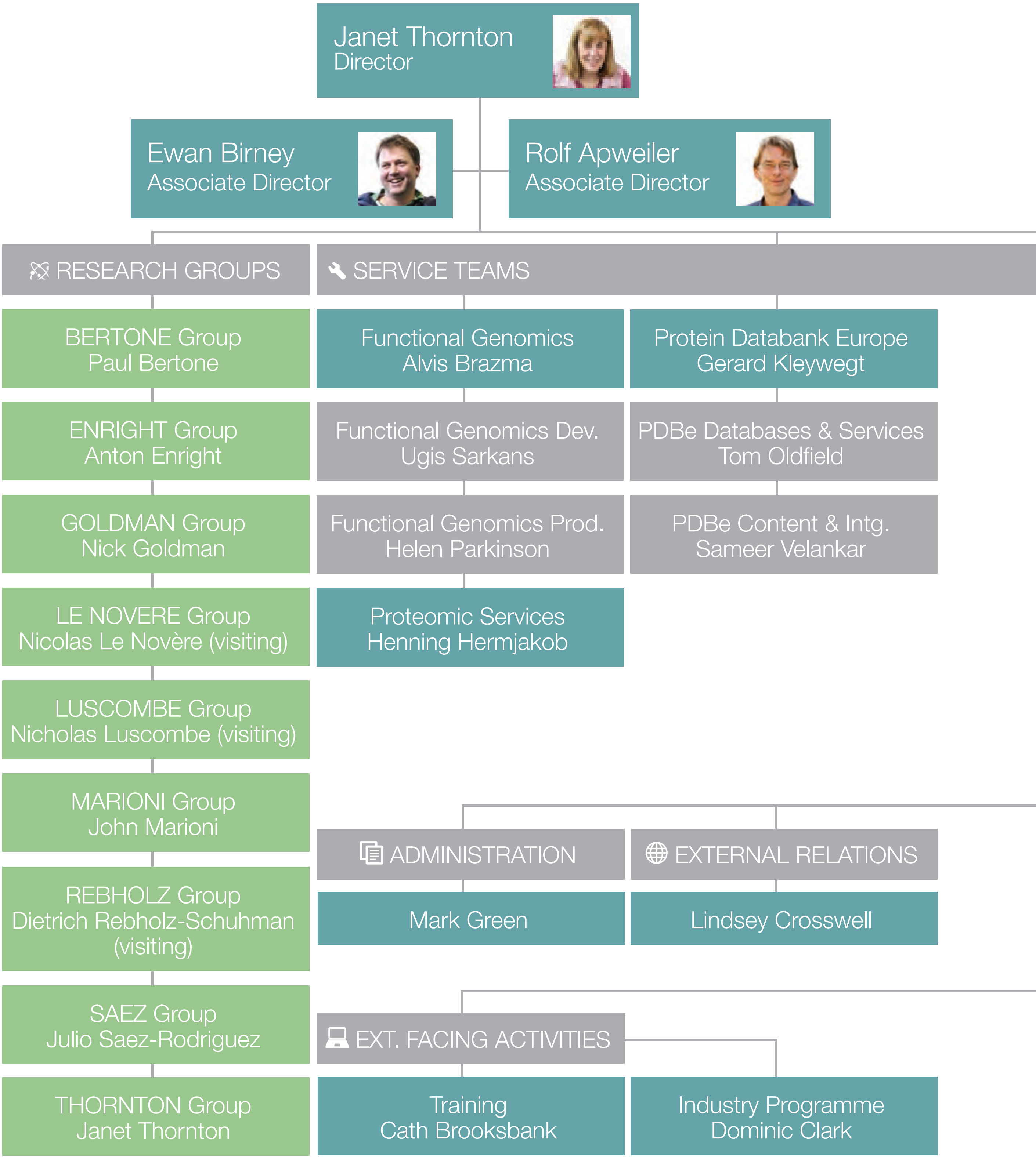
Acknowledgments

We thank our major funders: the member states of EMBL, the European Commission, the US National Institutes of Health, the Wellcome Trust and the UK Research Councils. Clearly, without these funds we would not be able to deliver the high-quality services and research for which we are known.

Selected references

- ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.
- The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.
- Pavelin, K. et al. (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comp Biol* 8, e1002554.
- Goncalves, A. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* 22, 2376–2384
- Schmidt, D. et al. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148, 335-348.
- Martincorena, I. et al. (2012) Evidence of non-random mutation rates suggests an evolutionary risk-management strategy. *Nature* 485, 95-98.

Organisation of EMBL-EBI Leadership 2012



Janet Thornton
Director

Ewan Birney
Associate Director

Rolf Apweiler
Associate Director

RESEARCH GROUPS

BERTONE Group
Paul Bertone

ENRIGHT Group
Anton Enright

GOLDMAN Group
Nick Goldman

LE NOVERE Group
Nicolas Le Novère (visiting)

LUSCOMBE Group
Nicholas Luscombe (visiting)

MARIONI Group
John Marioni

REBHOLZ Group
Dietrich Rebholz-Schuhman (visiting)

SAEZ Group
Julio Saez-Rodriguez

THORNTON Group
Janet Thornton

SERVICE TEAMS

Functional Genomics
Alvis Brazma

Functional Genomics Dev.
Ugis Sarkans

Functional Genomics Prod.
Helen Parkinson

Proteomic Services
Henning Hermjakob

Protein Databank Europe
Gerard Kleywegt

PDBe Databases & Services
Tom Oldfield

PDBe Content & Intg.
Sameer Velankar

ADMINISTRATION

Mark Green

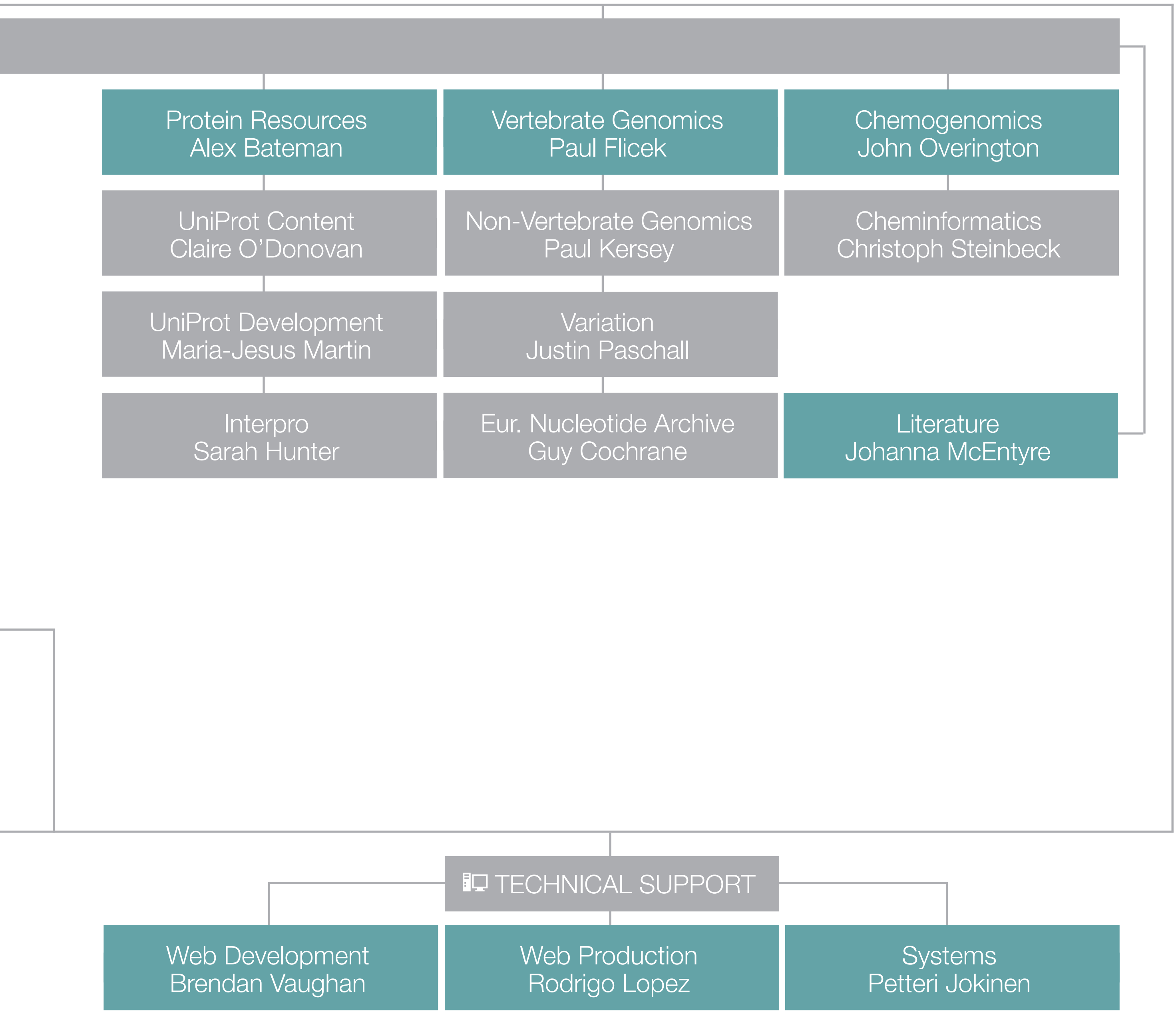
EXTERNAL RELATIONS

Lindsey Crosswell

EXT. FACING ACTIVITIES

Training
Cath Brooksbank

Industry Programme
Dominic Clark



Protein Resources
Alex Bateman

Vertebrate Genomics
Paul Flicek

Chemogenomics
John Overington

UniProt Content
Claire O'Donovan

Non-Vertebrate Genomics
Paul Kersey

Cheminformatics
Christoph Steinbeck

UniProt Development
Maria-Jesus Martin

Variation
Justin Paschall

Interpro
Sarah Hunter

Eur. Nucleotide Archive
Guy Cochrane

Literature
Johanna McEntyre

TECHNICAL SUPPORT

Web Development
Brendan Vaughan

Web Production
Rodrigo Lopez

Systems
Petteri Jokinen

Genes, genomes and variation

The European Nucleotide Archive (ENA) is at the forefront of managing of today's skyrocketing data submissions, coping with a staggering amount of data from the steady uptake of next-generation sequencing technology.

The team has produced innovative solutions to the challenge of serving and archiving these data and in late 2012 issued the first release of CRAM, an open software toolkit and file format for compressing sequence data. CRAM has been widely embraced by the NGS community.

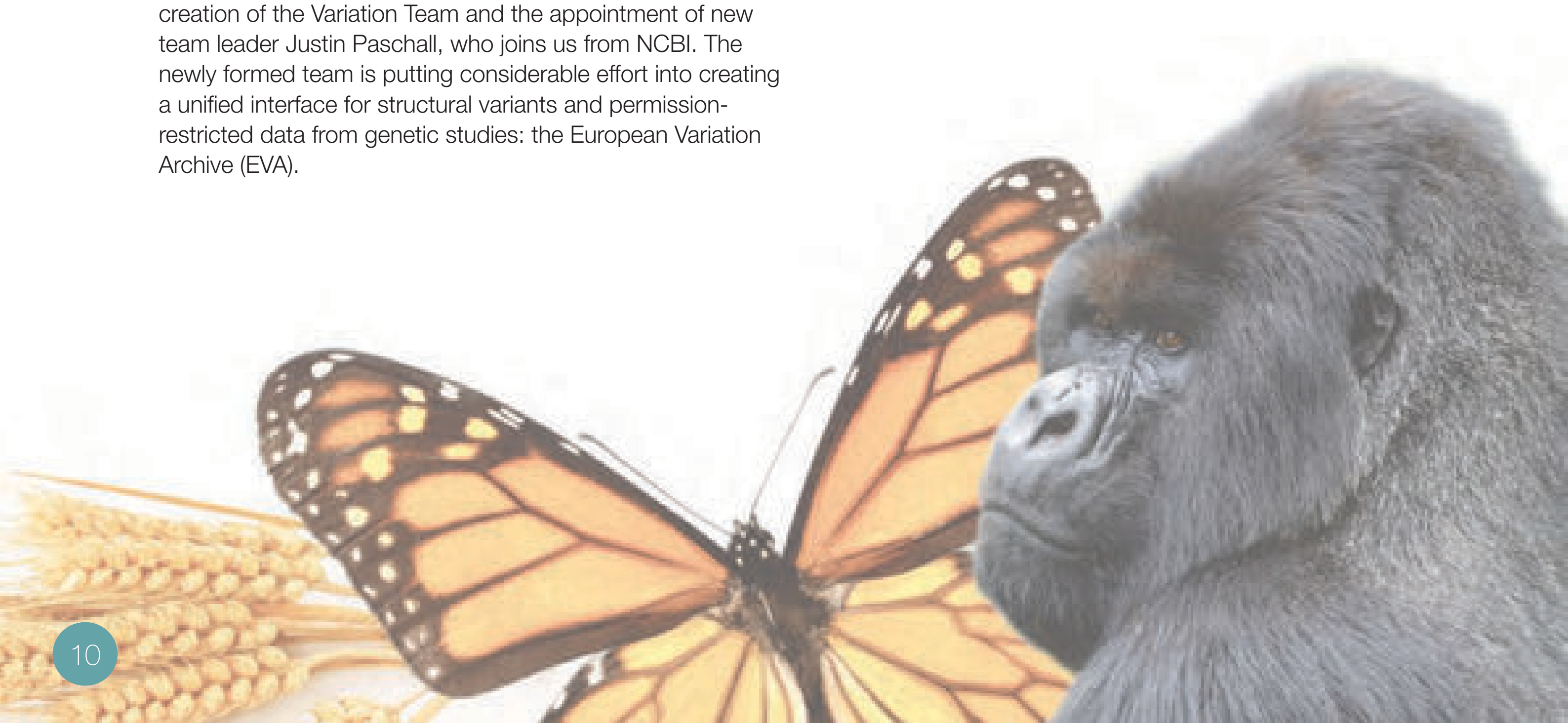
The 1000 Genomes Project published its data in 2012, presenting a map of normal human variation based on the genetic blueprints of more than 1000 healthy individuals from around the world. The success of the project was in no small part due to efficient data co-ordination, led by EMBL-EBI's Vertebrate Genomics team in collaboration with the National Center for Biotechnology Information (NCBI) in the US. The data are freely available from EMBL-EBI, Ensembl and from other locations including via the Amazon Web Services Cloud.

In addition to huge and valuable human genomics datasets from ENCODE and the 1000 Genomes Project, the mountain gorilla made a dramatic entrance into Ensembl in 2012 at the same time as the biting midge, butterflies, barley and wheat joined other nonvertebrate species in Ensembl Genomes. Our Nonvertebrate Genomics team took a step towards tackling crop diseases, launching a new portal for plant pathogen data: PhytoPath. Meanwhile, as key members of the transPLANT project we delivered an integrative portal for plant genomics data and made metabolic data for over 4000 bacterial genomes available through Microme.

Our variation services received a boost in 2012 with the creation of the Variation Team and the appointment of new team leader Justin Paschall, who joins us from NCBI. The newly formed team is putting considerable effort into creating a unified interface for structural variants and permission-restricted data from genetic studies: the European Variation Archive (EVA).

Analysis has become the major bottleneck in genome research, and members of the Vertebrate Genomics team addressed this problem by publishing Cortex, a new way to analyse genetic variants. Cortex is the first software capable of simultaneously assembling multiple eukaryotic genomes.

The ENA and InterPro teams officially launched EBI Metagenomics in 2012, giving researchers a means to submit, archive and analyse genomic information from environments containing many species. We now provide easy access to data from the gut flora of infants, bacteria in pickled cabbage and the troubled waters of the North Sea.



Paul Flicek

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

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

European Nucleotide Archive

The ENA provides globally comprehensive primary data repositories for nucleotide sequencing information. ENA content spans raw sequence reads, assembly and alignment information and functional annotation of assembled sequences and genomes. ENA's palette of services are provided over the web and through a powerful programmatic interface. ENA data and services form a core foundation upon which scientific understanding of biological systems has been assembled. Our exploitation of these systems will continue to develop. With ongoing focus on data presentation, integration within ENA, integration with resources external to ENA, tools provision and services development, the team's commitment is to the utility of ENA content and achieving the broadest reach of sequencing applications.

http://www.ebi.ac.uk/ena/

Ensembl

Ensembl produces and maintains both automatic and manually curated annotation on selected eukaryotic genomes. Automatic annotation is based on mRNA and protein information. Ensembl provides valuable insights into variation within and between species, and allows users to compare whole genomes to identify conserved elements. It is integrated with several other important molecular resources, for example UniProt, and can be accessed programmatically. Ensembl is developed as a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute.

http://www.ensembl.org/

Ensembl Genomes

The falling costs of DNA sequencing have led to an explosion of reference genome sequences and genome-wide measurements and interpretations. Ensembl Genomes (Kersey et al., 2012) provides portals for bacteria, protists, fungi, plants and invertebrate metazoa, offering access to genome-scale data through a set of programmatic and interactive interfaces, exploiting developments originating in the vertebrate-focused Ensembl project. Collectively, the two projects span the taxonomic space.

http://www.ensemblgenomes.org/

EGA

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

http://www.ebi.ac.uk/ega/

DGVa

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

http://www.ebi.ac.uk/dgva/

Metagenomics

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

http://www.ebi.ac.uk/metagenomics/

Genes, genomes and variation

Summary of progress 2012

Paul Flicek

Ensembl and 1000 Genomes

- Issued four major releases of Ensembl, featuring significant updates to our key resources for human and model organism genomics;
- Managed and made available data from the 1000 Genomes Project, which finalised and published its phase 1 analysis;
- Welcomed the mountain gorilla genome and four others to Ensembl;
- Began processing Blueprint and other reference epigenomic data, which complements reference human variation data in 1000 Genomes;
- Launched first version of the data portal for the International Mouse Phenotyping Consortium;
- Formed new Variation Team;
- Provided evidence for cohesin's role in tissue specific transcriptional regulation.

Paul Kersey

Ensembl Genomes

- Launched PhytoPath, a new portal for plant pathogen data;
- Offered metabolic data for over 4000 bacterial genomes through the Microme portal;
- Developed a new integrative portal for plant genomics data as part of the transPLANT project;
- Contributed to significant publications on the butterfly and wheat genomes;
- Provided the first integrated genome browser resource for the barley genome;
- Commenced participation in new projects to assemble, annotate and disseminate the genomes of wheat, barley and the biting midge;
- Issued four public releases of Ensembl Genomes.

Summary of progress 2012

Sarah Hunter

Metagenomics

- Formally launched the EBI Metagenomics portal to the public;
- Added new software analysis tools to pipeline for metagenomics data analysis.

Justin Paschall

EGA and DGVa

- Formed new Variation Team;
- Began development of the European Variation Archive (EVA), a unified interface for DGVa and EGA;
- Oversaw the substantial growth of EGA and DGVa following new ELIXIR partnerships formed with Spain and Finland.

Guy Cochrane

European Nucleotide Archive

- Enabled the capture, processing and presentation of major new raw read, read alignment, assembly and annotation datasets;
- Led community discussions on the use of lossless and lossy compression in sequence data management;
- Released version 1 of the CRAM sequence-data-compression toolkit;
- Built, tested and deployed the ENA advanced search service, supporting rich and granular interactive and programmatic discovery of ENA data;
- Enhanced the Webin data submission tool;
- Deployed new submission and data presentation services for genome assembly information;
- Developed training courses and online training materials and delivered training workshops.

European Nucleotide Archive

Our team develops and maintains the European Nucleotide Archive (ENA), which provides globally comprehensive primary data repositories for nucleotide sequencing information. We provide interactive and programmatic submission tools, curation support, advanced search services and rich integration with data resources beyond ENA.

Our team supports all ENA services through our helpdesk and the training programme, focusing on 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. As nucleotide sequencing becomes increasingly central to applied areas such as healthcare and environmental sciences, ENA data and services have become a core foundation upon which scientific understanding of biological systems are assembled. We focus on data presentation, data integration within ENA and with resources external to ENA, the provision of analytic tools and services development. Our commitment is to the utility of ENA content and to achieve the broadest reach of sequencing applications.

Major achievements

In 2012 we worked extensively on our web-submission applications and their programmatic equivalents. We focused on the reporting of sample information, enhancing the richness of information reported and simplifying the submitter experience. The Webin system is based on checklists of fields to be completed by submitters. We built and deployed several new checklists, including the complete series of MIxS checklists for both read- and sequence-data submissions from environmental sequencing studies. We also deployed a ‘default’ ENA checklist for sample information relating to reads where no other checklist is appropriate, providing the submitter with a guiding set of fields that are likely to be useful. We also added support for single-step multiple sample reporting, offering such functions as spreadsheet upload and editing.

In 2012 we developed and deployed a powerful data warehouse with an accompanying advanced search webservice and interactive query-builder interfaces. The warehouse integrates the many classes of ENA content and supports discovery of ENA content at a highly granular level. Our technology meets our troika of major challenges: high data volume, frequent data update cycles and the need for rapid query response. We have made the warehouse available under a RESTful service to which user queries are despatched under an expressive query language. We also provide a ‘query builder’ interactive web interface that helps users assemble their search.

The team achieved a major milestone with the formal release of CRAM sequence data compression software. CRAM, an open software toolkit and file format, provides highly efficient compression and direct computational access to sequence data in compressed form. The exponential growth in data produced in next-generation sequencing experiments is expected to continue in the coming decade as cancer genome sequencing efforts begin to scale up. CRAM provides an innovative way to store and use these data, overcoming a major bottleneck to progress. CRAM will provide the technology used by ENA to deal with this growth, and we have been working with major stakeholders to ensure that CRAM is usable in all areas of bioinformatics.

Our work on genome assemblies in 2012 involved adapting a number of our submission services to provide a new, simplified route for the submission of sequence contigs, assembly information and assembled sequence. We deployed a new presentation for this information: ENA browser, including discoverability and retrieval of assembly-related data objects (e.g., scaffolds and chromosomes) and integrated browsing between these objects.

Guy Cochrane

PhD University of East Anglia, 1999.

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

Future plans

In 2013 we will work extensively on interfaces for data discovery and retrieval. Our work on the warehouse and advanced search infrastructure is sufficiently mature to be leveraged for a number of general and use-specific interfaces. The current query builder is powerful, protecting the user from ENA's internal technical organisation; however, to be used to full advantage it requires knowledge of annotation and reporting conventions. We will maintain and improve this interface but will also work on more intuitive interfaces for less frequent users. We plan to build a number of such interfaces with an early focus on the phylogenetics community.

We will continue to develop and improve the CRAM sequence data compression software as a distributed codebase and work with user communities in its application. In addition, we will integrate CRAM technology into ENA read data processing pipelines. While submissions of data in CRAM format are already accepted, integration into our validation, processing, loading and data output pipelines will be a focus of development work in 2013.

By integrating ENA content in the ENA browser and associated RESTful services we have helped consumers understand and collate collections of data and interpret the nature of sequencing-based studies. In 2013 we will work on further integration that will be helpful to the submitting ENA user, simplifying the experience and lightening the task of internal data integration. Specifically, we will merge two codebases ('EMBL-Bank Webin' and 'SRA Webin') and withdraw an older codebase with redundant functionality.

We will extend our range of checklists, which support data submission and validation pipelines, and provide a basis for the systematic configuration of indexing tools and data presentation interfaces. We will continue to integrate ENA data with external resources, including literature.

Selected publications

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

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

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

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

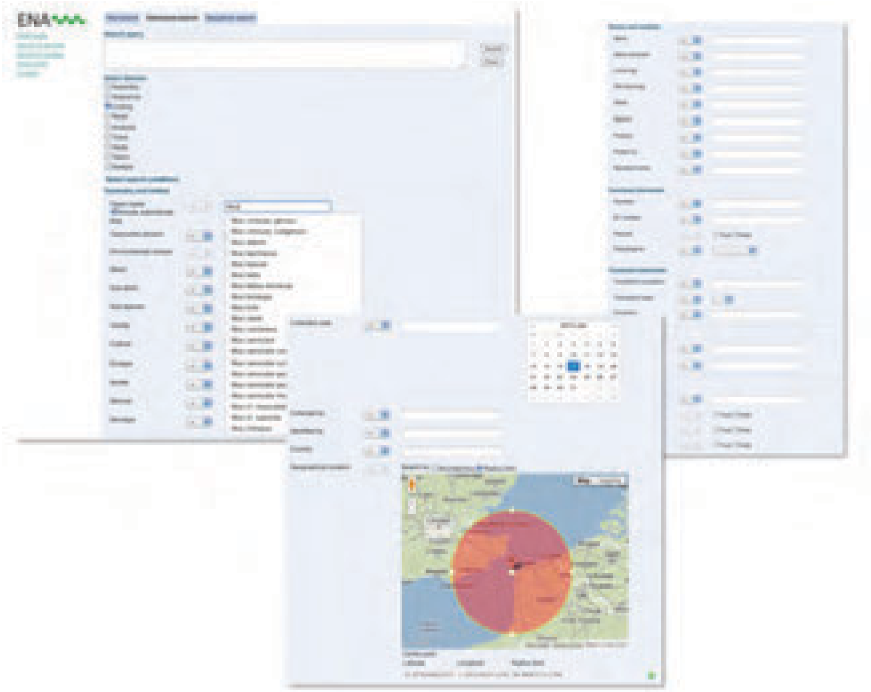


Figure. The ENA advanced search query builder: a page in which the submitter defines taxonomic, date and geographical constraints on data to be retrieved.

Vertebrate genomics

The Vertebrate Genomics team develops the Ensembl genome annotation resources and analysis infrastructure in collaboration with the Wellcome Trust Sanger Institute; creates informatics resources as part of worldwide efforts to distribute data and materials for mouse models, phenotyping and related research and leads data management activities for several large-scale genomics projects, including the 1000 Genomes Project and Blueprint as part of the International Human Epigenome Consortium (IHEC).

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

We also have an active research effort focused on the evolution of transcriptional regulation with an ultimate goal of understanding mechanisms and maintenance of cell-type specificity. Our major research collaboration with Duncan Odom's group at the University of Cambridge continues to pioneer methods of comparative regulatory genomics for biological discovery across a wide range of mammalian and vertebrate species. We are also interested in integrated analysis techniques to assess the interaction of ubiquitously expressed proteins and tissue-specific factors in tissue-specific gene regulation.

Major achievements

Ensembl's four major releases in 2012 included updating our mouse genome resources to the new GRCm38 assembly, the incorporation of several 'patch' release updates to the human genome assembly and extensive updates to our variation and regulation resources for human and other species. We added five new supported species, for example the 'living fossil' coelacanth, the economically important Nile tilapia and the melanoma research model *Xiphophorus maculatus*.

We led significant efforts internationally to standardise annotation of the functional consequences of sequence variation and made several important updates to key tools such as the Ensembl Variant Effect Predictor (VEP). We also launched a new REST API for Ensembl data to support new methods of programmatic access to the datasets and have invested in significant development for our cloud-based annotation and other cloud services. Ensembl continues to directly engage its users through a variety of channels, ranging from on-line video examples and social media to nearly 100 on-site training courses across Europe and around the world.

Our mouse informatics initiatives are continuing to collect and distribute resources in the context of Infrafrontier European Research Infrastructure and the International Mouse Phenotyping Consortium. This work is done in collaboration with Helen Parkinson in the Functional Genomics team and other partners.

We have received the first Blueprint data and are currently collecting data from other IHEC partners and processing this data through standard analysis pipelines before it is released to the scientific community. This reference epigenomic data complements the reference human variation data that we continue to manage and make available as part of the 1000 Genomes Project, which finalised and published its phase 1 analysis in late 2012. We are currently collecting and will then immediately release data for the final phase of the 1000 Genomes Project.

During 2012, EBI's variation database activities that had been a part of the Vertebrate Genomics team since 2007 were transferred to the newly created Variation Team under Justin Paschall's leadership. The transferred resources include the European Genome-phenome Archive (EGA) and the DGVa database of copy number and structure variation data. Both resources saw significant growth in 2012 with the EGA establishing partnerships with institutes in Finland and

Paul Flicek

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

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

Spain under the auspices of ELIXIR. We continue to interact closely with these key databases in the context of our existing activities.

Our research into the role of cohesin in tissue-specific transcriptional regulation, led at EMBL-EBI by PhD student Andre Faure, discovered a relationship between sites in the mouse genome occupied by a large number of transcription factors and the cohesin complex. Furthermore, this analysis showed evidence that cohesin may stabilise transcription factor binding to sites in the genome with relatively weak binding sequences. A separate project undertaken in collaboration with both the Odom group and Merlin Crossley's group at the University of New South Wales, Australia identified the transcription factor whose binding is disrupted in a rare haemophilia, ending a 20-year search for this missing transcriptional regulator. The causative mutation, within the Factor IX promoter, is the site of perfect functional conservation, but imperfect sequence conservation, which we were able to quantify by mapping transcriptional regulatory evolution in five mammalian species (see Figure).

Future plans

Genome sequencing and genome annotation is becoming increasingly relevant in clinical applications and ensuring that Ensembl's reference resources will be valuable for this application is an important challenge. Throughout genomics, high throughput sequencing enables the sequencing of new species and the creation of new and important datasets. Ensembl will continue to address these needs with increasingly flexible methods of data access, presentation and distribution. The other resources of the Vertebrate Genomics Team are working with specific communities and specific projects to maximise the value of the data that they generate, and part of this effort will include integration of the data into appropriate archives and resources within the team. We will continue to interact closely with Justin Paschall's Variation Team on DGVa and EGA.

Our research efforts with the Odom group are currently concentrated on closely related species, which enable us to probe the first steps of transcriptional regulatory evolution, with future expansions planned to other complex tissues. Additionally, we are seeking to apply the methods that we have developed for analysing genome regulation to datasets collected from multiple points on the differentiation cascade in specific tissues as part of other on-going projects.

Selected publications

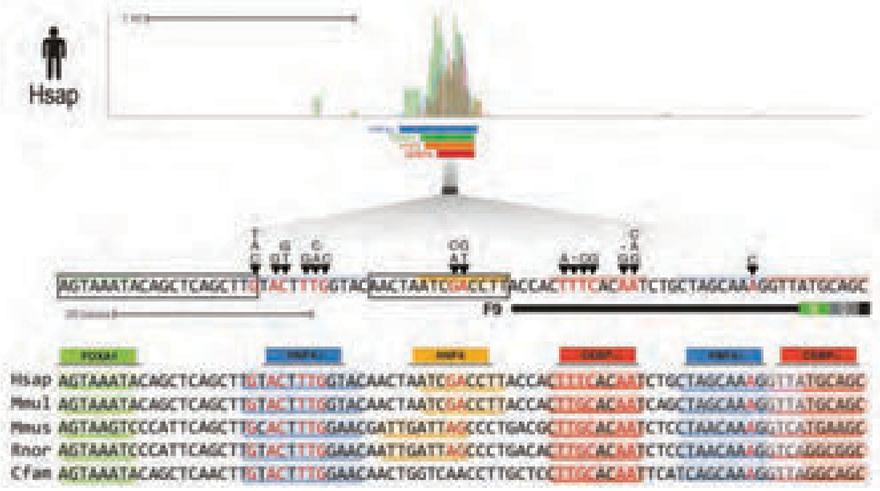
Flicek, P., et al. (2012) Ensembl 2012. *Nucleic Acids Res* 40, D84-D90.

Schmidt, D., Schwalie, P.C., et al. (2012) Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* 148, 335-348.

Mallon, A.M., Iyer, V., et al. (2012) Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. *Mamm Genome* 23, 641-652.

Faure, A.J., Schmidt, D., et al. (2012) Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Res* 22, 2163-2175.

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65.



Hsap
AGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTTCACAATCTGCTAGCAAAGGTTATGCAGC
F9
Hsap AGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTTCACAATCTGCTAGCAAAGGTTATGCAGC
Mmul AGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTGCACAATCAGCTAGCAAAGGTTATGCAGC
Mmus AGTAAGTCCCATTCAGCTTGCACTTTGGAACGATTGATTAGCCCTGACGCTTGCACAATCTCCTAACAAAGGTCATGAAGC
Rnor AGTAAATCCCATTCAGCTTGTACTTTGGAACAATTGATTAGCCCTGACACTTGCACAATCTCCTAACAAAGGTCAGGCGGC
Cfam AGTAAATACAGCTCAACTTGTACTTTGGAACAACTGGTCAACCTTGCTCCTTGCACAATTCATCAGCAAAGGTTAGGCAGC

Figure. Evolutionarily conserved transcription factor binding around the Factor IX gene promoter in five species (human, macaque, mouse, rat and dog). Mutations known to cause haemophilia are indicated by triangles about the genome sequence and are consistently found at sites of conserved transcription factor binding.

Nonvertebrate genomics

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

Our leading partnerships include VectorBase (Megy et al., 2012), a resource focused on the annotation of invertebrate vectors; WormBase (Yook et al., 2012), a resource for nematode biology; and PomBase (Wood et al., 2012), a resource focused on the fission yeast *Schizosaccharomyces pombe*. In the plant domain, we collaborate closely with Gramene in the US and with a range of European groups in the transPLANT project. Our major areas of interest include broad-range comparative genomics and the visualisation and interpretation of genomic variation, which is being increasingly studied in species throughout the taxonomy. We have developed a new portal for plant pathogen data: PhytoPath (launched in early 2012), and are involved in the development of Microme, a new resource for bacterial metabolic pathways.

By collaborating with EMBL-EBI and re-using our established toolset, small communities can store, analyse and disseminate data more cheaply and powerfully than if they develop their own tools. Our team helps very small communities with little informatics infrastructure perform highly complex and data-generative experiments—the type of work once the sole domain of large, internationally co-ordinated sequencing projects.

Major achievements

In 2012 we issued four public releases of Ensembl Genomes, and contributed to the regular data releases of VectorBase, WormBase and PomBase. We launched PhytoPath, a new public portal for plant pathogen data as part of a collaboration with Rothamsted Research. Ensembl Fungi and Ensembl Protists now contain 24 plant pathogens, and offer variation data for four species, EST alignments for 12 species, RNA-seq data for four species and DNA alignments for eight species.

We provided access to metabolic data for over 4000 bacterial species through the Microme portal and developed an integrated search facility over many European plant genomics resources through the transPLANT website. We developed a new version of the VectorBase website that will launch in 2013.

Our team made a significant contribution to the annotation of the genome of the butterfly *Heliconius melpomene*; and organised and integrated the emerging genomic data from wheat and barley. We also added several significant plant species in the Ensembl Genomes site, including tomato, potato and turnip—the first Brassica genome.

We initiated new projects to annotate and disseminate genomic data from the biting midge *Culicoides sonoerensis* as well as wheat and barley. As part of the VectorBase project, we led community efforts to complete the annotation of the genomes of the tsetse fly (*Glossina moristans*) and the kissing bug (*Rodnius prolixus*). We also worked on the primary annotation of the scuttle fly (*Megaselia scalaris*) and the salmon louse (*Lepeophtheirus salmonis*). Through our participation in the Assemblathon project we contributed to international efforts to assess standards and methods for genome assembly.

Meanwhile, we enhanced the range of services offered through Ensembl Genomes, creating new features such as a REST-ful API and annotation-aware gene trees. We developed a new, scalable pipeline for the management of variation data, which, at the time of publication, is in alpha testing with selected collaborators.

Paul Kersey

PhD University of Edinburgh 1992. Postdoctoral work at University of Edinburgh and MRC Human Genetics Unit, Edinburgh.

At EMBL-EBI since 1999.

Our team also began working with the Cochrane and Bateman groups to develop RNAcentral, a new resource for functional RNAs that will fill a significant gap in EMBL-EBI's service portfolio.

Future plans

The main focus of our work in 2013 will be accommodating the continued growth in the range of available genomic data. Early in the year we will undertake a major revision of Ensembl Bacteria, increasing its coverage to over 5000 species and providing access to genomic annotation through the standard Ensembl interfaces. We will also develop new data -mining tools to allow effective access to this enormous and fast-growing dataset. Our team will continue to collaborate on the development of resources for cereal genomes, and we expect to see the publication of reference annotations for the tsetse fly and kissing bug in 2013. As data from projects sequencing 18 anophelean species and 50 nematode species becomes available, we will embark on efforts to annotate vector and worm genomes on a larger scale. These larger data releases pose challenges in terms of both annotation method and provision of coherent methods for data access and search; however, they also offer opportunities to improve standards and discover interesting biological stories through the use of consistency metrics.

Selected publications

Brenchley, R., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491, 705–710.

Dasmahapatra, K.K., et al. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94-98.

The Gene Ontology Consortium (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res* 40, D559-D564.

Howe, K., et al. (2012) WormBase: annotating many nematode genomes. *Worm* 1, 14-20.

Kersey, P.J., et al. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res* 40, D91-D97.

Wood, V., et al. (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res* 40, D695-D699.

Yook, K., et al. (2012) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res* 40, D735-741.

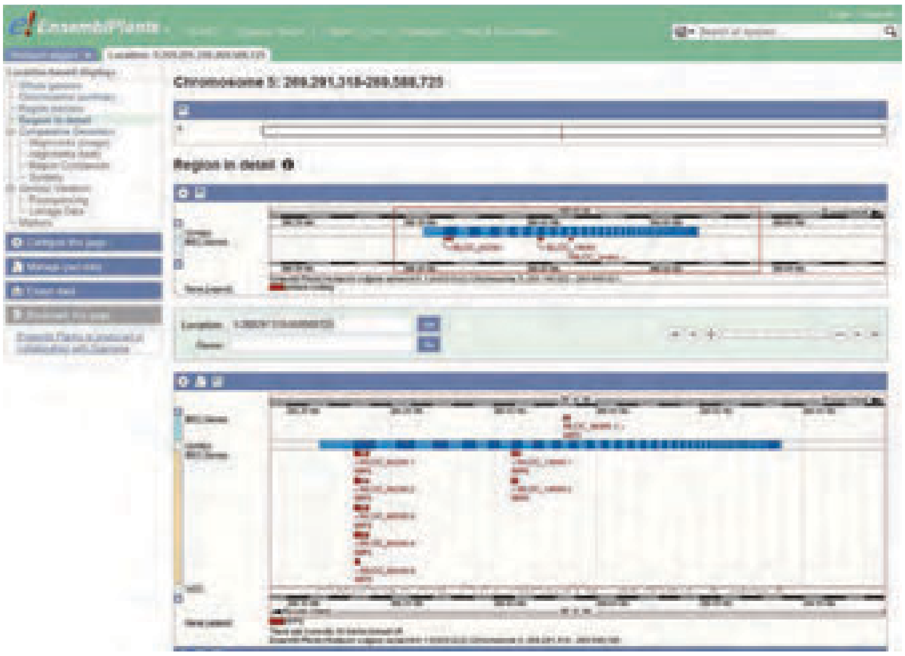


Figure. The barley genome in Ensembl Plants. Published in 2012, the barley gene-space has been sequenced and anchored to a physical map, but the complete genome is not yet fully sequenced or assembled. We collaborated with the International Barley Sequencing Consortium to organise the data and make it available through the Ensembl Plants interface.

Molecular atlas

Life scientists are increasingly using a range of different –omics techniques to study their subjects, integrating transcriptomics, proteomics and metabolomics data to achieve a systems-based understanding. Our ArrayExpress, Expression Atlas, PRIDE and MetaboLights resources are working towards offering a single, integrated view of –omics data.

In 2012 EMBL-EBI launched MetaboLights, a resource for metabolomics experiments and derived information. The Expression Atlas underwent a change, with Robert Petryszak stepping in to lead development of a prototype Baseline Expression Atlas, which uses sequencing-based expression data to report 'absolute' gene expression levels. Meanwhile, ArrayExpress accepted its millionth assay, and a number of substantial improvements were made to its user interface. The Proteomics Services Team now co-ordinates international data deposition strategies for mass spectrometry-based proteomics data through the ProteomeXchange consortium.

ArrayExpress

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

http://ebi.ac.uk/arrayexpress/

Expression Atlas

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

http://ebi.ac.uk/gxa/

PRIDE

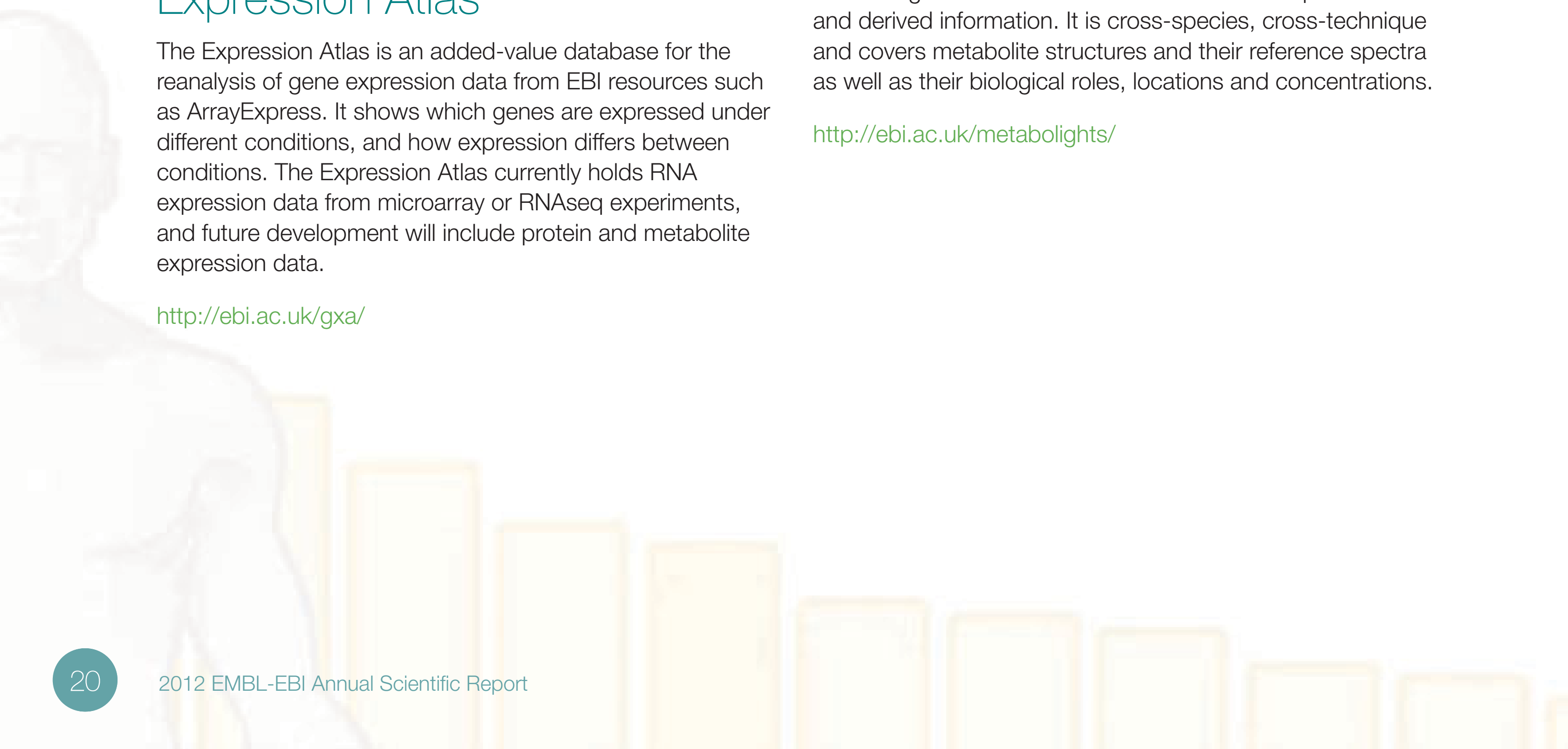
The PRoteomics IDEntifications database is a centralised, standards-compliant, public data repository for proteomics data. It includes protein and peptide identifications, post-translational modifications and supporting spectral evidence.

http://ebi.ac.uk/pride/

Metabolights

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

http://ebi.ac.uk/metabolights/



Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. Postdoctoral research at New Mexico State University, US.

At EMBL-EBI since 1997.

Summary of progress 2012

Ugis Sarkans

ArrayExpress

- Offered new ways of working with ArrayExpress data: R objects, integration with GenomeSpace, visualisation in Ensembl for sequencing-based data;
- Continued evolving the ArrayExpress user interface to provide a richer, more consistent service.

Helen Parkinson

ArrayExpress and Expression Atlas

- Produced EMBL-EBI's first RDF endpoint for gene expression data;
- Released Zooma, an automated annotation knowledgebase supporting ArrayExpress, the EBI BioSamples Database and the Expression Atlas;
- Welcomed the one millionth assay to ArrayExpress.

Henning Hermjakob

PRIDE

- Achieved full production status for the ProteomeXchange consortium, co-ordinating international data deposition strategies for mass spectrometry-based proteomics data;
- Handled over 80 submissions and more than 70 million new spectra;
- Released format-specific libraries (Griss et al., 2012; Reisinger et al., 2012);
- Released an updated deposition tool, PRIDE Converter (Côté et al., 2012);
- Released data analysis support, PRIDE Inspector (Wang et al., 2012);
- Implemented the capability to deposit large raw datasets in PRIDE.

Christoph Steinbeck

Metabolights

- Developed the MetaboLights database and archive, which was released in February 2012;
- Assumed co-ordination of COSMOS (European Commission co-ordination action) for metabolomics standards and data exchange.

Functional genomics

The Functional Genomics team provides bioinformatics services and conducts research in gene expression and high-throughput sequencing applications.

We participate in software development related to biomedical informatics and systems microscopy and are responsible for the Expression Atlas, which is growing to include proteomics and metabolomics data. Together with our Production and Development teams, we develop ArrayExpress, the archive of functional genomics data and the EBI BioSamples Database. We also contribute substantially to training in transcriptomics and the use of EMBL-EBI bioinformatics tools.

Our research efforts centre on developing new methods and algorithms for analysing gene expression data and integrating different types of data across multiple platforms. We are particularly interested in cancer genomics and transcript isoform usage, and collaborate closely with the Marioni group and others throughout EMBL.

Major achievements

The co-ordination of Expression Atlas development was taken over by Robert Petryszak, following the departure of Misha Kaphushesky in early 2012. The major focus was on planning and prototype development for the baseline Expression Atlas, which utilises high-throughput-sequencing-based expression data to report absolute (rather than relative) gene expression levels. We also started working on the close integration of gene-expression, protein-expression and metabolite-measurement data.

Major developments in the ArrayExpress Archive and the EBI BioSamples Database are described by Helen Parkinson and Ugis Sarkans (see also Rustici et al., 2013 and Gostev at al., 2012).

In 2012 we organised and participated in over 25 training events, including Bioinformatics Roadshows and on-site courses. These included the EMBO practical course on the analysis of high-throughput sequencing data, which was the most popular and oversubscribed training event in 2012.

Our team developed a prototype database for systems microscopy data and loaded seven datasets from project partners in the Systems Microscopy Network of Excellence, which was funded under the EU's Seventh Framework Programme (FP7).

As a part of our participation in the GEUVADIS project (funded by the FP7), we analysed mRNA and small RNA from lymphoblastoid cell lines of 465 individuals who participated in the 1000 Genomes Project. Our group led the analysis of transcript isoform use and fusion gene discovery. By integrating RNA and DNA sequencing data, we were able to link gene expression and genetic variation, and to characterise mRNA and miRNA variation in several human populations. All of the data generated in the project are available through ArrayExpress.

The human transcriptome contains in excess of 100 000 different transcripts. We analysed transcript composition in 16 human tissues and five cell lines to show that, in a given condition, most protein coding genes have one major transcript expressed at significantly higher level than others, and that in human tissues the major transcripts contribute almost 85% to the total mRNA. We also found that the same major transcript is often expressed in many tissues. These observations can help prioritise candidate targets in proteomics research and help predict the functional impact of the detected changes in variation studies. Our findings, submitted for publication in 2012, point towards a lower degree of transcriptome complexity than recently estimated. Other research includes an exploration of the utility of gene expression data in the public domain (Rung and Brazma, 2012).

Angela Goncalves, a PhD student in the Functional Genomics group, gained her doctorate in 2012 and published some of her major findings in *Nature Genetics* and *Genome Research* (Goncalves, et al. 2012).

Alvis Brazma

PhD in Computer Science, Moscow State University, 1987. Postdoctoral research at New Mexico State University, US.

At EMBL-EBI since 1997.

Future plans

In 2012 we began the process of integrating Expression Atlas data with PRIDE proteomics data, and this process will continue in 2013. We will also undertake to integrate the metabolomics data in MetaboLights with the Expression Atlas. Our research will continue to focus on large-scale data integration and systems biology. We will develop methods for RNA-seq data analysis and processing, and apply these to address important biological questions, such as the role of alternative splicing and splicing mechanisms. Together with our colleagues at the International Cancer Genome Consortium, we will investigate the impact of cancer genomes on functional changes in cancer development and explore fusion genes and their role in cancer development.

References cited

Rustici, G., et al. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. Nucleic Acids Res 41(D1), D987-D990.

Kutter, C., et al. (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nat Genet 43, 948-955.

Selected publications

Rung, J. and Brazma, A. (2012) Reuse of public genome-wide gene expression data. *Nat Rev Genet* doi: 10.1038/nrg3394.

Fonseca, N.A., et al. (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169-3177.

Goncalves, A., et al. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* 22, 2376-2384.

Gostev, M., et al. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res* 40 (D1), D64-D70. doi: 10.1093/nar/gkr937.

Kapushesky, M., et al. (2012) Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res* 40 (D1), D1077-D1081.

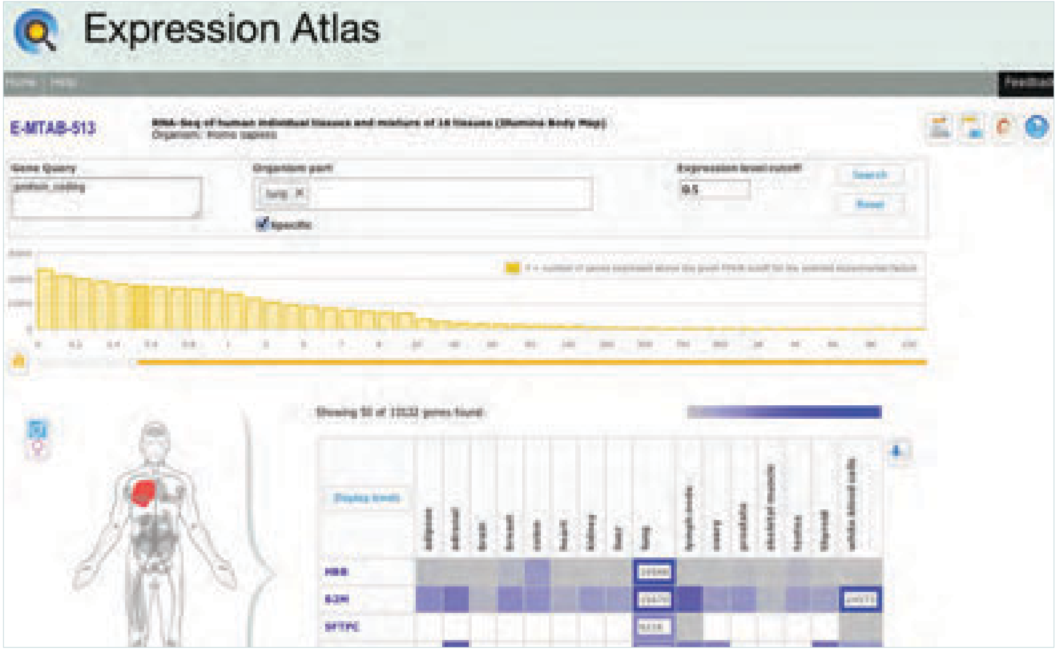


Figure. Prototype of the baseline Expression Atlas.

Functional genomics production

Our team manages data content, software delivery and usability for the ArrayExpress Archive, the Expression Atlas and the EBI BioSamples Database. The complex metadata in these resources represent experimental types, variables and sample attributes for which semantic mark-up in the form of ontologies is required. We develop ontologies and software for the annotation of complex biological data, including the Experimental Factor Ontology (EFO) for functional genomics annotation.

We are key collaborators in several projects funded by the European Commission, the US National Institutes of Health (NIH) and the US National Science Foundation (NSF). These range from Sybaris, which is investigating genome content and gene expression in fungal pathogens, to infrastructural projects such as BioMedBridges, which is developing integrative software to link different biomedical domains. As part of our collaboration with the National Center for BioOntology in the US, we support functional genomics data integration and analysis by developing tools for ontology manipulation and for the semantic web. With our partners in KOMP2 and the International Mouse Phenotyping Consortium, we manage, analyse and distribute complex phenotypic data from 20 000 knockout mouse lines. In the context of the InfraCOMP project, we promote mouse data integration in Europe.

Major achievements

We delivered a zoom-able, integrated browser for the National Human Genome Research Institute (NHGRI) genome-wide association study (GWAS) catalogue that uses semantic web technologies and EFO to support rich, dynamic queries. It allows users to query by trait, SNP (e.g., obesity, type II diabetes) and clinical measurement (e.g., blood glucose) to visualise associated regions of the genome. These in turn are integrated with Ensembl and the ontology term, synonyms and definitions.

As part of the EBI website redesign, our team is working to integrate our data more closely with other resources. By integrating the Expression Atlas dataset of under- and over-expressed genes with Ensembl, Reactome and Uniprot we have opened up possibilities for richer queries and data mining. This work represents EMBL-EBI's first production SPARQL endpoint.

EFO had two major releases in 2012: the first included numerous terms for the description of GWAS data and the second imported the Orphanet Ontology for genetic diseases. EFO supports Ensembl's annotation of phenotypes associated with variation.

A new NSF-funded collaboration with the Gramene plant database and the Ensembl Genomes team uses our team's curation and analysis expertise and the Atlas infrastructure to integrate gene expression datasets generated using Array and RNA-Seq technologies for the plant community. The results are generating new plant content for Atlas users.

Our team's three MSc students worked with the Literature Services team to evaluate ontologies and literature versus ArrayExpress database content. They also evaluated the use of ontologies in R to perform phenotype-driven analyses and compared RNA-Seq analysis methodologies with array-based technologies for the same samples. All of the students received distinctions for their work.

Future plans

In 2013 we will increase our use of semantic web technologies in production as we drive automatic annotation and support curation using RDF triple stores. We will also explore the evolution and repair of RDF knowledgebases in the new DIACHRON project. Our team plans to release the

Helen Parkinson

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

At EMBL-EBI since 2000.

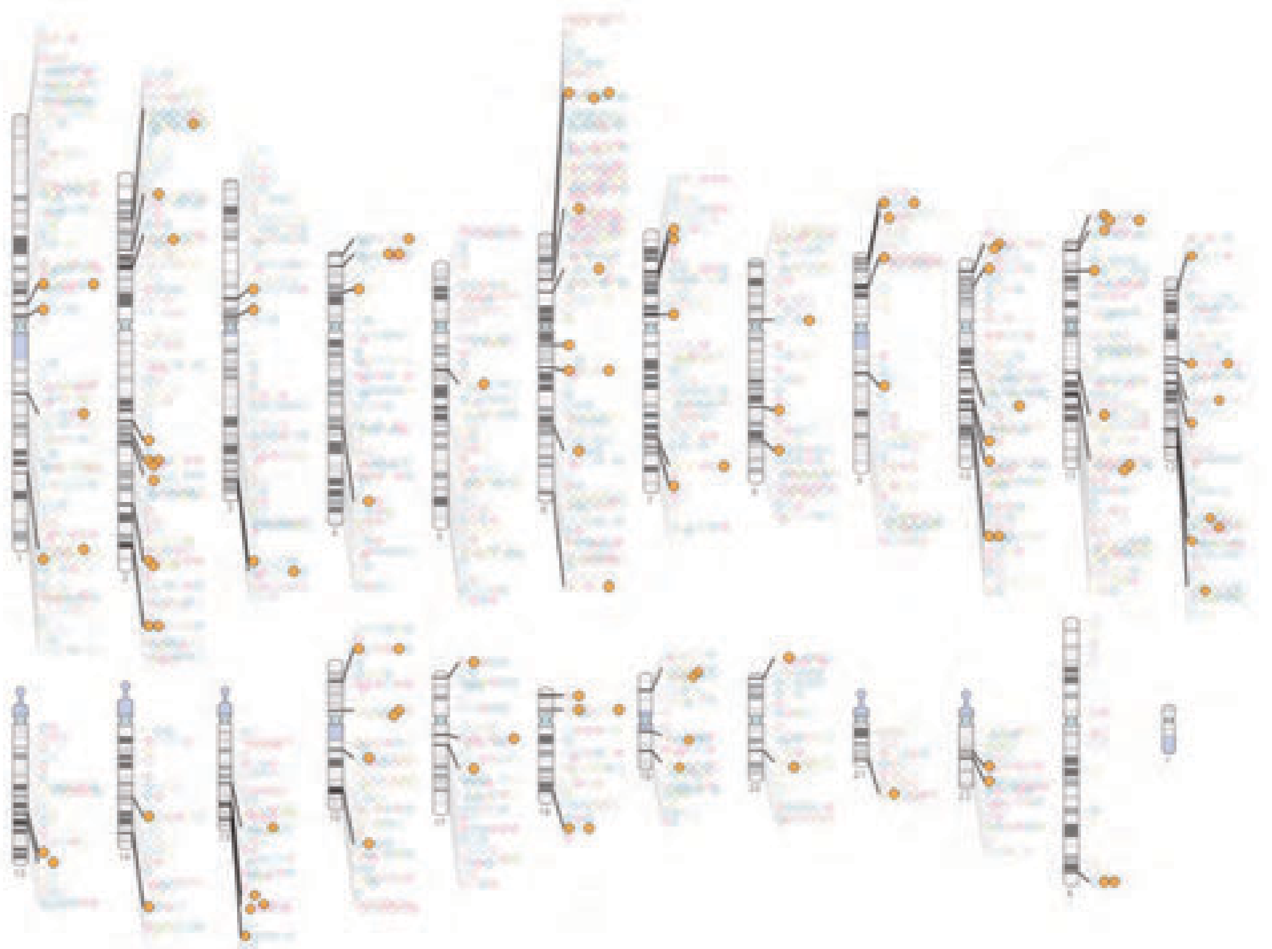


Figure. SNPs at 10-8. P value associated with Type II Diabetes, shown in orange against all SNPs from the GWAS Catalog, 2012. www.ebi.ac.uk/fgpt/gwas.

BioRDF-R package, which will enable enrichment analysis of RDF data. We will host MSc students exploring human-mouse data integration. We will contribute standards and service registries to the BioMedBridges project and will work towards semantic interoperability.

Selected publications

Rustici, G., et al. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res* 41 (D1), D987-D990.

Jupp, S., Parkinson, H. and Malone, J. (2012) Semantic Web Atlas: putting gene expression data into biological context. *SWAT4LS*.

Functional genomics development

Our team develops software for ArrayExpress, a core EMBL-EBI resource, and the EBI BioSamples Database, which centralises biological sample data.

Together with the Functional Genomics Production team, we build and maintain data management tools, user interfaces, programmatic interfaces, annotation and data submission systems for functional genomics resources. We also collaborate on a number of European 'multi-omics' projects in a data management capacity.

Major achievements

In 2012 we concentrated on data integration and quality. We implemented a link to GenomeSpace, a data-integration platform, making it easy for GenomeSpace users to retrieve and process data from ArrayExpress. To help users visualise sequencing-based ArrayExpress data, we began generating BAM files and linking them through a genome browser interface. We also enhanced the Bioconductor package for importing ArrayExpress data, and pre-processed a significant portion of ArrayExpress data; the resulting R objects being available to use on the ftp server.

We made several incremental improvements to the ArrayExpress user interface, including better visualisation of sequencing experiments. We invested significant effort aligning ArrayExpress with the look and feel of the new EMBL-EBI website, which will be launched in 2013. We also reviewed options for replacing MIAMExpress.

In 2012 we worked with other resources to ensure the EBI BioSamples Database provides useful, dependable sample data management services. We resolved some scalability problems, which made it possible to manage tens of millions of samples. In the process, we redeveloped the user interface, the programmatic interface and the backend layer. Because the user interface has the same underlying software architecture and components as ArrayExpress, it inherited functionality such as ontological query expansion.

The team participated in two large European projects: diXa and EU-AIMS. diXa is providing a sustainable data management infrastructure for the development of non-animal toxicity tests. Our role is to facilitate the reuse of ArrayExpress, PRIDE and MetaboLights and to build a portal linking studies in these resources. We are working on a solution that can be used for similar projects or for data that do not have a proper home because of the novelty of the technology used.

EU-AIMS is a five-year project centred on autism-spectrum disorder studies that began in 2012. Our role is to integrate genomics and imaging data.

Future plans

The next major milestone for ArrayExpress is replacing the aging MIAMExpress tool with Annotare, the next-generation data-submission and curation tool. Annotare will be adjusted for sequencing-based data submissions. The updated ArrayExpress interface will launch with the new EMBL-EBI website in early 2013, and will feature improvements such as batch queries and better representation of very large experiments.

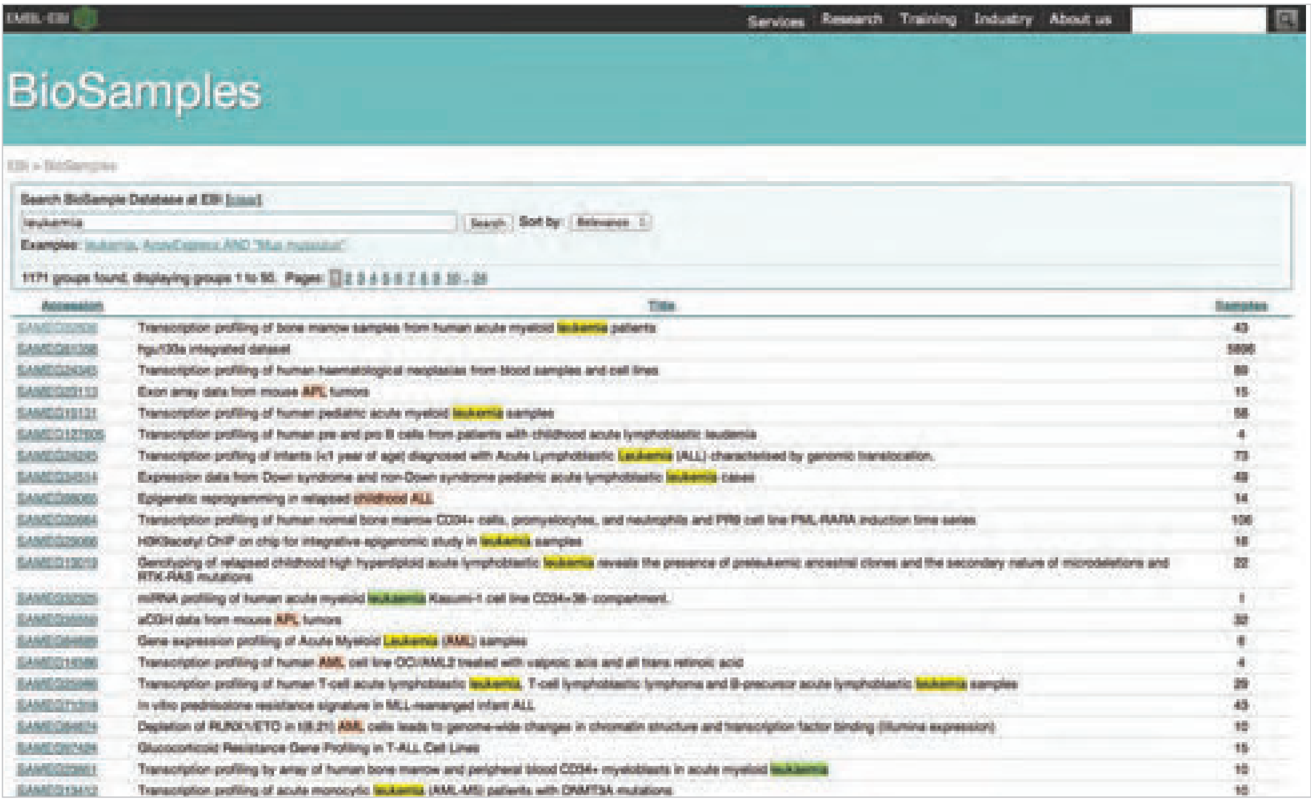
The new relational backend of the BioSamples database will be deployed in early 2013, along with an improved set of programmatic access interfaces. This will allow BioSamples to be used as the primary storage facility for biological sample information for EMBL-EBI resources. The first customer will be the European Nucleotide Archive. We will also enable restricted access to pre-publication samples and work on a data-exchange mechanism with NCBI in the US.

In 2013 we will deploy the first version of the diXa data warehouse, which will serve as a prototype 'BioStudy' database for semi-structured assay data. For the EU-AIMS project we will deploy the first version of the data-management infrastructure and will begin integrating relevant genomics data with MRI and fMRI imaging data.

Ugis Sarkans

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

At EMBL-EBI since 2000.



Search BioSample Database at EBI [clear]

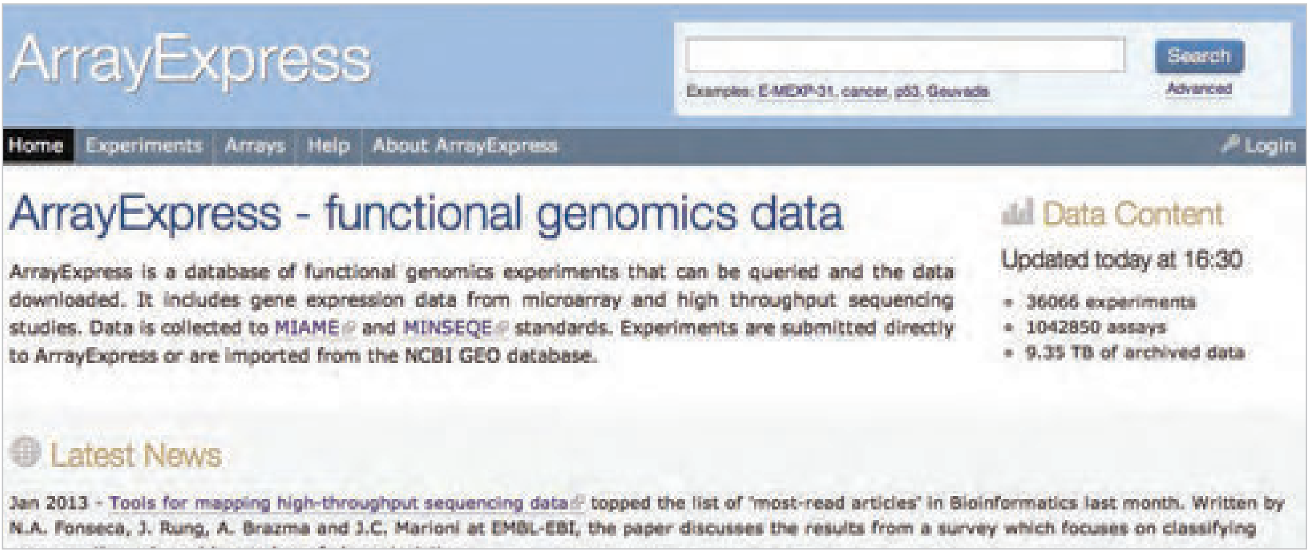
leukemia [Search] Sort by: Relevance

Examples: leukemia, ArrayExpress AND "Mus musculus"

1171 groups found, displaying groups 1 to 50. Pages: 1 2 3 4 5 6 7 8 9 10 .. 24

Accession	Title	Samples
[illegible]	Transcription profiling of bone marrow samples from human acute myeloid leukemia patients	43
[illegible]	hgu133a integrated dataset	5896
[illegible]	Transcription profiling of human haematological neoplasias from blood samples and cell lines	89
[illegible]	Exon array data from mouse APL tumors	15
[illegible]	Transcription profiling of human pediatric acute myeloid leukemia samples	58
[illegible]	Transcription profiling of human pre and pro B cells from patients with childhood acute lymphoblastic leukemia	4
[illegible]	Transcription profiling of infants (<1 year of age) diagnosed with Acute Lymphoblastic Leukemia (ALL) characterised by genomic translocation.	73
[illegible]	Expression data from Down syndrome and non-Down syndrome pediatric acute lymphoblastic leukemia cases	49
[illegible]	Epigenetic reprogramming in relapsed childhood ALL	14
[illegible]	Transcription profiling of human normal bone marrow CD34+ cells, promyelocytes, and neutrophils and PR9 cell line PML-RARA induction time series	106
[illegible]	H3K9acetyl ChIP on chip for integrative epigenomic study in leukemia samples	18
[illegible]	Genotyping of relapsed childhood high hyperdiploid acute lymphoblastic leukemia reveals the presence of preleukemic ancestral clones and the secondary nature of microdeletions and RTK-RAS mutations	22
[illegible]	miRNA profiling of human acute myeloid leukemia Kasumi-1 cell line CD34+38- compartment.	1
[illegible]	aCGH data from mouse APL tumors	32
[illegible]	Gene expression profiling of Acute Myeloid Leukemia (AML) samples	8
[illegible]	Transcription profiling of human AML cell line OCI/AML2 treated with valproic acid and all trans retinoic acid	4
[illegible]	Transcription profiling of human T-cell acute lymphoblastic leukemia, T-cell lymphoblastic lymphoma and B-precursor acute lymphoblastic leukemia samples	29
[illegible]	In vitro prednisolone resistance signature in MLL-rearranged infant ALL	43
[illegible]	Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding (Illumina expression)	10
[illegible]	Glucocorticoid Resistance Gene Profiling in T-ALL Cell Lines	15
[illegible]	Transcription profiling by array of human bone marrow and peripheral blood CD34+ myeloblasts in acute myeloid leukemia	10
[illegible]	Transcription profiling of acute monocytic leukemia (AML-M5) patients with DNMT3A mutations	10

Figure 1. The BioSamples Database in 2012.



Examples: E-MEXP-31, cancer, p53, Geuvadis

ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content

Updated today at 16:30

- 36066 experiments
- 1042850 assays
- 9.35 TB of archived data

Latest News

Jan 2013 - Tools for mapping high-throughput sequencing data topped the list of 'most-read articles' in Bioinformatics last month. Written by N.A. Fonseca, J. Rung, A. Brazma and J.C. Marioni at EMBL-EBI, the paper discusses the results from a survey which focuses on classifying

Figure 2. The ArrayExpress Database in 2012.

Proteins and protein families

UniProt, the unified resource of protein sequence and functional information, is closely integrated with Ensembl and Ensembl Genomes and in 2012 generated new reference proteome sets to match their genes in the reference genomes.

As the genome data collections broaden their taxonomic range, UniProt continues to refine its automatic annotation pipelines, as well as improving its tools for manual annotation.

User-experience design has been a recurring theme for our protein resources: the UniProt Development and InterPro teams spent significant time in 2012 creating user-experience-driven interfaces.

After 15 years of overseeing Protein sequence resources at EMBL-EBI, Rolf Apweiler welcomed Alex Bateman as his successor in late 2012. Alex and his team bring with them a portfolio of important resources, including Pfam, which will reside under the EMBL-EBI umbrella in 2013.

UniProt

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

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

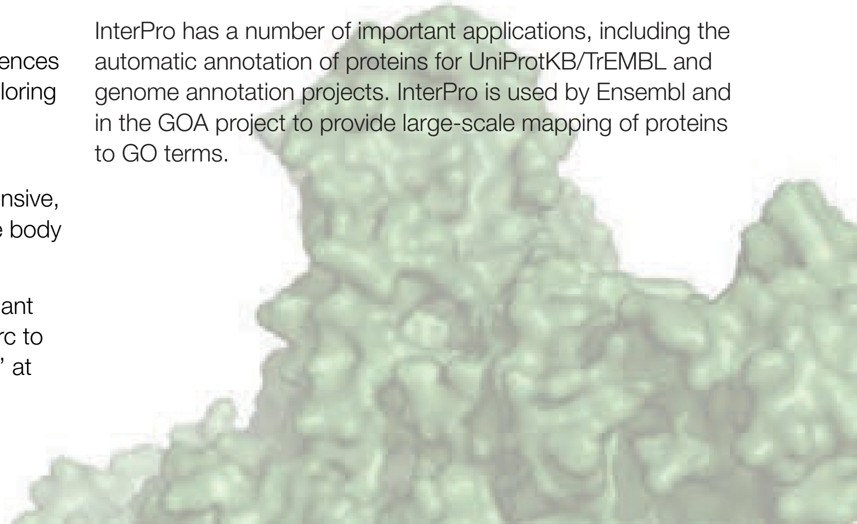
- The UniProt Knowledgebase (UniProtKB) is the central database of protein sequences and provides accurate, consistent and rich annotation about sequence and function.
- The UniProt Metagenomic and Environmental Sequences (UniMES) database serves researchers who are exploring the rapidly expanding area of metagenomics, which encompasses both health and environmental data.
- The UniProt Archive (UniParc) is a stable, comprehensive, non-redundant collection representing the complete body of publicly available protein sequence data.
- UniProt Reference Clusters (UniRef) are non-redundant data collections that draw on UniProtKB and UniParc to provide complete coverage of the 'sequence space' at multiple resolutions.

InterPro

InterPro is used to classify proteins into families and predict the presence of domains and functionally important sites. The project integrates signatures from 11 major protein signature databases: Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH-Gene3D, PANTHER and HAMAP. During the integration process, InterPro rationalises instances where more than one protein signature describes the same protein family or domain, uniting these into single InterPro entries and noting relationships between them where applicable.

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

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



Rolf Apweiler

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

Joint Associate Director since 2012.

Summary of progress 2012

Sarah Hunter

InterPro

- Issued five major releases of the InterPro database: created 1756 new entries and integrated 2355 signatures;
- Released a new version of InterProScan;
- Redesigned and re-launched the InterPro website;
- Incorporated a new InterPro search facility based on the central EBI search engine.

Claire O'Donovan

UniProt content

- Continued to manually annotate UniProtKB, with a particular focus on the human and other reference proteomes;
- Collaborated closely with other resources worldwide to ensure comprehensiveness, avoiding duplication of effort and achieving mutually beneficial exchange of data;
- Substantially progressed automatic annotation efforts, achieving a widening of the taxonomic and annotation depth as well as establishing more collaborations with external informatic- and laboratory-oriented groups;
- Increased manual and electronic GO annotation efforts: as of November 2012 there were 127 million GO annotations for 18.9 million UniProtKB entries, covering more than 370 000 taxonomic groups.

Maria-Jesus Martin

UniProt development

- Analysed different interface designs for accessing UniProt data, focusing on user interaction with the website;
- Integrated new species as Reference proteomes in collaboration with Ensembl and Ensembl Genomes to achieve consensus sequence annotation;
- Improved annotation tools (UniRule, Gene Ontology, proteome editors) to support curation of these resources;
- In collaboration with Ensembl, RefSeq and PRIDE, extended the data-import infrastructure to incorporate variation and proteomics data;
- Consolidated software and extended the databases to accommodate a rapidly growing volume of data.

UniProt content

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

UniProt manual curation

The curation methods we apply to UniProtKB/Swiss-Prot include manual extraction and structuring of experimental information from the literature, manual verification of results from computational analyses, quality assessment, integration of large-scale datasets and continuous updating as new information becomes available.

UniProt automatic annotation

UniProt has developed two complementary approaches in order to automatically annotate protein sequences with a high degree of accuracy. UniRule is a collection of manually curated annotation rules, which define annotations that can be propagated based on specific conditions. The Statistical Automatic Annotation System (SAAS) is an automatic, decision-tree-based, rule-generating system. The central components of these approaches are rules based on the manually curated data in UniProtKB/Swiss-Prot from the experimental literature and InterPro classification.

UniProt GO annotation (GOA)

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

Major achievements

As a core contributor to the Consensus CDS (CCDS) project, UniProt now has 18 854 manually curated human entries, of the total 20 248 records, in synch with the RefSeq annotation group (National Center for Biotechnology Information, NCBI) and the Ensembl and HAVANA teams (EMBL-EBI and the Wellcome Trust Sanger Institute). A component of this effort involves ensuring a curated and complete synchronisation with the HUGO Gene Nomenclature Committee (HGNC), which has assigned unique gene symbols and names to more than 33 500 human loci (over 19 000 of these are listed as coding for proteins).

We play a major role in establishing minimal standards for genome annotation across the taxonomic range, largely thanks to collaborations arising from the annual NCBI Genome Annotation Workshops, which are attended by researchers from life science organisations worldwide. These standards have contributed significantly to the annotation of complete genomes and proteomes and are helping scientists exploit these data to their full potential.

Our team is working on a ‘gold standard’ dataset to help users identify all experimental data for a given protein from a particular strain of a given organism, as well as all experimentally characterised annotations/proteomes from a proteome or protein family. This work is undertaken in collaboration with model organism databases and the Evidence Code Ontology (ECO).

Our curators are key members of the GO Consortium Reference Genomes Initiative for the human proteome and provide high-quality annotations for human proteins. In 2012, the electronic GO annotation pipeline was reviewed and improved, with particular focus on the UniPathway, Ensembl and InterPro collaborations.

Claire O'Donovan

BSc (Hons) in Biochemistry, University College Cork, 1992. Diploma in Computer Science University College Cork, 1993.

At EMBL since 1993, at EMBL-EBI since 1994. Team Leader since 2009.

Future plans

In 2013 we will continue work on a gold-standard dataset across the taxonomic range to fully address the requirements of the biochemical community. We will also continue to expand and refine our Ensembl and Genome Reference Consortium collaborations to ensure that UniProtKB provides the most appropriate gene-centric view of protein space, allowing a cleaner and more logical mapping of gene and genomic resources to UniProtKB. We also plan to extend our nomenclature collaborations to include higher-level organisms. We will prioritise the extraction of experimental data from the literature and extend our use of data-mining methods to identify scientific literature of particular interest with regard to our annotation priorities. We are committed to expanding UniRule by adding feature annotation and extending the number and range of rules with additional curator resources, both internal and external.

Selected publications

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

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

Alcantara, R., et al. (2013) The EBI enzyme portal. *Nucleic Acids Res* 41 (Database issue), D773-D780.

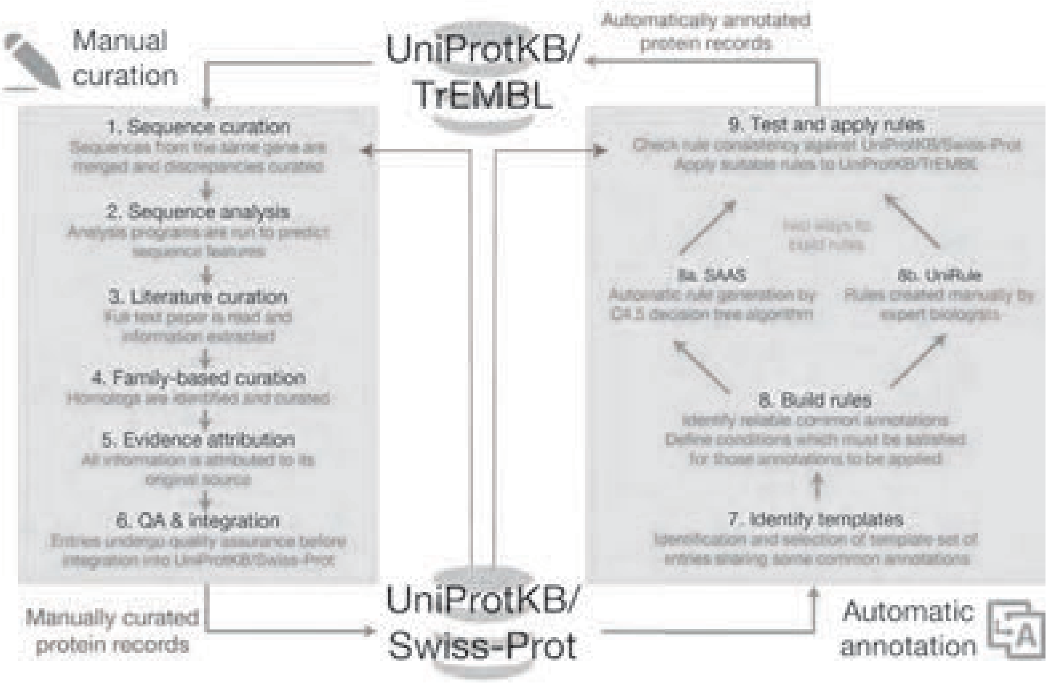
Pedruzzi, I., et al. (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res* 41 (Database issue), D584-D589.

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

Eberhardt, R.Y., et al. (2012) AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database* (Oxford) 2012, bas003.

Burge, S., et al. (2012) Biocurators and Biocuration: surveying the 21st century challenges. *Database* (Oxford) 2012, bas059.

Gaudet, P., et al. (2012) Recent advances in biocuration: Meeting Report from the fifth international Biocuration Conference. *Database* (Oxford) 2012, bas036.



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

Figure. Organisation of data and information in the Universal Protein Resource. www.uniprot.org/help/biocuration

UniProt development

Our team provides the bioinformatics infrastructure for UniProt databases and services. We are also responsible for the development of tools for UniProt curation and the study of novel, automatic methods for protein annotation.

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

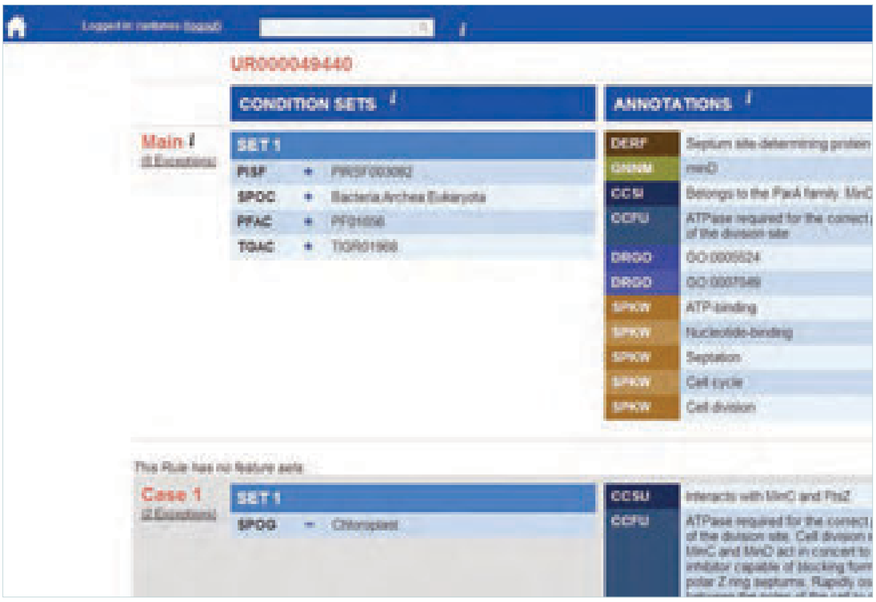
Major achievements

The UniProt website facilitates the search, identification and analysis of gene products. The team has been implementing new web interfaces as a result of the user feedback gathered in a number of website reviews. Consecutive interactions with users have allowed us to refine these interfaces, and the implementation of a much-improved website has been initiated in collaboration with our colleagues at the Swiss Institute of Bioinformatics (SIB). User feedback analysis has highlighted new requirements in sequence variation and protein expression, and we are addressing these needs.

In an attempt to integrate user-community annotation efforts into UniProt, the team developed a procedure to analyse the relevance and quality of Wikipedia articles describing human genes and their corresponding gene products. In collaboration with wiki projects, including GeneWiki, we developed a procedure to map UniProtKB records to relevant Wikipedia articles and established a pipeline for cross-referencing these resources. Other user engagement activities include the use of social media (primarily Twitter and Facebook) to share UniProt news with a broader community.

In order to help users to access a wide range of completely sequenced genomes, we increased the number of reference proteome datasets. We worked with our user community to provide new reference proteomes and to maintain well-annotated model organisms and others of interest for biomedical and biotechnological research. The extension of the data import pipelines to Ensembl and Ensembl Genomes organisms has allowed us to complete the proteome sets and include new species with no coding sequence annotations in the INSDC nucleotide databases.

Our team organised and distributed 557 reference proteomes. All of these datasets are now associated with their corresponding genome assembly. New species produced in 2012 include *Pan troglodytes* (chimpanzee), *Pongo abelii* (Sumatran orangutan), *Cavia porcellus* (Guinea pig), *Takifugu rubripes* (Japanese pufferfish), *Apis mellifera* (honeybee),



UR000049440

CONDITION SETS i

ANNOTATIONS i

Main i
(4 Exceptions)

SET 1

PISF	+	PIRSF003092
SPOC	+	Bacteria,Archea,Eukaryota
PFAC	+	PF01656
TGAC	+	TIGR01968

DERF	Septum site-determining protein
GNNM	minD
CCSI	Belongs to the ParA family. MinC
CCFU	ATPase required for the correct [illegible] of the division site
DRGO	GO:0005524
DRGO	GO:0007049
SPKW	ATP-binding
SPKW	Nucleotide-binding
SPKW	Septation
SPKW	Cell cycle
SPKW	Cell division

This Rule has no feature sets

Case 1
(2 Exceptions)

SET 1

SPOG	-	Chloroplast

CCSU	Interacts with MinC and FtsZ
CCFU	ATPase required for the correct [illegible] of the division site. Cell division i[illegible] MinC and MinD act in concert to [illegible] inhibitor capable of blocking form[illegible] polar Z-ring septums. Rapidly os[illegible]

Figure. The UniRule annotation tool.

Maria-Jesus Martin

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

At EMBL-EBI since 1996. Team Leader since 2009.

Bombyx mori (silk moth), *Gibberella zeae* (wheat head blight fungus), *Solanum lycopersicum* (tomato), *Glycine max* (soybean), and *Phytophthora ramorum* (sudden oak death agent), amongst many others.

Collaborations with Ensembl and Ensembl Genomes have allowed us to create data links between DNA sequences and the functional proteins they encode. We have developed a pipeline to map protein sequences in UniProt to their corresponding transcripts and genes for most of the species in Ensembl. This is key to facilitating further integration of genome-related data from the 1000 Genomes Project. We are developing a pipeline to incorporate variation data from Ensembl sources and to annotate functional variants in Ensembl.

In 2012 we implemented UniRule, a system for automatic annotation of a large volume of uncharacterised proteins; it now contains most of the annotation rules produced by the three consortium members. We also extended the UniRule curation tool to include the annotation of sequence-related features. This makes it easier for curators to manage prediction rules and perform statistical assessment of existing and new rules. UniRule annotates over 6 million sequences in UniProtKB/TrEMBL.

In 2012 the team also worked on a new user interface for QuickGO, the UniProt GO browser. The initial user-testing phase is complete and we expect the new interface to be released by Spring 2013. We develop the web-based Protein2GO tool, which UniProt curators use to contribute annotations to the GOA project. We extended this tool in 2012 to include new functionalities as requested by the GO Consortium curators.

Future plans

In 2013 we will serve the rule-annotation tool as the central mechanism for rule predictions for all UniProt curators in the Consortium. We will provide a UniRule XML format for rule-exchange within UniProt and explore this format for public distribution and data-exchange mechanisms with user communities interested in functional prediction. We will implement a UniProt website with a new look and feel and improved functionality that will facilitate easy access to reference and complete proteomes. We will continue to focus on usability issues and engage with our users to ensure we maintain a global genome/proteome- and gene-product-centric view of the sequence space, and to make it easier for our users to explore in-depth the variations and annotations for each specific protein within our resources. We will continue to co-operate with diverse data providers (e.g., Ensembl, RefSeq, PRIDE) to integrate relevant genome and proteome information.

Selected publications

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40 (Database issue), D71-D75.

Quest for Orthologs Consortium. Towards community standards in the quest for orthologs (2012) *Bioinformatics* 28, 900-904.

Ruth Y.E., et al. (2012) AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database*, bas003.

Gaudet, P., et al. (2012) Recent advances in biocuration: Meeting Report from the fifth International Biocuration Conference. *Database*, bas036.

The UniProt Consortium. (2012) UniProtKB amid the turmoil of plant proteomics research. Front. *Plant Sci*; doi: 10.3389/fpls.2012.00270.

The UniProt Consortium. (2012) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res* 41 (Database issue), D584-D589.

Velankar, S., et al. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 41(D1), D483-D489.

Salazar, G., et al. (2012) MyDas, an Extensible Java DAS Server. *PLoS One* 7.9, e44180.

Alcántara R, et al. (2013) The EBI Enzyme Portal. *Nucleic Acids Res* 41 (D1), D773-D780.

InterPro

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

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

Metagenomics is the study of the sum of genetic material found in an environmental sample or host species, typically using next-generation sequencing (NGS) technology. The Metagenomics Portal, a resource established at EMBL-EBI in 2011, enables metagenomics researchers to submit sequence data and associated descriptive metadata to the public nucleotide archives. Deposited data is subsequently functionally analysed using an InterPro-based pipeline, and the results generated are visualised via a web interface.

Major achievements

We redesigned and re-launched the InterPro website in late 2012, and played a key role in the EMBL-EBI website redesign process. We also built a new InterPro search facility that utilises the central EBI search engine. Search results are now much easier to interpret and browse: the engine behaves in a Google-like manner, allowing users to enter wildcards (e.g., * and ?), use logic (AND or NOT), search with single words or phrases and quickly select subsets of the results using faceted filtering. InterPro results are now paginated and highlight the context of the query terms.

The new EMBL-EBI website, which will launch in early 2013, features improved discoverability of InterPro and other resources. Global EBI search results are shown in categories on local search pages to encourage users to explore the data in different ways.

In 2012 we moved the InterPro DAS and BioMart services to the London Data Centres; the main InterPro website will join them there shortly.

The InterPro database continues to benefit from improved coverage of UniProtKB proteins, increasing to 80.8% in the latest release (v. 40.0). This is partly due to significant data curation and integration efforts, which led to an additional 2355 signatures being incorporated into the database in 2012.

Focussed curation of InterPro2GO term associations led to 334 additional entries being assigned GO terms; 44% of entries now have at least one term associated. The total number of GO mappings has increased by 838, despite a concerted effort to remove terms that are too general (and therefore uninformative) or erroneously mapped. In 2012 we published the first paper describing how this highly utilised annotation resource is created and maintained.

InterProScan5 is poised to take over as the main InterPro scanning software in 2013. Multiple release candidates were made publicly available in 2012, each containing new features and improved implementation.

InterPro Scan 5: release candidate 4 features

- Search all 11 member databases, plus four additional algorithms: Phobius, TMHMM, Coils and SignalPv4;
- Predict potential membership of a protein in a pathway based on InterPro results;
- Use a BerkeleyDB-based protein match look-up service that reduces calculation overheads by only searching sequences not already found in UniProtKB (install this locally or query the EBI-hosted service);

Sarah Hunter

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

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

- Use multiple output formats: HTML, GFF3, XML, TSV and SVG;
- Run it 'out of the box' on any Linux machine with minimal configuration, and utilise cluster-queuing technologies;
- Handle both protein and nucleotide sequences, with results mapped back to the original sequence.

EBI Metagenomics reached 20 public metagenomics projects in 2012, comprising 131 separate samples and a significant number of privately held studies. In collaboration with the European Nucleotide Archive, we developed a system for the submission of sequence files and minimum-standards-compliant metadata. We expanded the initial analysis pipeline from quality control, clustering, CDS prediction and functional classification steps to include an rRNA prediction step (using rRNAselector) and taxonomic diversity estimation, using the Qiime software. We are investigating Taverna for the structuring and managing the complex workflows used in the analysis pipeline (see Figure) and in 2012 developed a utility to integrate Taverna processes with the LSF queue system.

Our work on the organisation and display of data on the website has made it easier for users to access analysis results. In addition, we developed a metagenomics 'GO slim' (a subset of GO terms particularly useful to metagenomics) to assist users in their interpretation of function prediction results. The data can be downloaded in a variety of formats, and we have made it possible to download sequences that are functionally classified by the resource or remain of unknown function.

Future plans

To facilitate the move of the InterPro website to the London Data Centres in early 2013, we have re-written the InterPro relational database into a data warehouse structure. This simplifies the web application code written to access the data, and greatly reduces the amount of down-time experienced by our curation team during release. Together with the official release of InterProScan5, we expect these developments to simplify our data-production processes. InterProScan5 will be used by the EBI-hosted installation, completing the five-year effort to re-architecture the InterPro resource.

We are designing and testing new EBI Metagenomics webpages that will help users visualise taxonomic prediction data from a variety of experiment types (i.e., shotgun metagenomics, amplicon-based marker gene analysis, metatranscriptomics). We believe these changes, to be implemented in 2013, will provide a more complete suite of analysis tools, bringing us in line with competing resources. We will transition our pipeline fully into the Taverna software, simplifying maintenance and offering multiple workflows, depending on the environment that has been sequenced. Finally, we will encourage data submission to the repository to increase the coverage of the experiments carried out by the metagenomics community.

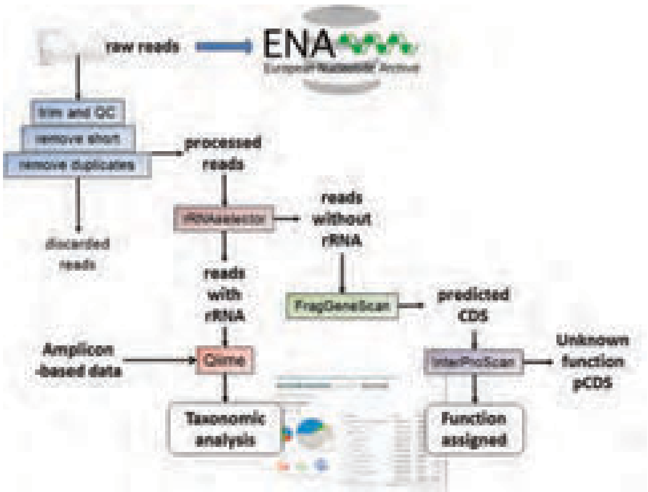


Figure. The analysis workflow for a shotgun metagenomics experiment, as processed by EBI Metagenomics.

Selected publications

Burge, S., et al. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database* (Oxford) 2012, bar068.

Lewis, T.E., et al. (2012) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res* 41 (D1), D499-507.

Salazar, G.A., et al. (2012) MyDas, an Extensible Java DAS Server. *PLoS One* 7, e44180.

Hunter, C., et al. (2012) Metagenomic analysis: the challenge of the data bonanza. *Brief Bioinform* 13, 743-746.

Molecular and cellular structure

Understanding molecular structure is key to understanding function. PDBe, Europe's arm of the Worldwide Protein Data Bank collaboration (wwPDB), made significant improvements in 2012 to the ways in which it delivers structural information.

Special attention was paid to integrating structural data with other types of information, including sequence data, through the SIFTS service, a collaboration with the UniProt team.

Other important developments were the enhanced analysis of NMR entries and many improvements to EMDB, the European resource for electron microscopy-based models. EMDB now has a new search service and an interactive viewer for electron tomograms.

Outreach, training and communications activities for structural biology at EMBL-EBI received a boost in 2012 with the appointment of Gary Battle as the dedicated PDBe Outreach Co-ordinator.

Protein Data Bank in Europe

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

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

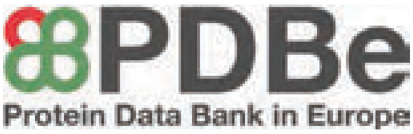
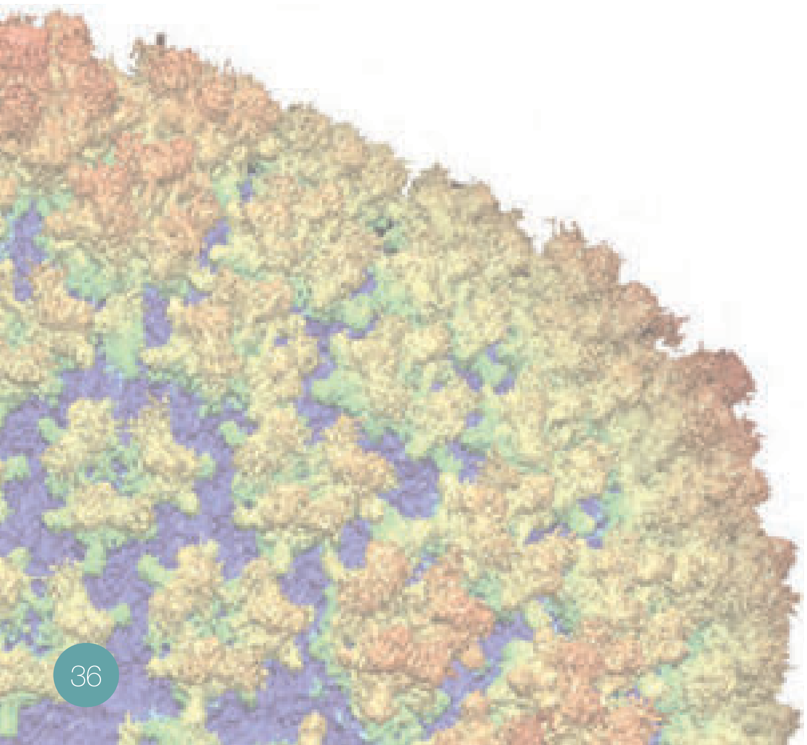


Figure. In 2012 PDBe designed a new logo to be introduced in 2013. Our aim is for the new logo to become synonymous with high-quality information about 3D molecular and cellular structure. The four-fold symmetry of the logo evokes associations with 3D structure at a variety of scales, from electrons (the symmetry of certain d-orbitals), to atoms (e.g., co-ordination around a metal ion) to molecular complexes and assemblies (e.g., tetramers of voltage-gated ion channels) and finally to clusters of dividing cells.



Gerard Kleywegt

PhD University of Utrecht, 1991. Postdoctoral researcher, then independent investigator, University of Uppsala, 1992-2009. Co-ordinator, then Programme Director of the Swedish Structural Biology Network, 1996-2009. Research Fellow of the Royal Swedish Academy of Sciences, 2002-2006. Professor of Structural Molecular Biology, University of Uppsala, 2009.

At EMBL-EBI since 2009.

Summary of progress 2012

Gerard Kleywegt

PDBe

- Enhanced visualisation, analysis and validation features on EMDB entry pages, including an interactive viewer for electron tomograms;
- Enhanced analysis and validation features for NMR entries in the PDB;
- Launched EMsearch: a flexible EMDB search service;
- Released a new module for PDBeXplore that allows exploration and analysis of proteins in the PDB archive based on GO classification;
- In collaboration with our wwPDB and EMDataBank partners, released remediated versions of the PDB and EMDB archives. The EMDB ftp archive was merged with the PDB ftp archive and PDBj began to process EMDB depositions;
- Expanded our outreach, training and communications activities and appointed a dedicated Outreach Co-ordinator.

Tom Oldfield

PDBe databases and services

- Moved the main PDBe resources to the London infrastructure;
- Enhanced PDBe web resources, particularly the quaternary structure analysis server (PDBePISA) and the structure alignment server (PDBeFold);
- Launched PDBeStatus and enhanced existing status services;
- Continued to develop an integrated wwPDB deposition and annotation system;
- Developed new infrastructure for future PDBe systems.

Sameer Valenkar

PDBe content and integration

- Processed 1676 depositions to the PDB and 253 depositions to EMDB, including the 1000th EMDB entry annotated at EMBL-EBI;
- Improved the representation of structure data in response to feedback from the structural biology community, and established a wwPDB working group to improve format specifications;
- Enhanced the data integration (SIFTS) infrastructure, making it easier to track changes to cross-references to PDB data;
- Improved data content and representation in several PDBe services;
- Implemented a new X-ray structure validation pipeline for use by all wwPDB sites;
- Published six new interactive 'Quips' blog posts about topical structures in the archive.

Protein Data Bank in Europe

The major goal of PDBe is to provide integrated structural data resources that evolve with the needs of biologists. Our team handles deposition and annotation of structural data expertly; provides an integrated resource of high-quality macromolecular structures and related data and maintains in-house expertise in X-ray crystallography, Nuclear Magnetic Resonance spectroscopy and 3D Electron Microscopy. Our focus areas are: advanced services, ligands, integration, validation and experimental data.

Major achievements

In 2012 PDBe annotation staff curated a record number of entries, both in absolute numbers (1676 PDB entries, up 12% from 2011) and in relative terms (17% of all deposited PDB entries worldwide, up from 16% in 2011). In addition, 253 EMDB entries were deposited at PDBe (up 56% from 2011, and representing 58% of worldwide depositions).

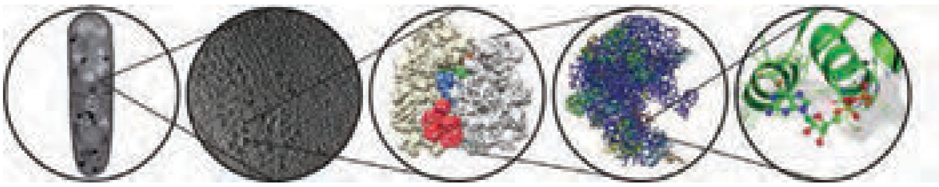
PDBe and its partners in the US and Japan are developing a common tool for handling the deposition and annotation of structural data on biomacromolecules—obtained using any technique or combination of techniques—by all wwPDB and EMDataBank partners. Our team is a major contributor to system design and software development in this project and is responsible for the workflow manager, the validation modules and the deposition interface. The new tool will be rolled out during 2013.

The wwPDB and EMDataBank partners continually review the quality and integrity of the PDB and EMDB archives. In 2012 work began on PDB remediation releases for 2013 and 2014. In the latter effort, PDBe is responsible for improving the description and annotation of protein modifications.

PDBe's Electron Microscopy staff worked with the EMDataBank partners and the EM community to improve and extend the EMDB data model, and introduced EMsearch, a flexible EMDB search service. We enhanced facilities on the EMDB entry pages for analysis, visualisation and validation of EM data, and in collaboration with the OME team in Dundee, UK launched an interactive viewer that allows experts and non-experts alike to study the contents of electron tomograms.

PDBe's NMR staff began to develop a validation pipeline for NMR structures and data, which will become part of the common wwPDB deposition and annotation system. We also enhanced the Vivaldi NMR analysis, visualisation and validation tool for PDBe.

The PDBe team organises outreach and training activities and participates in EMBL-EBI Bioinformatics Roadshows. We



Fiigure 1. This series shows the thinking behind PDBe's new logo, which symbolises our ambition to bring structure to biology by archiving, enriching and disseminating 3D structural data at many scales. These span (pictured from right to left): the atomic details of interactions between drugs and other ligands with biomacromolecules, the overall folds and structural details of individual molecules and domains, the shape and organisation of large complexes and machines and, finally, the localisation of such complexes and machines in the context of the living cell. Atomistic data in the PDB and volumetric data in EMDB cover most of these scales. Together with community experts, PDBe is exploring the archival needs and opportunities of 3D cellular structure data produced by emergent techniques such as electron and soft X-ray tomography and 3D scanning-electron microscopy.

Gerard Kleywegt

PhD University of Utrecht, 1991. Postdoctoral researcher, then independent investigator, University of Uppsala, 1992-2009. Co-ordinator, then Programme Director of the Swedish Structural Biology Network, 1996-2009. Research Fellow of the Royal Swedish Academy of Sciences, 2002-2006. Professor of Structural Molecular Biology, University of Uppsala, 2009.

At EMBL-EBI since 2009.

appointed a dedicated PDBe Outreach Co-ordinator in 2012. We organised an expert workshop on emerging structural biology techniques, such as electron tomography, soft X-ray tomography and 3D scanning-electron microscopy in order to assess archiving needs and opportunities. In response to a recommendation of this workshop, PDBe will seek funding and start pilot archives for these new types of data. We also co-organised the EMBO course on Computational Structural Biology. We delivered 23 lectures and 17 training/outreach presentations, and contributed to more than 20 publications. We published six Quips (interactive structure tutorials) and released a timely and informative article about Kobilka and Lefkowitz's Nobel Prize-winning structures.

Future plans

Our goal is to make PDBe the logical first stop on any quest for information about 3D molecular and cellular structure. To transform the structural archives into a truly useful resource for biomedical and related disciplines, we will continue to focus our developments on: advanced services (e.g., PDBePISA, PDBeFold, PDBeMotif and the new PDB browsers); annotation, validation and visualisation of ligand data; integration with other data resources; validation and presentation of information about the quality and reliability of structural data; and exposing experimental data in ways that help all users understand the extent to which the data support the structural models and inferences. In 2013 the new joint wwPDB deposition and annotation system will be tested and released. Much of the groundwork for a complete redesign of the PDBe website and search system as well as the PDB and EMDB entry pages has been carried out in 2012, and the new website will be launched in 2013. This event will also mark the start of a publicity campaign to make non-expert users of structures aware of the wealth of 3D structural information available from PDBe.

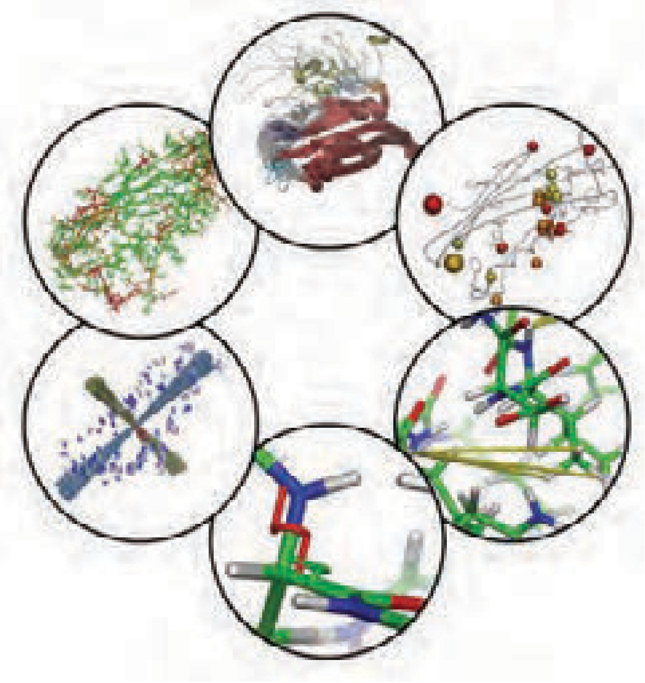


Figure 2. Visualisation and analysis of experimental data and validation information for NMR-derived entries in the PDB, using protein PA1076 from P. aeruginosa as an example (PDB entry 2k4v). Clockwise from top: NMR structure ensemble, coloured by domain; atoms with unusual chemical shift values shown as coloured spheres; violations of distance constraints shown as yellow lines between the corresponding hydrogen atoms; visualisation of experimental torsion-angle constraints and violations; visualisation of residual dipolar coupling data; representative model, with residues coloured according to the Red/Orange/Green quality score assigned by the NRG-CING resource.

Selected references

Velankar, S., et al. (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 40, D519-D524.

Gore, S., et al. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr* D68, 478-483.

Patwardhan, A., et al. (2012) Data management challenges in three-dimensional EM. *Nat Struct Mol Biol* 19, 1203-1207.

Berman, H.M., et al. (2012) The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure* 20, 391-396.

PDBe content and integration

Our goal is to ensure that PDBe truly serves the needs of the biomedical community. We are constantly improving PDBe's web interface and designing new tools to make structural data available to all.

In the context of the SIFTS project, we integrate structural data with other biological data to facilitate discovery. These integrated data form the basis for many query interfaces that allow biomacromolecular structure data to be presented in its biological context. Our specific focus areas are: data integrity, data quality, integration and data dissemination to the non-expert biomedical community.

Major achievements

In 2012 PDBe annotation staff curated a record number of PDB entries (1676 entries, up 12% from 2011), the majority of which were annotated within one working day of being deposited. PDBe staff also annotated 253 EMDB entries - 58% of all EMDB entries deposited in 2012. One of these entries was the 1000th EMDB entry to be processed at EMBL-EBI since the EMDB archive was established here in 2002.

PDBe staff, in collaboration with other wwPDB members, carried out a remediation project to improve the representation of the biologically important peptide-like antibiotic and inhibitor molecules in the PDB. The remediation efforts addressed issues related to consistent representation and functional annotation of these molecules. A collection of almost 600 reference dictionary files, containing detailed chemical and functional description of antibiotic and inhibitor molecules, was made available through the PDB ftp area.

We remediated EMDB data to improve the quality and consistency of the archive. Our team addressed issues related to taxonomy information, author and citation information, map-symmetry records and consistent representation of all EMDB map files.

Validation information about newly deposited structures (as well as those already in the public archive) is a critical element of usability. PDBe implemented an initial version of the validation-software pipeline as part of the new Deposition and Annotation system, developed in collaboration with the other wwPDB partners.

Since 2000, PDBe has provided valuable integration services via SIFTS (Structure Integration with Function, Taxonomy and Sequences), which is maintained in close collaboration with the UniProt team. SIFTS supports several PDBe tools and services and is used by many other major bioinformatics resources. We improved the SIFTS infrastructure and the coverage of mapping information, upgrading the versioning and adding new data on GO and InterPro assignments.

Our team undertook the redesign of the PDBe website and search infrastructure, which is scheduled to launch in 2013. We worked with many users in both academic and industry sectors to better understand how they use and navigate structure-related information; our findings formed the basis of the redesign. We are also revamping the search infrastructure for PDB and EMDB data and developing an API for accessing structure data.

Future plans

We strive to make biomacromolecular structure data available in useful ways to a broad biomedical community. We continually improve the quality and integrity of the data in our care, and endeavor to integrate them with other sources of biomedical information.

We will work with our wwPDB partners to simplify the quality assessment of crystal structure entries in the PDB by progressing an X-ray validation pipeline, which was developed at PDBe. We will also develop similar pipelines for NMR and 3DEM data, again making quality assessment easier for non-expert users. We will design a more intuitive interface to display validation information for all kinds of 3D biomacromolecular structure data, and will launch a completely redesigned web site in 2013.

A new wwPDB Deposition and Annotation system will be released in 2013, and we expect this to result in a major change in annotation practices. We will work closely with the other wwPDB partners to ensure an efficient transition.

Sameer Velankar

PhD, Indian Institute of Science, 1997. Post-doctoral researcher, Oxford University, UK, 1997-2000.

At EMBL-EBI since 2000. Team leader since 2011.

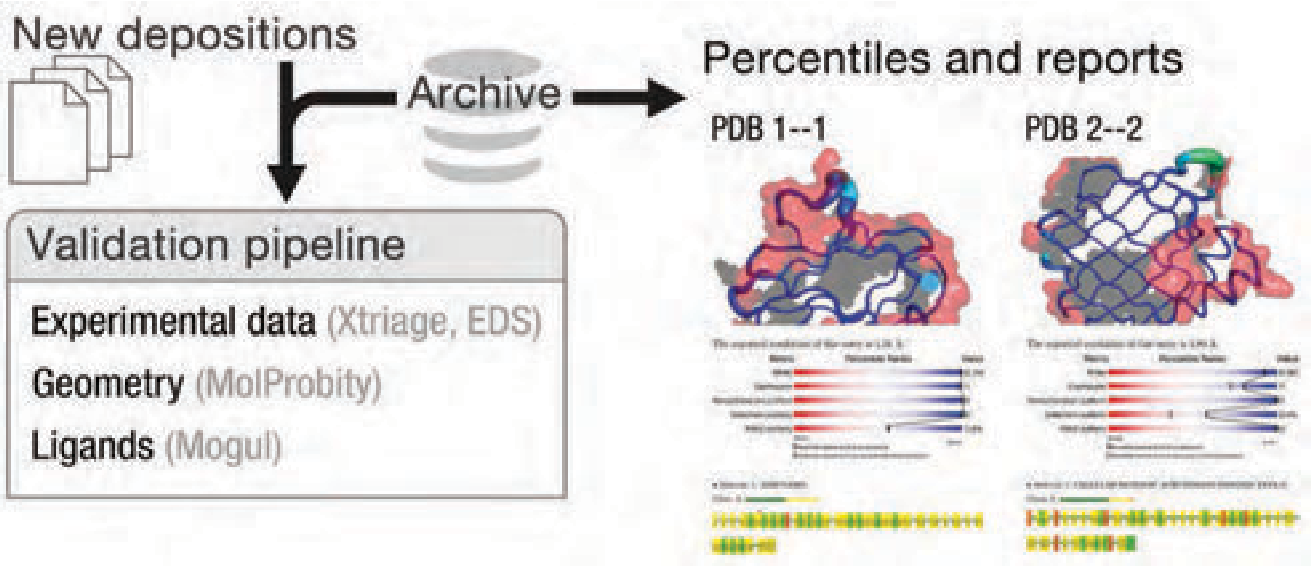


Figure. The wwPDB X-ray validation pipeline assesses the quality of experimental data, the atomic model and the fit of the model to the data, using community-standard software. The pipeline will be run for all structures already in the PDB archive and also for all newly deposited structures. It produces high-quality statistics, graphs, plots and a summary report to help expert and non-expert users assess the global and local quality of individual PDB entries so they can decide if it is suitable for their needs. The validation results can be used to compare and select the most suitable model of a particular biomacromolecule in the entire PDB.

Selected references

Velankar, S., et al. (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 40, D445-D452.

Gore, S., Velankar, S. and Kleywegt, G.J. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr* D68, 478-483.

PDBe databases and services

Our team manages two production systems: the weekly update of deposited data and the weekly increment of newly released data. These data systems are managed within multiple Oracle databases and support a large number of integrated web resources to collect data and disseminate information to the wider life-science community.

Major achievements

During 2012 the PDBe databases and services team migrated 27 web services to the London infrastructure. This required the weekly synchronisation of five terabytes of data with the London infrastructure and the addition of six production database servers to the PDB release cycle. The system had to be automated and capable of validating each stage. We created remote management systems for the databases and web services make it easier to review the status of the PDBe infrastructure in London.

Our team made significant contributions to the co-ordinated development of the new wwPDB deposition and annotation system, most of which is in place and ready for internal testing. During 2012 we developed a user interface for deposition that is integrated with the workflow system, updated the workflow manager for annotators and improved the underlying workflow system by implementing self-diagnostics and automated recovery. We managed a unified installation at all the wwPDB partner sites so that we can manage all changes simply—even in third-party software.

PDBeStatus launched in 2012, allowing users to search for PDB entries based on their status. For example, it can give a quick overview of all X-ray crystal structures removed from the archive in 2005, or show information about newly deposited entries that have not been released.

We improved data quality and dissemination of several PDBe services in 2012, including PDBeFold and PDBePISA. We modified these services to read mmCIF files and, in response to user feedback, are improving the quality of value-added data. We retired the EuroCarbDB website and infrastructure and replaced it with a portal web page, making it easier to locate the many carbohydrate services provided by former participants in the EuroCarbDB project.

In 2012 we specified and designed a number of new infrastructure components in the context of a new PDBe website, which will launch in early 2013. These components include hardware servers, Lucene search systems, an API to collate data and a unified search system with faceted output.

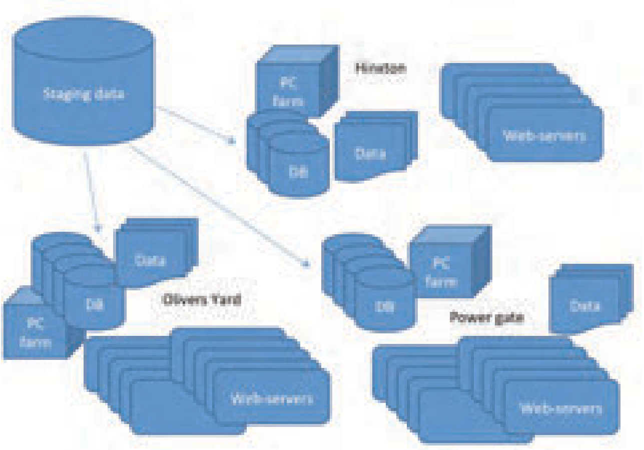


Figure 1. PDBe maintains three copies of fail-over production databases that are re-created at 00:00 every Wednesday. This involves rebuilding databases, running hundreds of processes and moving 5 TB of data in a window of 38 hours in time for the weekly release. The data and remote database servers are managed with automated systems.

Tom Oldfield

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

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

Future plans

A number of PDBe resources await the move to London: PDBeMotif, a key resource for searching the PDB at the local structure level; the EM-related infrastructure and the NMR-related infrastructure. Procedures to support the migration are in place, and the process will be completed in the first half of 2013.

The new wwPDB deposition and annotation system will go into production in 2013, after many rounds of testing. Because depositions will be carried out with far more data checking before submission, wwPDB and PDBe technical staff will need to invest considerable effort in ensuring that the final improvements to the code base are all working properly.

The new PDBe website will be released in 2013. We will make every effort to create an integrated system that allows all the underlying PDBe infrastructure to appear as a single, unified resource, easily accessible for casual and novice users while providing expert knowledge to advanced users.

Selected publication

Velankar, S., et al. (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 40, D445-D452.

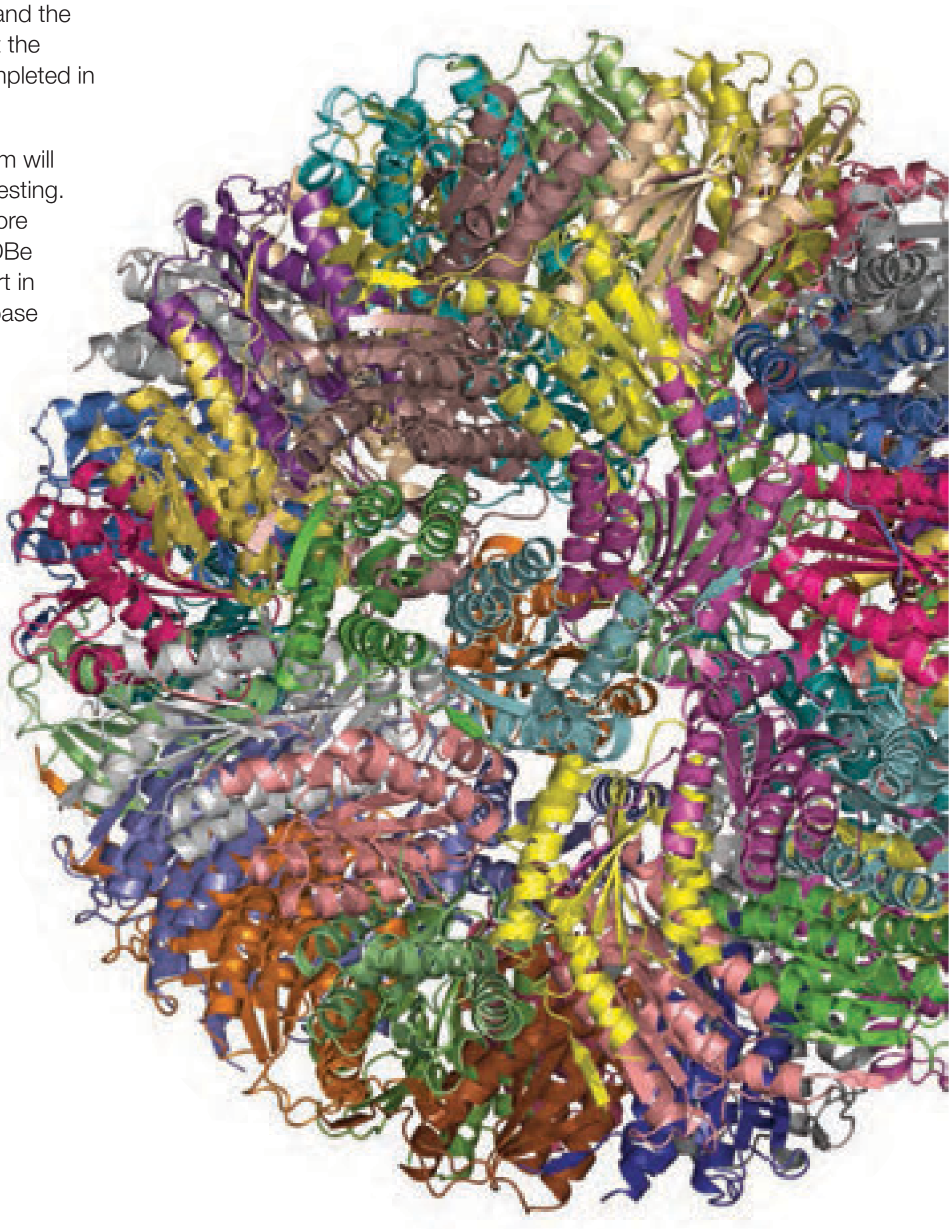


Figure 2. The ever-larger and increasingly complex structures deposited in the PDB require robust data-handling and archiving systems to manage the weekly release of approximately 200 new entries.

Molecular systems

The genes and gene products encoded by genomes do not act in isolation but do so in co-ordinated systems, often containing protein, small molecule and oligonucleotide or oligosaccharide components.The Molecular Systems Cluster groups together resources that cover systems biology: from enzymes and their mechanisms, through protein—protein interactions and networks, to pathways and approaches to quantitatively model entire complex biological systems.

The Enzyme Portal, launched in February 2012, draws together enzyme-related data from ten different databases. It is the first of EMBL-EBI's resources to be entirely user-led, from conception to interface design (Pavelin et al., 2012). The Enzyme Portal allows easy and reliable searching across many enzyme resources, including two developed and maintained directly in the Steinbeck group: IntEnz and Rhea.

In 2012 IntAct, our database of molecular interactions, topped 300 000 pieces of evidence supporting interactions between molecules. The IntAct team also led international efforts to co-ordinate the curation of molecular interaction data (Orchard et al., 2012).

Following the departure of Nicolas Le Novère to The Babraham Institute, Cambridge, the leadership of BioModels at EMBL-EBI is now in Henning Hermjakob's hands. The contents of this resource grew by two orders of magnitude in 2012, owing to the integration of a large set of models from the Path2Models project. This represents a new, 'top-down' approach to building quantitative models, starting with pathways from data resources.

Enzyme Portal

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

http://ebi.ac.uk/enzymeportal/

Reactome

Reactome is an open-source, open-access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff, and cross-referenced to many bioinformatics databases.

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

IntAct

IntAct provides a freely available, open-source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions.

http://ebi.ac.uk/intact/

BioModels

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

http://ebi.ac.uk/biomodels/

John Overington

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

At EMBL-EBI since 2008.

Summary of progress 2012

Henning Hermjakob

Biomodels, IntAct & Reactome

- Published the IMEx consortium strategy for internationally co-ordinated curation of molecular interaction data (Orchard et al, 2012);
- Reached 300 000 binary interaction evidences in the IntAct molecular interaction database (Kerrien et al, 2012);
- Reached the milestone of 6000 reactions in the Reactome pathway resource;
- Integrated the Path2Models data, increasing the data content of BioModels Database by two orders of magnitude to more than 140 000.

Christoph Steinbeck

Enzyme Portal, IntEnz & Rhea

- Developed and launched the Enzyme Portal;
- Issued releases 28 to 36 of the Rhea enzyme resource;
- Issued releases 75 to 83 of the IntEnz enzyme resource;
- Redesigned the Enzyme Portal, IntEnz and Rhea to fit with the new EMBL-EBI website, scheduled to launch in early 2013.

Proteomics services

The Proteomics Services team develops tools and resources for the representation, deposition, distribution and analysis of proteomics and systems biology data.

We follow an open-source, open-data approach: all of the resources we develop are freely available. The team is a major contributor to community standards, in particular the Proteomics Standards Initiative (PSI) of the international Human Proteome Organisation (HUPO) and systems biology standards (COMBINE Network). We provide public databases as reference implementations for community standards: the PRIDE proteomics identifications database, the IntAct molecular interaction database, the Reactome pathway database and BioModels Database, a repository of computational models of biological systems.

As a result of long-term engagement with the community, journal editors and funding organisations, data deposition in our standards-compliant data resources is becoming a strongly recommended part of the publishing process. This has resulted in a rapid increase in the data content of our resources. Our curation teams ensure consistency and appropriate annotation of all data, whether from direct depositions or literature curation, to provide the community with high-quality reference datasets.

We also contribute to the development of data integration technologies, using protocols like Distributed Annotation System (DAS) and semantic web technologies, and provide stable identifiers for biomolecular entities through identifiers.org.

Major achievements

A major success in 2012 was achieving full production mode for the ProteomeXchange consortium. In this EU-funded consortium, PRIDE works with a number of international partners (e.g., PeptideAtlas, UniProt, University of Ghent, University of Liverpool, ETH Zurich, University of Michigan, Wiley-VCH) to co-ordinate data deposition and dissemination strategies for mass spectrometry data, providing a single entry point for data deposition, a shared accession number space and a deposition metadata format. In spring 2012, ProteomeXchange started full production and achieved good user acceptance, reaching 77 submissions with more than 20 million spectra by November 2012 (see Figure).

ProteomeXchange submission strategies include the capability to deposit large raw datasets, and in 2012 we completely redeveloped PRIDE data deposition support, publishing Java libraries (Griss et al, 2012; Reisinger et al, 2012), an updated PRIDE Converter submission tool (Côté et al, 2012), and the PRIDE Inspector data access tool (Wang et al, 2012).

IMEx is an international consortium of major molecular interaction data providers that globally synchronise their data deposition and curation efforts (Orchard et al, 2012). IMEx partners share formats, identifier spaces and curation strategies and directly share the web-based IntAct curation infrastructure (Kerrien et al., 2012). This avoids redundant development while retaining the value of each individual resource. IMEx partners include UniProt (Switzerland and the United Kingdom), I2D (Canada), InnateDB (Ireland and Canada), Molecular Connections (India) and MechanoBio (Singapore).

Reactome provides review-style, curated and peer-reviewed human pathways in a computationally accessible form. In 2012 we collaborated with the Ouwehand group to provide a detailed update of platelet-related pathways (Jupe et al, 2012), as well as many other high-profile pathways. We curate disease variants of normal physiological pathways to provide users with information about mutations and how they affect pathways. We also focus on cancer-related signalling pathways.

Our BioModels team contributed to a large-scale collaborative effort for the semi-automated generation of systems biology models based on KEGG pathways, Biocarta, MetaCyc and SABIO-RK data. More than 142 000 models, covering 1852 species, were generated, opening up new opportunities

Henning Hermjakob

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

At EMBL-EBI since 1997.

to explore and refine models. We rose to the challenge of integrating these models into BioModels Database in 2012, increasing the number of available models by two orders of magnitude.

Future plans

Following the upgrade of the PRIDE submission system, we have turned our attention to redeveloping the core database and web interface. This is essential for maintaining good response times and helping users access increasingly large and complex proteomics datasets. We will begin providing quality-controlled subsets and derived datasets from PRIDE, evolving this resource from a primary database into a systems biology source of protein expression data.

IntAct will increasingly provide confidence-scored interaction datasets derived from integration of individual publications. The same strategies will be applied, where possible, to interaction data from multiple sources, which we access through the PSICQUIC interface. This will ensure we can provide integrated, up-to-date interaction datasets. We will redesign the IntAct website in 2013 to improve accessibility.

In 2013 we will make IntAct datasets accessible through Reactome. This will enhance Reactome's visualisation of molecular interactions in the context of molecular pathways. The next Reactome website release will include a new interactive pathway viewer, and will feature improved overlay of external information such as expression data. Reactome curation will focus on disease-induced modifications of pathways, supported by improved visualisation tools and close collaboration with disease-oriented research communities.

We will develop the new storage infrastructure for the BioModels Database so that the resource can cope with the influx of computationally generated models. We will develop new features (e.g., full model versioning, support for more modelling formats), update the user interface, enhance search facilities and improve overall performance.

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

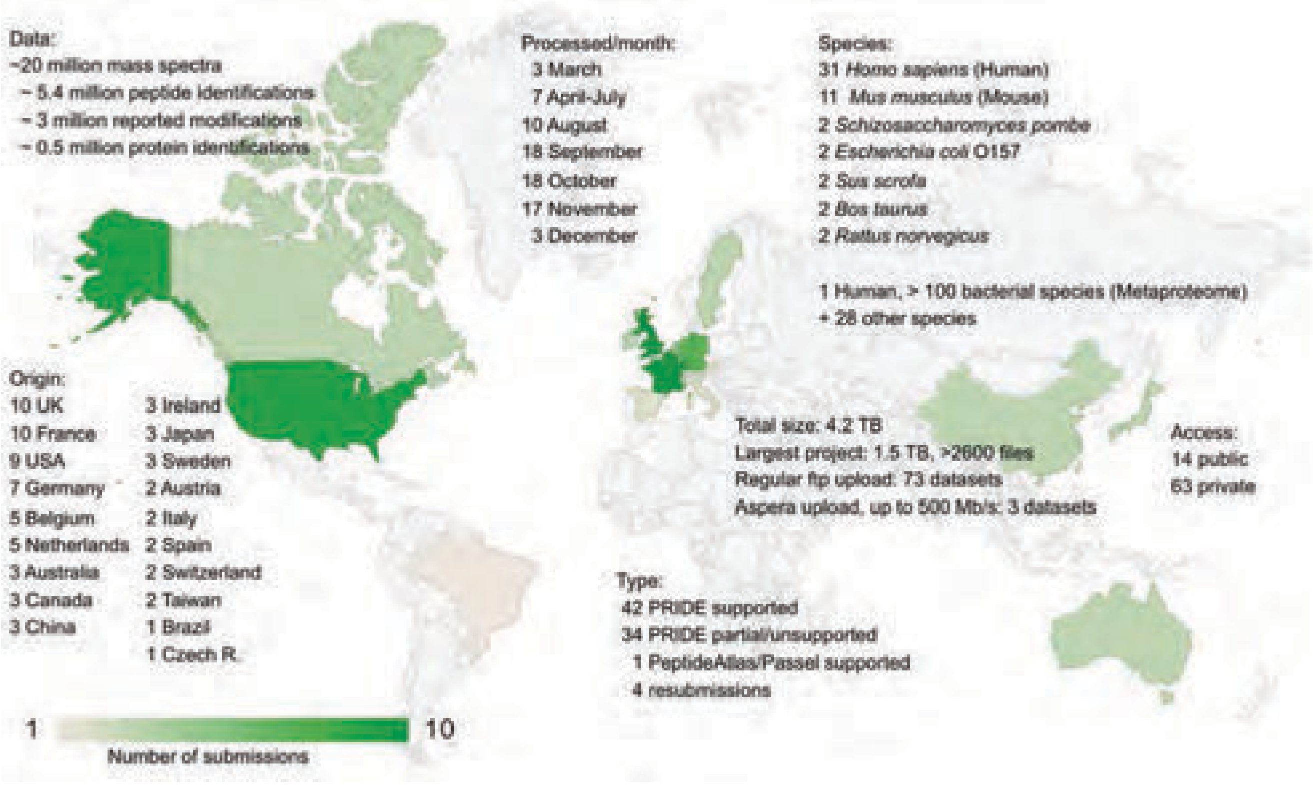


Figure. Distribution of ProteomeXchange submissions in 2012

Chemical biology

The importance of small molecules in life-science informatics has never been greater. The ability to probe, modulate and control biomolecules with chemicals has huge economic and healthcare impacts. EMBL-EBI has a number of teams focussed on the informatics aspects of chemical biology—specifically small molecules and their roles and effects on biological systems.

In 2012 ChEMBL team extended its reach to the neglected disease community, providing a single access point to all data from the open-access MalariaBox from the Medicines for Malaria Ventures, joining other new, high-value malaria and tuberculosis datasets. In order to provide timely, accurate cross references across all EMBL-EBI chemistry data resources, the team also developed the UniChem web service

Meanwhile, the Cheminformatics and Metabolism team turned its attention to natural products in ChEBI, developing a stand-alone tool for classifying compounds that resemble natural products as well as a new version of the SENECA tool, which helps to elucidate natural product structures.

ChEMBL

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

ChEBI

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

John Overington

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

At EMBL-EBI since 2008.

Summary of progress 2012

John Overington

ChEMBL

- Expanded the integration of EMBL-EBI data with partner resource data;
- Engineered robust update and distribution routes for ChEMBL;
- Extended outreach to neglected disease community;
- Extended and optimised REST web services based on user feedback;
- Developed UniChem, a pan-EBI chemistry resource lookup and registration system;
- Included domain-level binding annotation to improve data mining of ChEMBL activity data;
- Developed an open-software ChEMBL virtual machine capable of substructure and similarity searching, increasing the number of researchers with local extensible access to bioactivity data and tools;
- Achieved broad impact, with many medically relevant papers published using ChEMBL data.

Christoph Steinbeck

ChEBI

- Issued 12 releases of ChEBI;
- Developed a stand-alone tool for the classification of ‘natural product likeness’ and integrated it into the Taverna workflow tool;
- Developed libraries for mass spectroscopy data and enumeration of lipid structures;
- Released the first version of Metingear, a tool for genome-scale reconstruction;
- Upgraded the CDK plugin for the KNIME workflow tool and developed additional nodes to process chemical information;
- Started to develop the second version of SENECA, a tool for computer-assisted structural elucidation of natural products;
- Developed a stand-alone web-based tool for ChEBI ontology enrichment that can be used to help interpret the biological context of metabolites by showcasing which ChEBI biological roles are over-represented in a set of chemicals.

ChEMBL

The ChEMBL team develops and manages EMBL-EBI's database of quantitative small molecule bioactivity data focussed in the area of biotherapeutic drug discovery. Although great progress has been made in developing medicines, synthetic small molecule and natural product-derived drugs still form the majority of novel life-saving drugs.

The process, complexity and costs of discovering new drugs has recently risen to the point where public–private partnerships are coming to the fore as a practical, cost-effective solution. Central to the success of this is data sharing and the availability of structure, binding, functional and ADMET data. The ChEMBL database stores curated chemical structures and abstracted quantitative bioactivity data alongside calculated molecular properties. The majority of the ChEMBL data is derived by manual abstraction and curation from the primary scientific literature, and thereby covers a significant fraction of the structure–activity relationship (SAR) data for the discovery of modern drugs. Our associated research interests focus on data-mining ChEMBL for data that can be applied to drug-discovery challenges.

Major achievements

ChEMBL data content continued to expand in 2012, with a 40% increase in the number of compounds and double the number of experimental bioactivities. Deeper annotation of ChEMBL, as well as partnership and sharing of best practices with key laboratories, have improved data interoperability of chemical and bioactivity data.

Usage of ChEMBL's web interface grew approximately three-fold in 2012, with development of the interface reflecting feedback from users. Downloads of the entire database and integration with other systems continue to be strong, and a number of public services now rely on ChEMBL. We continued our successful webinar series,which was complemented by a series of training courses, both on-site in Hinxton, Cambridge, UK and at host institutes throughout the world.

Our new UniChem service allows us to integrate and query across other EMBL-EBI and global chemistry resources on the basis of shared InChI structures. At launch it contained over 26 million structures and currently allows immediate cross-linking of new data to stable resources. This separation of integration registration has sharpened focus on curation and quality assessment of the core ChEMBL data.

One of the most important applications of the data contained within ChEMBL is in the assessment and scoring of genes and proteins as targets for drug discovery. We built an infrastructure for consistent large-scale scoring of targets for their potential 'druggability' alongside our well-established structure-based scoring of binding-site data.

Although the majority of drug discovery focuses on 'small molecule' structures, there is an increasing interest in biological drugs such as monoclonal antibodies, solubilized receptors and replacement enzymes. We expanded our infrastructure to include these, and assembled an extensive set of curated, clinical-stage monoclonal antibodies, which are generally beyond the scope of resources such as UniProt.

Data query privacy is extremely important, especially as more commercial organisations rely on public services. We published our data privacy policy in 2012, and within UniChem built in the ability to securely partition data into public and private volumes, aiding analysis of results prior to publication for key datasets.

To expand the data captured within ChEMBL we developed cloud-based data entry portals, for example for distributed data entry for non-proprietary names and structures for drugs. We also developed infrastructure for in-line curation of ChEMBL using controlled vocabularies, adopted ontologies and enhanced structure entry—including from images.

We participate in eTox, an IMI project to predict toxicity by building an unprecedented collaborative database of compound-related toxicity data relevant to drug development and performing analyses and software development. We are a key partner in: the EU-funded diXa chemical safety data infrastructure; the ELIXIR BioMedBridges project and the EU-OpenScreen research infrastructure, helping researchers identify compounds affecting new targets by integrating high-throughput screening platforms, chemical libraries, bio- and

John Overington

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

At EMBL-EBI since 2008.

chem-informatics support, screening results, assay protocols and chemical information. We are also a partner in SMSDrug.net, a networking project to catalyse collaboration at the synthetic chemistry–drug discovery interface within the United Kingdom.

The group's social media outreach continued to expand, with the ChEMBL-og blog and @ChEMBL Twitter feeds gathering a large and diverse following.

Research

We apply ChEMBL data to important drug-discovery challenges. Our projects are typically built on data-mining ChEMBL and working with collaborators on integration with other data classes or disease/patient datasets.

We analyse the design of peptide-derived drugs, which fall between the properties of Lipinski-compliant small molecules and biological drugs. In 2012 we extracted and analysed a library of all ChEMBL alpha-amino acids within peptide ligands, developed QSAR approaches to suggest property optimisation strategies and placed this set of amino acids within the context of all possible (and reasonable) alpha-amino acids. We established that the physicochemistry affinity properties of peptide ligands are distinct from synthetic small molecule ligands.

We developed an approach to annotate the likely binding domain for a ligand within a complex multi-domain protein, greatly reducing the false-positive rate for sequence search matches for contaminating common spectator domains. We also started the reconstruction of assay cascades from the ChEMBL data and began exploring changes in affinity that occur as a molecular-to-phenotypic scale is traversed.

We explored the differential expression of all known, pharmacologically responsive genes in mouse development, from pre-birth to natural old age. We observed changes in the expression patterns of this gene set that could point towards differential efficacy and safety in paediatric and geriatric human patients.

In studying the incorporation of resistance mechanisms within drug design, we applied integrated sequence, structural modelling, molecular dynamics simulations and QSAR models to identify less mutable, functional parts of binding pockets within a number of antiviral target systems. We found physicochemical and structural properties of allosteric regulators that may offer opportunities to tackle classically difficult drug targets.

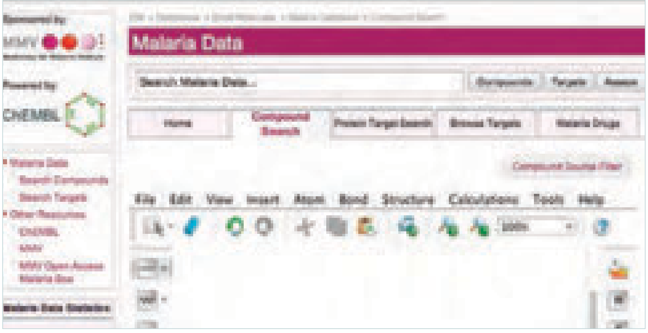


Figure: Medicines for Malaria Venture data are now in ChEMBL.

Future plans

In 2013 we will focus on translational and safety biology and build our community around open data for neglected diseases. We will increase the throughput and effectiveness of target and bioassay annotation in ChEMBL by including key ontologies (e.g., the BioAssay Ontology), fostering data-sharing partnerships and leading international efforts to standardise the curation of bioactivity data. We plan to complete linkage to targets, deepen the data model to deal with post-translational modifications and macromolecular assemblies, and integrate population-variation data. Finally, we will enhance the delivery of ChEMBL via web front-end, REST web services with an RDF triple-store representation.

Selected publications

Gaulton, A., et al. (2012) ChEMBL: A large-scale bioactivity database for chemical biology and drug discovery. *Nucleic Acids Res* 40 (D1), D1100-1107.

Hersey, A., Senger, S., Overington, J.P. (2012) Open data for drug discovery – learning from the biology community. *Future Med Chem* 5, 1865-1867.

Krüger, F.A., Rostom, R., Overington, J.P. (2012) Mapping small molecule binding data to structural domains. *BMC Bioinform* 13, S11.

Chambers, J., et al. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform*, doi:10.1186/1758-2946-5-3.

Cheminformatics and metabolism

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

We develop methods to decipher, organise and publish the small molecule metabolic content of organisms. We also develop algorithms to process chemical information; predict metabolomes based on genomic and other information; determine the structure of metabolites by stochastic screening of large candidate spaces and enable the identification of molecules with desired properties. This requires algorithms based on machine learning and other statistical methods for the prediction of spectroscopic and other physicochemical properties represented in chemical graphs.

Major achievements

In 2012 we focused on developing and launching two new resources: the Enzyme Portal and MetaboLights, a general-purpose, open-access repository for metabolomics data. We released a beta version early in the year and the production version in June 2012. We are now working on the MetaboLights Reference Layer, which will be launched in 2013.

The Enzyme Portal unifies access to all enzyme-related information at EMBL-EBI. It removes artificial boundaries for viewing and using enzyme-related information, combining reputable resources to characterise different types of data. Drawing on the usability features and back-end technology of the EBI Search, it can be queried using enzyme, gene, compound name or reaction ID and presents the results in an organised fashion.

The ChEBI team continued to curate high-quality small molecule data and improve the web interface. ChEBI entries now display the introductory paragraph of and link to related Wikipedia articles when available. Conversely, ChEBI entities have been added to relevant Wikipedia pages. The number of unique visitors to the ChEBI website every month has increased by approximately 50% from 2011, and the resource had more than 430 000 visits – a 45% increase over 2011 (these figures exclude traffic generated by robots). ChEBI now offers novel ontology visualisation and the team invested considerable effort in improving the classification of natural products and carbohydrates in the ontology.

In October 2012, the European Co-ordination of Standards in MetabOlomicS (COSMOS) consortium, comprising 14 partners and co-ordinated by Dr Steinbeck and Dr Salek, officially started its work on metabolomics data standardisation, publication and dissemination workflows. The MetaboLights database is a key component in this effort.

Future plans

We will continue to develop and stabilise the Enzyme Portal and the MetaboLights experimental repository. We will work on the MetaboLights Reference Layer, a comprehensive knowledgebase organised around a metabolite-centric view that features reference spectra, biological reference data, protocols, cross-references to other resources and advanced search and download functionality. We will populate it with comprehensive, manually curated data including chemical structures and characteristics from ChEBI, metabolic pathways, reference spectroscopy and chromatography. MetaboLights will interface with the BioSamples Database and the Expression Atlas to offer accessible information about the reference biology, metabolites and their occurrence and concentration in species, organs, tissues and cellular compartments in healthy and diseased conditions. We will also make publication references and protocols available, helping experimentalists to gain a comprehensive view on known metabolites. We will work closely with the

Christoph Steinbeck

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

At EMBL-EBI since 2008.

metabolomics community on data exchange formats and mechanisms.

We will adapt Enzyme Portal tools to enable searching by protein sequence and chemical structure of compounds related to enzymes.

Our team will develop curation and data-submission components that can be re-used across EMBL-EBI's cheminformatics databases. As part of our efforts to develop metabolomics and metabolism resources, we will continue to work on the first release of a natural product collection in ChEBI that will feature information about approximately 5000 natural products, including their structure and detailed biological source.

COSMOS will work to develop policies to ensure that metabolomics data is encoded in open standards, tagged with a community-agreed and complete set of metadata, supported by a communally developed set of open-source data management and capturing tools, disseminated in open-access databases adhering to these standards, supported by vendors and publishers who require deposition upon publication and properly interfaced with data in other biomedical and life science e-infrastructures.

Selected publications

Steinbeck, C., et al. (2012) MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics* 8, 757-760.

Alcántara, R., et al. (2012) The EBI Enzyme Portal. *Nucleic Acids Res*, doi: 10.1093/nar/gks1112.

Jayaseelan, K.V., et al. (2012) Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* 13, 106.

Pavelin, K., et al. (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comp Biol* 8, e1002554.

Figure 1. Metabolights, launched in June 2012.

MetaboLights
ASPARTIC ACID - MTBLC22660
An alpha-amino acid that consists of succinic
a single alpha-amino substit...
MetaboLights Download
Submit a new study
Update an existing study

Figure 2: Enzyme Portal, launched in February 2012.

Cross-domain tools and resources

Scientific literature is central to the research landscape and is increasingly linked to the underlying data. At EMBL-EBI, we use the literature and related resources as a force for integration to empower researchers to navigate, search and retrieve scientific information.

Several activities in 2012 have contributed towards this integrative goal. The Europe PubMed Central brand was launched in November 2012, reflecting the inclusion of the European Research Council as its 19th funding organisation. Having one brand covering both abstracts and full text articles provides clarity and focus for future development regarding the literature.

A new interface, API and submission-accessioning service have been developed for the EBI BioSamples database. This database provides organisational infrastructure for data that arise from the same biological samples. Many datasets often derive from the same sample, so it is important that all those datasets, in whichever public database they are deposited, point to the same metadata.

In the area of ontology development and application, the Experimental Factor Ontology (EFO), originally developed to categorise gene expression datasets, was extended in 2012 to support annotation of genome-wide association studies and the integration of genomic and disease data.

Europe PubMed Central

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

EBI BioSamples Database

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

Gene Ontology

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

Experimental Factor Ontology

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

Rolf Apweiler

PhD 1994, University of Heidelberg.

At EMBL since 1987, at EMBL-EBI since 1994.
Associate Director since 2012.

Summary of progress 2012

Johanna McEntrye

Literature services

- Re-launched UKPMC as Europe PubMed Central in November 2012;
- Operated Europe PubMed Central with 99.7% uptime and met speed performance targets, with page loads taking 1-2 seconds on average;
- Oversaw a significant increase in Europe PubMed Central services: the number of IP addresses accessing the site increased in 2012 from about 25 000 to about 60 000.

Gene Ontology (GO)

- Expanded the functionality of GO's direct ontology submission tool, TermGenie, which now includes a 'free form' input for experienced users;
- Completed integration of GO and CHEBI ontologies;
- Implemented a continuous integration platform, harnessing powerful OWL reasoners, for improved quality assurance over the ontologies;
- Expanded the logical definitions of GO terms: over 35% of terms now have axioms associated with them, including those referring to external ontologies such as CHEBI.

Helen Parkinson

Experimental Factor Ontology (EFO) and EBI BioSamples

- Issued a major release of the EFO that supports the annotation of the NHGRI GWAS catalogue and includes the Orphanet genetic disease classification for OMIM and Ensembl integration;
- Released a new, dynamic, karyotype-based browser for the GWAS catalogue using the EFO;
- The BioSamples team implemented a new user interface using EFO for searches, delivered improved tabular layouts for sample description, and provided a new BioSamples API;
- The BioSamples database now includes more than a million samples from EBI assay databases and 70 000 from reference collections such as Coriell Cell Lines.

Literature services

Scientific literature is an essential component of the scientific data infrastructure and often represents both the start and end point of a scientific project.

As the volume of scientific articles published continues to grow, it becomes increasingly important to manage the scientific literature effectively and find ways to optimise it in order to support scientific discovery.

There are three main components to this endeavour that EMBL-EBI Literature Services is addressing. First, there is the basic service requirement to deliver the ability to search, display and download abstracts and full-text articles in a consistent manner, and make these data available for different levels of use. Second, given this core resource, there is the challenge of integrating it with biomolecular data resources to create effective data navigation for the research community. Third, it is essential to address wider community needs by aligning the core resource with other related literature-based efforts, such as author disambiguation and funding-source attribution.

This year, EMBL-EBI Literature Services has worked in collaboration with 19 funders of life science research across Europe, lead by the Wellcome Trust, to consolidate existing literature databases (CiteXplore and UK PubMed Central) into Europe PubMed Central (http://europepmc.org). Europe PubMed Central contains over 26 million abstracts and includes PubMed as well as data from Agricola and patents from the European Patent Office. The full text article component consists of about 2.4 million full-text articles, of which about 500 000 are open access.

Layered on top of the abstracts and articles are several features that add value to the content and ibegin to address the integration of the literature with other scientific databases hosted by EMBL-EBI:

- We calculate citation network information for the records we hold: over 11 million articles have been cited at least once, representing the largest public-domain citation network in the world.
- We link the literature to databases hosted by EMBL-EBI data resources by using the references appended to database records by curators and submitters. Over 1 million articles are linked to databases by this method.
- We use text-mining techniques to identify terms of interest, such as gene symbols, organisms, diseases, Gene Ontology terms, chemicals and accession numbers, using these to link to appropriate databases. More recently, we have extended the mining of accession numbers from full text articles, feeding this information back to the appropriate databases in order to further enrich those resources with literature-data cross links.

All this information is shared as widely as possible via three different Europe PubMed Central services: interactive use via the website, programmatic access via SOAP web services, and by bulk download (for the Open Access set of articles) by ftp. Supporting the reuse of article content in these ways will help to stimulate the development of new tools and services by the scientific and developer communities.

Figure 1. Europe PubMed Central home page.

Johanna McEntyre

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

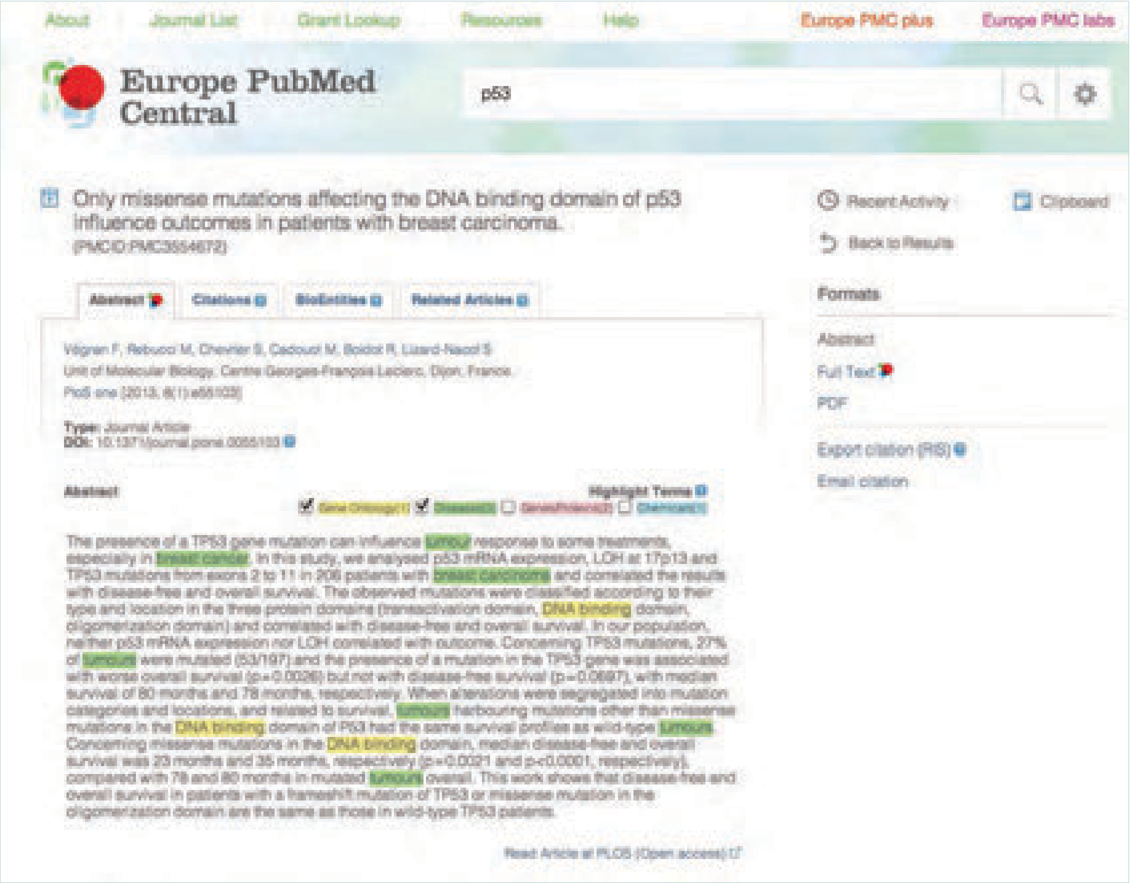
Team Leader at EMBL-EBI since 2009.

Major achievements

The major achievement of Literature Services in 2012 was the consolidation of existing literature resources into the single Europe PubMed Central brand, which supersedes both CiteXplore and UK PubMed Central. Europe PubMed Central is now the single source of literature-based content at EMBL-EBI, running at a professional level of service provision and meeting all performance targets. Developing this website was a significant rebranding exercise in which the University of Manchester and the British Library also played major roles alongside EMBL-EBI. We have made many improvements to the Europe PubMed Central website, including the introduction of a citation count sort order and a more intuitive way to search for journal titles. The most notable development of 2012 was the release of a public web service that allows programmatic access to all of the content hosted by Literature Services.

Future plans

Building on the consolidated resource of Europe PubMed Central, we plan to develop an infrastructure that allows third parties to link related resources to articles in the database. In particular we hope to engage with text mining research groups, allowing them to publish the results of new text-mining algorithms in the context of Europe PubMed Central. In support of this, we will also be redesigning some of the website architecture to make these and other unique features of Europe PubMed Central more prominent. We hope to extend the content base to include the full text of books and reports alongside the research article collection, an activity that was deferred due to the release of Europe PubMed Central. We will also be using text mining to find data citations in articles in order to enrich cross-linking throughout EMBL-EBI resources.



Only missense mutations affecting the DNA binding domain of p53 influence outcomes in patients with breast carcinoma.
(PMCID:PMC3554672)

Végran F, Rebucci M, Chevrier S, Cadouot M, Boidot R, Lizard-Nacol S
Unit of Molecular Biology, Centre Georges-François Leclerc, Dijon, France.
PloS one [2013, 8(1):e55103]

Type: Journal Article
DOI: 10.1371/journal.pone.0055103

Abstract

The presence of a TP53 gene mutation can influence tumour response to some treatments, especially in breast cancer. In this study, we analysed p53 mRNA expression, LOH at 17p13 and TP53 mutations from exons 2 to 11 in 206 patients with breast carcinoma and correlated the results with disease-free and overall survival. The observed mutations were classified according to their type and location in the three protein domains (transactivation domain, DNA binding domain, oligomerization domain) and correlated with disease-free and overall survival. In our population, neither p53 mRNA expression nor LOH correlated with outcome. Concerning TP53 mutations, 27% of tumours were mutated (53/197) and the presence of a mutation in the TP53 gene was associated with worse overall survival (p=0.0026) but not with disease-free survival (p=0.0697), with median survival of 80 months and 78 months, respectively. When alterations were segregated into mutation categories and locations, and related to survival, tumours harbouring mutations other than missense mutations in the DNA binding domain of P53 had the same survival profiles as wild-type tumours. Concerning missense mutations in the DNA binding domain, median disease-free and overall survival was 23 months and 35 months, respectively (p=0.0021 and p<0.0001, respectively), compared with 78 and 80 months in mutated tumours overall. This work shows that disease-free and overall survival in patients with a frameshift mutation of TP53 or missense mutation in the oligomerization domain are the same as those in wild-type TP53 patients.

Figure 2. Europe PubMed Central search results page, illustrating the citation count sort order.

Research

2012 has seen the further transformation of biology towards a big science, involving thousands of laboratories worldwide generating and analysing big data. The complexity of biological systems is being revealed, with many levels of control and interaction. We are only just beginning to translate this information into a deeper understanding of molecular and cellular processes, which will be the basis on which we construct models of whole cells, organs and organisms.

The role of bioinformatics in this process cannot be overstated. At EMBL-EBI our research is embedded in a data-rich environment, surrounded by technical experts who are involved in many of the core international data resources that empower biological research. This unique environment provides the backdrop against which we address some of the most challenging biological problems of our time.

In 2012 we said farewell to three talented and experienced group leaders, who were at, or approaching, the end of their nine-year tenure at EMBL. Nick Luscombe took up a new role as Chair in Computational Biology in the University College London Genetics Institute, and holds a joint appointment as Senior Group Leader at the Cancer Research UK London Research Institute. His lab will join the Francis Crick Institute when it opens in 2015. Nicolas Le Novère is now a Group Leader for the signalling programme at the nearby Babraham Institute, Cambridge, UK where he continues to study neuronal signalling and develop community services to facilitate research in computational systems biology. Dietrich Rebholz-Schuhmann is leading a research team at the University of Zurich's Institute of Computational Linguistics, where he seeks new ways to identify gene–disease associations by text-mining the scientific literature and co-ordinates Mantra, the multilingual terminologies project.

We were pleased to welcome new group leaders who have already brought fresh inspiration to our research programme: Pedro Beltrao, Oliver Stegle and Sarah Teichmann. Pedro joins us from the Krogan and Lim labs at the University of California, San Francisco, and continues to explore the evolution of cellular interactions. Oliver Stegle, formerly of the Max-Planck Institutes in Tübingen, Germany, uses statistical and mechanistic models to interrogate –omics data such as the '1001 genomes' Arabidopis dataset to understand the effect of genetic background and environment on an organism's characteristics and disease. Sarah Teichmann joins us from the MRC Laboratory of Molecular Biology, Cambridge, bringing with her a well-established group. The Teichmann group, which has a wet-lab component at the Wellcome Trust Sanger Institute, maps the evolution of protein families in the context of their functional and structural neighbours.

Regulation

The ENCODE project, often described as a continuation of the Human Genome Project, dominated the news in September. Led by Ewan Birney at EMBL-EBI and funded by the National Human Genome Research Institute in the US, ENCODE comprised over 400 scientists in 32 labs throughout the world who worked in concert to produce a detailed map of genome function that identifies four million gene 'switches'. ENCODE broke the mould for scientific publishing: upwards of 30 papers were published under open-access license in several different journals, with the contents linked by topic and united for optimum exploration in a single interface provided by *Nature*. A virtual machine allows readers to explore the data in context and reproduce the experimental conditions—a notable innovation in the reporting of Materials and Methods. Coverage of the story was extensive, with approximately 13 000 articles on the web, around 450 of which were featured in leading news publications throughout the world.

Nick Luscombe's group, working with the lab of Asifa Akhtar at the Max-Planck Institutes in Freiburg, showed that the transcriptional machinery of male fruit flies works twice as hard on its single chromosome as the female machinery, which has the luxury of two chromosomes (Conrad, T. et al., 2012).

John Marioni's group, working with Duncan Odom at Cancer Research UK and EMBL-EBI colleagues in the Brazma and Flicek groups, combined RNA-sequencing with DNA-sequence data and delivered fascinating insights into the cellular decisions involved in making a specific isoform in different cell types (Goncalves et al., 2012).

Paul Bertone's group, working in mouse embryonic stem cells, discovered an important 'go' signal that tells pluripotent cells when to commit themselves to becoming a specific type of cell. This type of work is fundamental to progress in regenerative medicine.

Evolution

Understanding the evolution of cellular mechanisms is another common research theme at EMBL-EBI. Petra Schwalle in Paul Flicek's research group, working with Duncan Odom's group at the University of Cambridge, developed a new, integrated model of the evolution of the transcription factor CTCF (Schmidt et al., 2012). The model explains the origin of some 5000 highly conserved CTCF binding events in mammals, providing insight into how these binding sites are moved and amplified by 'jumping' repeats called retro-elements.

Christophe Dessimoz in Nick Goldman's group, working with researchers from the Swiss Institute of Bioinformatics (SIB), confirmed a long-held and fundamental assumption in biology that had recently been called into question. The 'ortholog conjecture' posits that studying related genes (orthologs) from other species, even distant ones, can inform us about our own biology. Dessimoz and his colleagues used data from tens of thousands of papers to demonstrate that the ortholog conjecture does indeed hold true.

Nick Luscombe's lab showed that an organism's most valuable assets are protected most carefully. They demonstrated how mutation rates in bacteria vary by more than an order of magnitude, with highly expressed genes having the lowest mutation rates (Martincorena, I. et al., 2012).

In Janet Thornton's group, postdoc Nick Furnham collaborated with Christine Orengo's group at University College London to develop a new resource called FunTree and used it to study 276 enzyme superfamilies. Using FunTree they determined the extent to which enzyme functions have changed over the course of evolution – findings that have important implications for the development of new therapeutics.

Disease

The molecular stratification of diseases that look medically similar is potentially a very powerful application of bioinformatics. Paul Bertone's group identified a gene expression signature in tumourigenic stem cells from glioblastoma multiforme patients that strongly correlated with patient survival (Engstrom et al., 2012).

Anton Enright's group discovered tumour-suppressor-like activity in breast cancer cells in the micro-RNA miR-9 (Selcuklu et al., 2012), whereas the micro-RNA mIR-22 may be protective against cardiac failure (Gurha, P. et al., 2012).

Tools

In pursuing a novel line of enquiry we often find that the tools or databases we need to answer a scientific question do not yet exist. So we build them.

FunTree, described above, is a powerful tool that is potentially useful in the design of new enzymes.

Tools developed by the Marioni and Brazma groups make it possible to quantify how different sequence analysis tools affect gene expression measurements derived from RNA-sequencing data (Fonseca et al., 2012), allowing sequencing groups to identify the best analytic tool for their specific problem.

Julio Saez-Rodriguez's group developed CellNOpt, a new platform for the statistical modelling of signalling networks, and DvD, a pipeline for repurposing medicines using public repositories of gene expression data.

Bertone group

Pluripotency, reprogramming and differentiation

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

Embryonic stem (ES) cells are similar to the transient population of self-renewing cells within the inner cell mass of the pre-implantation blastocyst (epiblast), which are capable of pluripotential differentiation to all specialised cell types comprising the adult organism. These cells undergo continuous self-renewal to produce identical daughter cells, or can develop into specialised progenitors and terminally differentiated cells. Each regenerative or differentiative cell division involves a decision whereby an individual stem cell remains in self-renewal or commits to a particular lineage. The properties of proliferation, differentiation and lineage specialisation are fundamental to cellular diversification and growth patterning during organismal development, as well as the initiation of cellular repair processes throughout life.

The fundamental processes that regulate cell differentiation are not well understood and are likely to be misregulated in cancer. One focus in the lab is the study of neural cancer stem cells derived from human glioblastoma multiforme tumours. Using neural stem (NS) cell derivation protocols, it is possible to expand tumour-initiating, glioblastoma-derived neural stem (GNS) cells continuously in vitro. Although the normal and disease-related counterparts are highly similar in morphology and lineage-marker expression, GNS cells harbour genetic mutations typical of gliomas and give rise to authentic tumours following orthotopic xenotransplantation. We apply genomic technologies to determine transcriptional changes and the chromosomal architecture of patient-derived GNS cell lines and their individual genetic variants. These data provide a unified framework for the genomic analysis of stem cell populations that drive cancer progression, and contribute to the molecular understanding of tumorigenesis.

Summary of progress

- Mapped genome-wide binding sites of key pluripotency regulators and chromatin modifications in mouse embryonic stem cells;
- Determined the transcriptional regulation of midbrain dopaminergic progenitor cell differentiation in early neural development;
- Identified a gene expression signature of tumorigenic stem cells in glioblastoma multiforme strongly correlated with patient survival.

Major achievements

In 2012 we investigated the function of the Nucleosome Remodelling and Deacetylation (NuRD) complex in the control of lineage commitment in ES cells. NuRD is required for proper embryonic development, and NuRD-deficient embryos fail to form primary germ layers in vivo. ES cells lacking NuRD are viable, but are unable to commit to differentiation upon withdrawal of factors that promote self-renewal in culture. We utilised a knockout ES cell line in which Mbd3, a core structural component of the NuRD complex, had been ablated. Using this system we showed that NuRD contributes to considerable transcriptional heterogeneity in ES cells, marked by a loss of regulatory inhibition of pluripotency-associated genes. Further experiments demonstrated a dependence of NuRD function for PRC2 recruitment.

Foxa1 and Foxa2 are important transcription factors in early neural differentiation, associated with the specification of midbrain dopaminergic (mDA) neurons and the formation of the neural floor plate in the developing embryo. Using ChIP-seq we determined the regulatory activity of Foxa2 in ES-cell derived mDA neural progenitors, providing mechanistic insights into Foxa2-mediated regulatory events. We found that Foxa2 acts directly to activate essential determinants of mDA neurons in the floor plate, while negatively regulating transcription factors expressed in the ventrolateral midbrain. We also identified a set of Foxa2-regulated enhancer elements that promote axon trajectories around the midline of the developing embryo.

Paul Bertone

PhD Yale University, 2005.

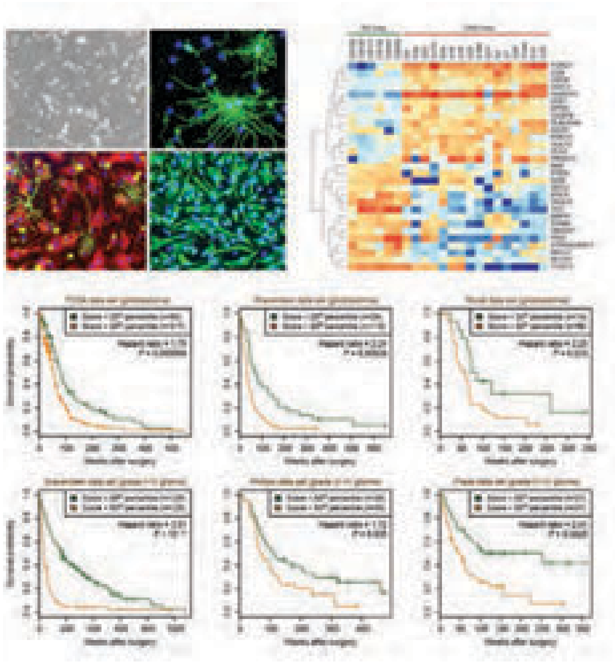
At EMBL–EBI since 2005. Joint appointments in EMBL Genome Biology and Developmental Biology Units. Associate Investigator, Wellcome Trust—Medical Research Council Stem Cell Institute, University of Cambridge.

In the realm of cancer stem cells, we applied high-throughput sequencing methods to characterise the unique transcriptional properties of GNS cell lines derived from individual patient tumours. We found novel oncogene and tumour suppressor candidates that are clearly misregulated in GNS cells, and were able to identify a gene expression signature that defines this population in parental GBM tumours. Our results generalise well across a large panel of independent GNS cell lines and hundreds of tumour biopsies, where a number of novel marker genes expressed in GNS cells correlate with both tumour grade and patient survival. These results indicate that greater populations of compromised stem cells may contribute to poorer clinical outcomes, and that studies of cancer stem cell function are an essential complement to tumour-based research initiatives.

Future plans

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

Tumour-based cancer studies are limited by cellular diversity of tissue biopsies, lack of corresponding reference samples and inherent restriction to static profiling. Cancer stem cells constitute a renewable resource of homogeneous cells that can provide insights leading to therapeutic opportunities. We will analyse our GNS cell bank and use the data to develop methods to stratify glioblastoma classes based on the molecular attributes and differentiation capacities of tumour-initiating stem cells. Tumour subtypes are associated with diverse clinical outcomes, but previous results have suffered from sample heterogeneity. We are refining existing subtype classification to improve the diagnostic utility of this approach and performing functional experiments to identify alterations in GNS cells that impart tumourigenic potential.



Selected publications

Engström P.G., et al. (2012) Digital transcriptome profiling of normal and glioblastoma-derived neural stem cells identifies genes associated with patient survival. *Genome Med* 4, 76.

Metzakopian E., et al. (2012) Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors. *Development* 139, 2625–2634.

Reynolds N., et al. (2012) NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. *Cell Stem Cell* 10, 583–594.

Figure. Malignant brain tumours are driven by abnormal neural stem cells. Top left: GNS cells propagate indefinitely in culture and can differentiate into the major cell types of the central nervous system, such as astrocytes and oligodendrocytes. Top right: Detailed analysis of these cells from independent glioblastoma cases identified a molecular signature strongly correlated with patient survival, such that an increase in primary tumour biopsies is associated with more limited prognoses (bottom).

Birney group

Nucleotide data

DNA sequence remains at the heart of molecular biology and bioinformatics. The Birney research group focuses on sequence algorithms and using intra-species variation to explore elements of basic biology.

The Birney group has a long-standing interest in developing sequencing algorithms. Over the past four years our primary focus has been on compression, with theoretical and now practical implementations of compression techniques. The group's 'blue skies' research includes collaboration with Dr Nick Goldman on a method to store digital data in DNA molecules. We continue to be involved in this area as new opportunities arise—including the application of new sequencing technologies.

We are also interested in the interplay of natural DNA sequence variation with cellular assays and basic biology. Over the past five years there has been a tremendous increase in the use of genome-wide association to study human diseases. However, this approach is very general and need not be restricted to the human disease arena. Association analysis can be applied to nearly any measureable phenotype in a cellular or organismal system where an accessible, outbred population is available. We are pursuing association analysis for a number of both molecular (e.g., RNA expression levels and chromatin levels) and basic biology traits in a number of species where favourable populations are available, including human and *Drosophila*. We are beginning to expand this to a variety of other phenotypes in other species, including establishing the first vertebrate, near-isogenic wild panel in Japanese rice paddy fish, or Medaka (*Oryzias latipes*).

Summary of progress

- Co-ordinated the large-scale integrated analysis of the first scale-up stage of the ENCODE project, culminating in a consortium integrative analysis and the parallel publication of over 30 manuscripts in a novel, open-access publishing platform;
- Developed a robust algorithm to store data digitally in DNA molecules and demonstrated its practical use;
- Conducted large-scale analysis of the patterns of variation at transcription factor binding sites (TFBS) in Drosophila and human, providing evidence for the functional buffering of TFBS mutations.

Major achievements

In 2012 Ewan Birney led the analysis of the Encyclopedia of DNA elements (ENCODE) project data, leading to a collection of high-profile publications in September 2012. The ENCODE project performed a large number of cellular assays, in particular chromatin immuno-precipitation (ChIP) and DNaseI Hypersensitivity Analysis, which probe the chromatin state of genomes. The Birney group was involved in co-ordinating the integrated analysis of ENCODE data and performed a series of specific analyses, for example the integration of different machine learning techniques to classify genome components and generation of combined element collections.

The group was also involved in the curation of the analytical products of the ENCODE project, and in making these available in novel forms alongside more traditional publication, including as interactive figures, cross-manuscript research threads and the ENCODE analysis Virtual Machine (VM). The VM is a novel approach to reporting methods, providing a facility for researchers to install an exact replica of the analytical environment and software used in the preparation of the integrative analysis paper. After delivering an integrated analysis at an unprecedented scale, the group moved on from its involvement in the ENCODE project in September 2012.

The Birney group develops theoretical and practical implementations of DNA compression techniques. The project with Nick Goldman led to a high-profile publication demonstrating a scalable method to reliably store more information in DNA, offering a realistic technology for large-scale, long-term and infrequently accessed digital

Ewan Birney

PhD 2000, Wellcome Trust Sanger Institute.

At EMBL since 2000. Joint Associate Director since 2012.

archiving. This paper was published in 2013 and will be the focus of future reports.

In the area of sequence variation, we pursued a series of analyses of specialised genetic panels. In the realm of the human genome we collaborated with colleagues at the Wellcome Trust Sanger Institute and others to analyse genetic effects on transcription factor binding and transcription in a panel of lymphoblastoid cell lines. In collaboration with the Furlong lab at EMBL, we studied the effect of variation on transcription levels as measured by RNA-seq in developing Drosophila embryos from the inbred lines of the Drosophila Genetic Reference Panel. In Medaka, we collaborated on the development of the first vertebrate near-isogenic wild panel from Japanese rice paddy fish. Initial analysis of the progenitors of this panel is under way, with extensive sequencing of the founding trios.

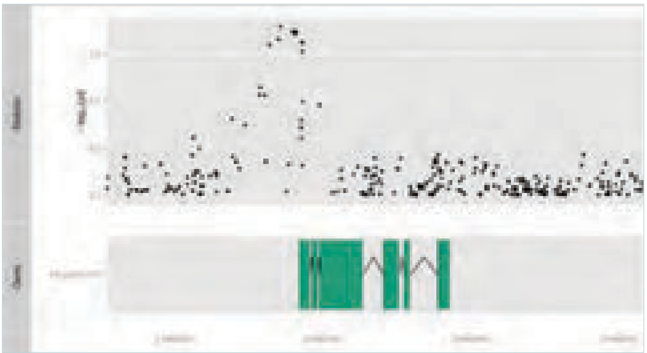


Figure 2. The association of Drosophila SNPs across a 20 kb genomic region with RNA expression levels of the FBgn0031191 gene contained in the region. The level of association is shown as –log10 of the P value on the Y-axis with genomic position on the X axis in the top panel. The genomic structure of the FBgn0031191 gene is shown in the bottom panel, with the direction of transcription from left to right. The most strongly associated SNPs in this case are clustered at the start of the gene, consistent with an effect through modification of promoter efficiency.

Figure 1. The ENCODE papers published in 2012 inspired a new wave of analysis and exploration.

Future plans

The Birney group will continue to work on sequence algorithms and intra species variation. In humans there will be work on molecular phenotypes in an iPSC panel generated as part of the HipSci consortium. In *Drosophila* we will look at multi-time-point developmental biology measures, and we will assess the near isogenic panel in Japanese rice paddy fish.

Selected publications

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74, doi:10.1038/nature11247 (2012).

Goldman, N., et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, doi:10.1038/nature11875 (2013).

Spivakov, M., et al. Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol* 13, R49, doi:10.1186/gb-2012-13-9-r49 (2012).

Enright group

Functional genomics and analysis of small RNA function

Complete genome sequencing projects are generating enormous amounts of data. Although progress has been rapid, the function of a significant proportion of genes remains either poorly characterised or entirely unannotated.

Our group aims to predict and describe the functions of genes, proteins and regulatory RNAs as well as their interactions in living organisms. Regulatory RNAs have recently entered the limelight, as the roles of a number of novel classes of non-coding RNAs have been uncovered. Our work involves the development of algorithms, protocols and datasets for functional genomics.

We focus on determining the functions of regulatory RNAs, including microRNAs, piwiRNAs and long non-coding RNAs. We collaborate extensively with laboratories on commissioning experiments and analysing the data produced. Some laboratory members take advantage of these close collaborations to gain hands-on experience in the wet lab.

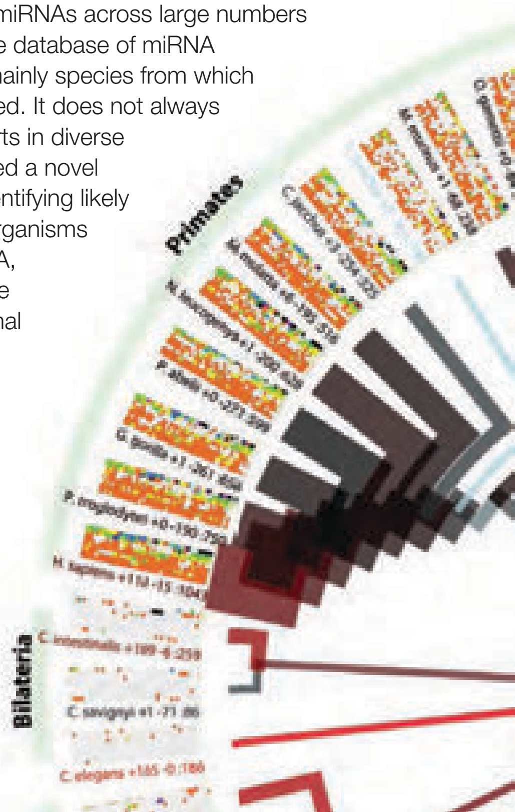
Major achievements

We obtained funding from the BBSRC to embark on a large-scale project to characterise and analyse the function of long non-coding RNAs (lncRNAs) in animal systems. These lncRNAs are widespread molecules whose functions are largely unknown. They have been implicated across a number of areas of genetic and epigenetic regulation. We have been working with the O'Carroll lab in EMBL Monterotondo to characterise and profile these molecules from high-throughput sequencing data in the Mouse germline. Matthew Davis and Harpreet Saini are developing the tools and analysis techniques required for this project. Additionally, we are working with the Furlong laboratory at EMBL Heidelberg to explore the roles of these non-coding molecules in *Drosophila melanogaster*. As part of this funding we are developing and will make available a set of novel pipelines and computational protocols for NGS data analysis. Matthew Davis and Stijn van Dongen have developed the Kraken pipeline, which includes a novel algorithm called REAPER. The system is an extremely fast and lightweight approach written in C that scales to the kind of extremely large datasets that are being more commonly encountered in modern genomics. We have also been working to understand the functions of adult (Pachytene) piwi-associated RNAs. Nenad Bartonicek has been working together with the laboratory of Duncan Odom (University of Cambridge) to assess their function and evolution in animal systems.

Summary of progress

- Obtained BBSRC funding to undertake a large-scale analysis of long non-coding RNAs (lncRNAs) in collaboration with the O'Carroll Lab in EMBL Monterotondo.
- A number of publications on the function and evolution of microRNAs.

We have also been undertaking evolutionary analysis of miRNAs, led by José Afonso Guerra-Assunção. Small, non-coding molecules do not lend themselves well to standard sequence-based phylogenetic approaches to understanding their evolution. Nevertheless, a great deal can be learned about the evolution of small RNA regulation in vertebrates. Our first approach was to identify likely orthologs and paralogs of known miRNAs across large numbers of species. The miRBase database of miRNA sequences comprises mainly species from which a miRNA was first isolated. It does not always search for its counterparts in diverse organisms. We developed a novel mapping strategy for identifying likely miRNA loci in multiple organisms based on a query miRNA, and mapped all miRBase miRNAs across the animal genomes available in Ensembl. We developed a system for large-scale exploration of the syntenic arrangement



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.

of miRNAs. We collaborated on a number of projects to assess the impact of single nucleotide polymorphisms (SNPs) between individuals at the level of miRNAs or their targets.

Future plans

Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. Leonor Quintais will develop strategies for dealing with large-scale CLIP assays for microRNA target analysis. We will continue to build an accurate database of piRNA loci in animals and explore the importance and evolution of these molecules. We are extremely interested in the evolution of regulatory RNAs and developing phylogenetic techniques appropriate for short non-coding RNA. We will continue to build strong links with experimental laboratories that work on miRNAs in different systems. This will allow us to build better datasets with which to train and validate our computational approaches. The use of visualisation techniques to assist with the interpretation and display of complex, multi-dimensional data will continue to be an important parallel aspect of our work.

We will begin a new collaboration to study the effects of small and non-coding RNAs that are transported between cells with the laboratory of Stefano Pluchino at Cambridge University, Neuroscience. New predoc, Tomasso Leonardi aims to quantitate the levels of small and non-coding RNAs in vesicles and to explore their uptake and effect in recipient cells.

We have also secured further BBSRC funding as part of the BioLayout Express 3D project, which we will use to build new graphical interfaces for the analysis and exploration of next-generation sequence datasets in collaboration with Tom Freeman at the Roslin Institute in Edinburgh.

Selected publication

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

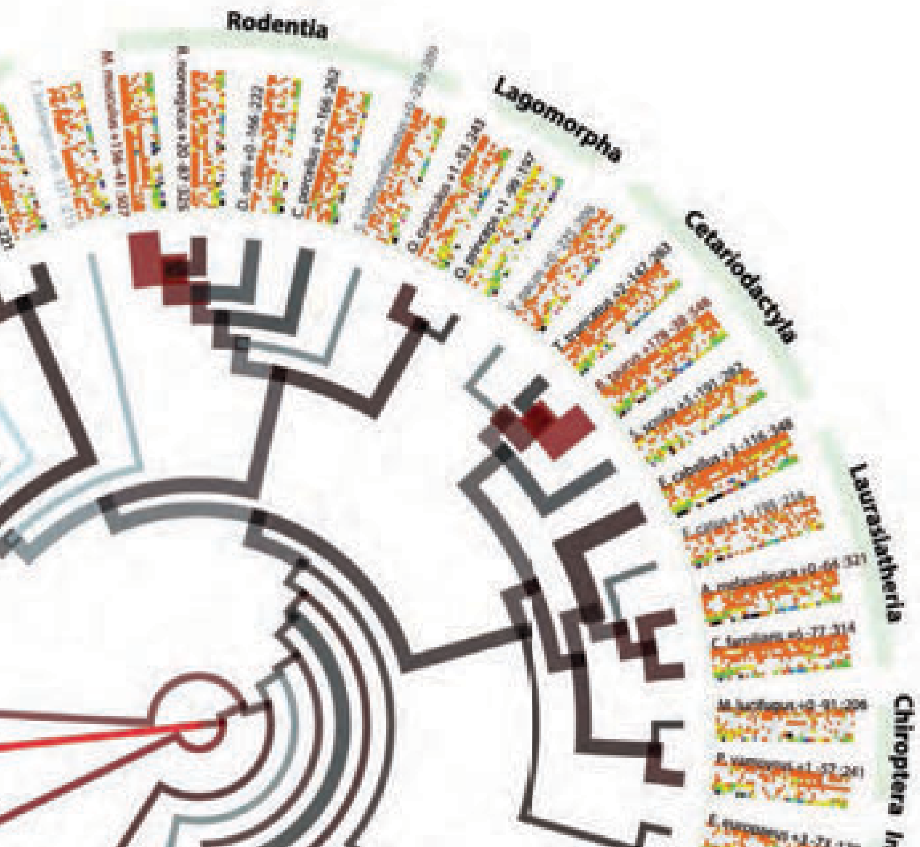


Figure. Evolutionary distribution of miRNA families. Phylogenetic tree representing miRNA family gains and losses. Branch width represents the number of miRNA families present among leaves of the branch, while the colour represents significant miRNA family loss (blue) or gain (red). For each of 408 miRNA families present at multiple loci in at least two species, we also build a graphical 'glyph'. This glyph can be used to quickly assess presence, absence or expansion of families between clades. Each square represents a specific miRNA family. Squares are coloured as follows: white indicates that this species does not contain a particular family; black indicates that this species contains at least 10 copies of miRNAs within that family. Numbers of copies between 1 and 10 are indicated as a rainbow gradient (red through violet). Groups of species are labelled according to the name of the evolutionary branch preceding them.

Goldman group

Evolutionary tools for genomic analysis

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

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

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

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

Summary of progress

- Developed novel computational methods to perform evolutionary analysis of ovarian cancer progression within individuals;
- Analysed protein-coding evolution in the African great apes, identifying genes undergoing accelerated evolution and estimating genome-wide levels of evolutionary constraint;
- Confirmed the 'ortholog conjecture,' which posits that studying related genes (orthologs) from other species, even distant ones, can inform us about human biology;
- Expanded our phylogeny-aware alignment method PAGAN to perform extension of existing alignments with new data;
- Designed new algorithms for orthologous group inference and assessed quality of Gene Ontology database functional annotations.

Major achievements

We played an important role in the publication of the gorilla genome in 2012—a major undertaking that involved the comprehensive analysis of ca. 5000 protein-coding genes in six primate species. We confirmed the expected overall similarity in evolutionary processes between the African great apes; identified genes in each species subject to changing evolutionary rate and found examples of notable genome-wide differences in selection pressures throughout the history of primates and the African great apes, consistent with historical variation in population size and demography.

We developed a new algorithm for efficient hierarchical orthologous group inference to pinpoint the emergence of particular gene copies in species' evolution and demonstrated its effectiveness on both simulated and empirical data. When the validity of the 'ortholog conjecture' was questioned, based on Gene Ontology (GO) classifications of orthologs and paralogs, Christophe Dessimoz and his colleagues at the Swiss Institute of Bioinformatics were prompted to examine data from tens of thousands of papers. By correcting for several biases affecting functional annotations, they demonstrated that the conjecture does indeed hold true. Furthermore, in an assessment of the quality of automatically inferred annotations in GO, which are often perceived as unreliable and disregarded, we found their reliability to be competitive with that of curated annotations, particularly when they can use non-experimental evidence from the literature.

In 2012 we investigated the prevalence of false positives and false negatives introduced by alignment error, and the ability of alignment filters to improve performance. Our own phylogeny-aware aligner, PRANK consistently performed better than other widely used aligners and, unlike them, did not benefit from the application of alignment filters.

Nick Goldman

PhD University of Cambridge, 1992. Post-doctoral work at National Institute for Medical Research, London, 1991-1995, and University of Cambridge, 1995-2002. Wellcome Trust Senior Fellow, 1995-2006.

At EMBL-EBI since 2002. EMBL Senior Scientist since 2009.

The accurate alignment of large numbers of sequences is a growing computational challenge, and one attractive solution is the addition of new sequences to existing alignments without full re-computation. We generalised the approach used in our PRANK tool so that it uses sequence graphs that, compared to profile-based aligners, allow the storage of more information about alternative putative indel events and a reference alignment. We have shown that the new method, called PAGAN, provides greater accuracy than its alternatives, especially for short sequence reads.

Using newly gathered sequences, we successfully validated predictions made in a previous study of caecilian amphibians to identify genes that should provide the most information relevant to phylogenetic reconstruction. We also completed experimental designs of the relationship between phylogenetic information and evolutionary rate and of the strategies for the phylogenetic placement of metagenomic data. The abundance of repetitive elements in genomes makes finished assembly of eukaryotic genomes a technical challenge. We developed a solution, NG-SAM (Next Generation Sequence Assembly aided by Mutagenesis), which introduces random mutations into copies of the repetitive region, allowing it to be inferred by consensus methods. NG-SAM relies only on PCR and dilution steps and can be scaled to an NGS workflow.

Future plans

The group remains dedicated to using mathematical modelling, statistics and computing to enable biologists to draw as much scientific value as possible from modern molecular sequence data. We shall continue to concentrate on linked areas that draw on our expertise in phylogenetics, genomics and next-generation sequencing. Basic to all our work are the fundamentals of phylogenetic analysis, where we are investigating the use of non-reversible models of sequence substitutions and problems such as ‘long branch attraction’. We remain committed to keeping abreast of evolving NGS technologies and exploiting them for new experiments. We will continue to develop novel data analysis methods, for example how to detect and represent the discordant evolutionary histories of different genes in an organism’s genome, as well as continuing our work making the most advanced evolutionary multiple sequence alignment methods available to the broadest possible range of researchers. We are beginning to look to medical applications of NGS and phylogenetics as a source of inspiring collaborations, and hope to start to bring molecular evolution into a clinical setting where it may soon be applicable in ‘real time’ to help inform doctors’ decisions and treatments.

Selected publications

Altenhoff, A.M., et al. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comp Biol* 8, e1002514.

Jordan, G. and Goldman, N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol* 29, 1125–1139.

Löytynoja, A., Vilella, A.J. and Goldman, N. (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* 28, 1684–1691.

Massingham, T. and Goldman, N. (2012) Error-correcting properties of the SOLiD Exact Call Chemistry. *BMC Bioinformatics* 13, 145.

Scally, A. et al. (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483, 169–175.

Sipos, B., et al. (2012) An improved protocol for sequencing of repetitive genomic regions and structural variations using mutagenesis and next generation sequencing, *PLoS One* 7, e43359.

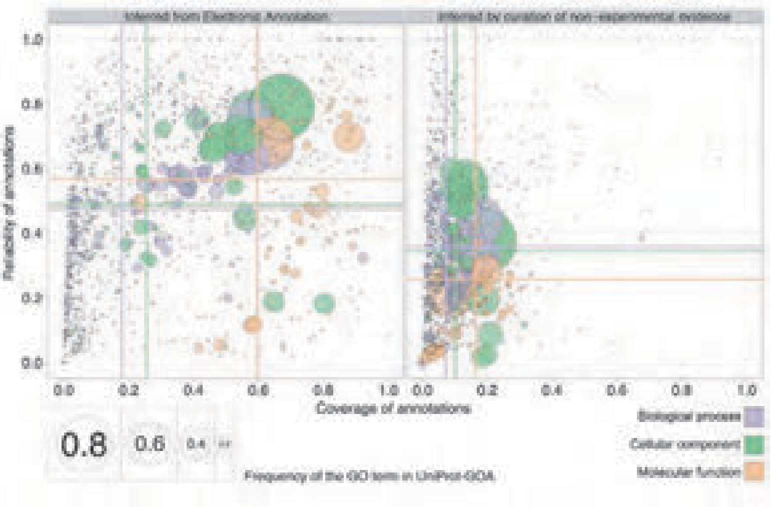


Figure. Quality of electronic and curated annotations on a common set of GO terms: Electronic annotations are more reliable than generally believed; their reliability (y axes) rivals that of curated annotations. They also have greater coverage (x axes). Terms associated with the three ontologies are drawn in different colours, with the area of the disc reflecting the frequency of the GO term in the UniProt-GOA release. The coloured lines correspond to the mean values for the respective ontologies and axes.

Le Novère group

Computational systems neurobiology

Our research interests revolve around signal transduction in neurons, ranging from the molecular structure of proteins involved in neurotransmission to signalling pathways and electrophysiology.

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 co-operativity, pathway switch and bi-stability.

The group provides community services that facilitate research in computational systems biology. For example, we lead the development of standard representations, encoding and annotating schemes, tools and resources for kinetic models in chemistry and cellular biology. The Systems Biology Markup Language (SBML) is designed to facilitate the exchange of biological models between different software. The Systems Biology Graphical Notation (SBGN) is an effort to develop a common visual notation for biochemists and modellers. We also develop standards for model curation (e.g., MIRIAM) and controlled vocabularies (e.g., the Systems Biology Ontology) to improve model semantics. To manage perennial cross-references, we run the MIRIAM Registry and its associated Identifiers.org Uniform Resource Identifier (URI) scheme. BioModels Database, developed in our group, 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

- Gained insights into the complex equilibria and kinetic events involved in calcium signalling by progressing the modelling of signalling pathways involved in synaptic plasticity;
- Increased the number of models provided by BioModels database by >500%: >900 models coming from literature, 140 000 models generated from pathway databases publicly distributed in 2011.

Major achievements

In 2012 we focused on modelling signalling pathways that are involved in synaptic plasticity. Progress in this area led to a deeper understanding of the complex equilibria and kinetic events involved in calcium signalling. In particular, we shed light on the mechanisms by which calcium, calmodulin and calcium/calmodulin kinase II interact.

We continued to develop the BioModels Database, which saw explosive growth: an increase of over 500% in 2012. This resource now offers to the community more than 900 models from scientific literature and 140 000 models generated from pathway databases. This represents more than 10 million mathematical relations and 400 million cross-references linking models and data. Computational models of biological processes can be used in a variety of formats, or to build other models.

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 from 2003 to 2012.

Future plans

The group split up in October 2012. The basic research component moved to the Babraham Institute, located three miles from Hinxton in Cambridgeshire. The Le Novère Lab will develop there as part of the computational and mathematical biology activity of the institute. Strong emphasis will continue to be placed on signalling, with other fields of investigation arising from various collaborations.

The service side of the group will be integrated with the Proteomics Service Team, led by Henning Hermjakob. BioModels Database and Identifiers.org will join the production side of EMBL-EBI, which will present both an opportunity and a challenge.

Selected publications

Juty, N., et al. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res* 40, D580-D586.

van Iersel, M.P., et al. (2012) Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinformatics* 28, 2016-2021.

Stefan, M. I., et al. (2012) Structural analysis and stochastic modelling suggest a mechanism for calmodulin trapping by CaMKII. *PLoS One* 7, e29406.

Adams, R.R. (2012) The Input Signal Step Function (ISSF), a standard method to encode input signals in SBML models with software support, applied to circadian clock models. *J Biol Rhythms* 27, 328-332.

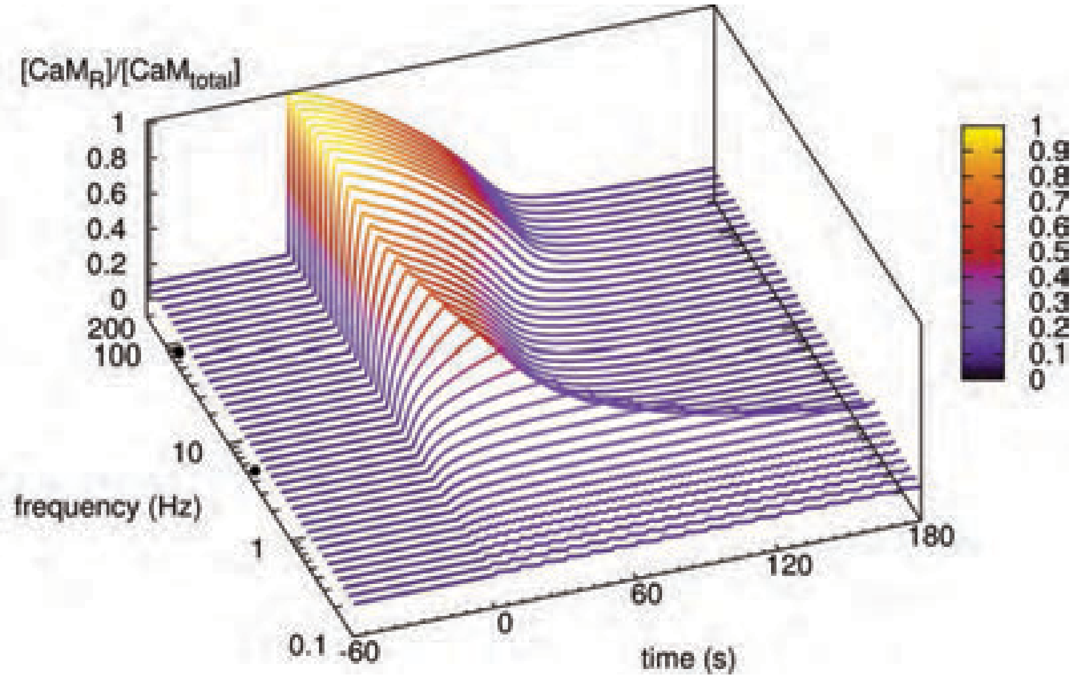


Figure. Effect of calcium stimulations of a dendritic spine on the activity of calmodulin. Calcium is provided at different frequencies but fixed total amount, resulting in variable signal duration. Calmodulin activity is displayed as the fraction of molecules in the open (R) conformation. When the frequency passes 5 Hz, the open state of calmodulin is stabilised by its binding to CaM Kinase II, resulting in an activation outlasting the initial signal. Li et al. Calcium input frequency, duration and amplitude differentially modulate the relative activation of calcineurin and CaMKII. Figure reproduced from PLoS ONE (2012) 7, e43810.

Luscombe group

Genomics and regulatory systems

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.

Advances in sequencing and other laboratory techniques have given rise to unprecedented volumes of information describing the function and organisation of regulatory systems. Yet many observations made with these data are unexpected and appear to complicate our view of gene expression control. We are interested in how dysfunctions in the regulatory system, specifically in transcription factors (TFs), give rise to a third of human developmental disorders and yet play an important role in evolutionary adaptation and innovation.

The combination of computational biology and genomics enables us to uncover general principles that apply to many different biological systems; any unique features of individual systems can then be understood within this broader context. Our work applies computational and genomic methods to explore how gene expression is regulated, how these mechanisms control interesting biological behaviours and how gene regulation interacts with evolutionary processes.

Summary of progress

- Compared the sequences of 34 *E. coli* genomes and demonstrated that local mutation rates have been evolutionarily optimised (Martincorena et al., 2012; Martincorena and Luscombe, in press), settling a 50-year old question and revising a central tenet of evolutionary theory;
- Dissected chromosome-wide mechanisms involved in the precise two-fold up-regulation of male X-linked genes during fly dosage compensation, and demonstrated that dosage compensation is controlled by enhanced initiation (Conrad et al., 2012);
- Developed truly predictive statistical models that for the first time accurately reproduce even-skipped expression;
- Discovered how competitive binding between U2AF65 and hnRNP C protects the transcriptome from the detrimental exonisation of thousands of Alu elements (Zarnak et al., 2012).

Major achievements

Evolution of local mutation rates

By comparing the sequences of 34 *E. coli* genomes, we demonstrated that local mutation rates have been evolutionarily optimised (Martincorena et al., 2012; Martincorena and Luscombe, in press). Mutation rates follow a 'risk management' strategy, prioritising the cells' resources to protect 'important' genes. The study settled a 50-year old question and revised a central tenet of evolutionary theory. It also raised intriguing new questions about how DNA-repair mechanisms are deployed across the genome.

Epigenetic mechanisms for dosage compensation

In collaboration with the Akhtar laboratory at the Max Planck Institute for Immunobiology and Epigenetics in Germany, we dissected chromosome-wide mechanisms involved in the precise two-fold up-regulation of male X-linked genes during fly dosage compensation. Previously, we discovered the first context-dependent activity of a histone acetyltransferase (Kind et al., 2008) and reported the one of the first genome-wide

Nick Luscombe

BA University of Cambridge 1996. PhD University College London 2000. Anna Fuller Postdoctoral Fellow, Yale University 2000-2004.

At EMBL-EBI from 2005 to 2012.

roles of nucleoporin subunits in gene regulation (Vaquerizas et al., 2010). In 2012 we accurately measured the two-fold increase in Pol II-binding at promoters, demonstrating that dosage compensation is controlled by enhanced initiation (Conrad et al., 2012).

Statistical models of gene expression in fly development

Using compiled-in-situ hybridisation images from the Virtual Embryo dataset, we developed statistical models that, for the first time, accurately reproduce even-skipped expression. The models precisely forecast behaviours beyond the training data, making them truly predictive of, for example, the effects of regulatory perturbations. The study generated experimentally testable hypotheses and provided new insights into the underlying mechanisms of transcriptional regulation (Ilsley et al., submitted).

Prevention of aberrant exonisation of Alu elements

In previous years we collaborated with Jernej Ule's laboratory (MRC Laboratory of Molecular Biology, Cambridge, UK) to develop nucleotide-resolution, genome-wide techniques to identify protein-RNA interactions (Konig et al., 2010) and demonstrate how hnRNP C binds to enhanced and repressed splice sites. In 2012 we discovered how competitive binding between U2AF65 and hnRNP C protects the transcriptome from the detrimental exonisation of thousands of Alu elements (Zarnak et al., 2012).

Future plans

Nick Luscombe is now Chair in Computational Biology at the University College London Genetics Institute, and holds a joint appointment as Senior Group Leader at the Cancer Research UK London Research Institute. The Luscombe group will join the Francis Crick Institute when it opens in 2015. Their work will continue to focus on the nuclear organisation of chromosomes, gene regulation in disease states, gene regulation and DNA-damage repair.

Selected publications

Conrad, T., et al. (2012) Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters. *Science* 337, 742-746.

Jolma, A., et al. (2013) Binding specificities of human transcription factors. *Cell* 152, 327-339.

Martincorena, I., et al. (2012) Evidence of non-random mutation rates suggests a risk management strategy for evolution. *Nature* 485, 95-98.

Tan-Wong, S.M., et al. (2012) Gene loops enhance transcriptional directionality. *Science* 338, 671-675.

Zarnack, K., et al. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the uncontrolled exonization of Alu elements. *Cell* 152, 453-466.

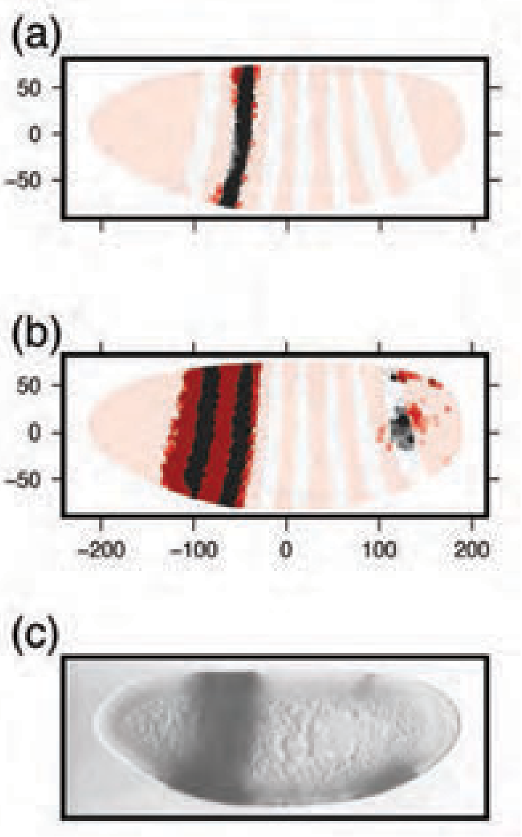


Figure. Neutral mutation rates in the E. coli genome are heterogeneous and non-randomly distributed.

Marioni group

Computational and evolutionary genomics

Gene expression levels play a critical role in evolution, developmental processes and disease progression.

Variability in the transcriptional landscape can help explain phenotypic differences both between and within species. As a result, identifying and characterising the regulatory mechanisms responsible for changes in gene expression is critically important. Next-generation sequencing technology has revolutionised our ability to do this. By facilitating the creation of unbiased, high-resolution maps of genomes, transcriptomes and regulatory features such as transcription factor binding sites, these experimental techniques have given rise to a detailed view of gene expression regulation in both model and non-model organisms.

To make the most of these technological developments, we devise effective statistical and computational methods for analysing the vast amounts of data generated. Harnessing both experimental and computational biology helps us truly understand complex biological processes such as gene regulation. Our group focuses on the development of computational methods for interrogating high-throughput genomics data. We concentrate on modelling variation in gene expression levels in different contexts: between individual cells from the same tissue; across different samples taken from the same tumour and at the population level, where a single, large sample of cells is taken from the organism and tissue of interest. We apply these methods to a range of biological questions, from examining the regulation of gene expression levels in a mammalian system to the development of the brain in a marine annelid. We collaborate with outstanding experimental groups, both within and beyond EMBL.

Summary of progress

- Combined RNA-sequencing with DNA-sequence information to interrogate gene and isoform regulatory mechanisms in a mammalian system (Goncalves et al., 2012);
- Developed approaches for quantifying how different sequence analysis tools effect gene expression measurements derived from RNA-sequencing data (Fonseca et al., 2012);
- Developed statistical tests for assessing the form of row and column covariance matrices by exploiting matrix-variate distributions and applied these approaches to gene expression data derived from glioblastoma samples;
- Modelled the variability present in single-cell RNA-sequencing data using a computational and experimental approach;
- Developed a computational approach to infer the kinetics of stochastic gene expression from single-cell RNA-sequencing data (Kim and Marioni, 2013, in press).

Major achievements

Gene regulation in mammals

In collaboration with the Odom group at Cancer Research UK-Cambridge Research Institute, the Brazma group and the Flicek group we collected and developed models for analysing RNA-sequencing data obtained from liver samples derived from both the parents and F1 crosses of two different strains of mice. This system allows us to categorise genes into different sets depending upon the mechanism by which they are regulated. We found that a surprisingly large proportion of genes were regulated by both cis and trans regulatory variants, suggesting extensive compensatory cis-trans regulation in the evolution of mouse gene expression (Goncalves et al., 2012).

Modelling RNA-sequencing data

A key step in the analysis of RNA-sequencing data is accurate alignment of the short reads to the relevant reference genome to quantify gene expression levels. In 2012 we collaborated with the Brazma group to develop the quantification and analysis tools necessary to establish a baseline atlas of gene expression levels, which will be released as an EMBL-EBI service. The baseline expression atlas will use RNA-seq data to present a simple interface that can be used to determine the set of tissues in which a gene is expressed and the sets of

John Marioni

PhD in Applied Mathematics, University of Cambridge, 2008. Postdoctoral research in the Department of Human Genetics, University of Chicago.

At EMBL since 2010.

genes with expression profiles that characterise each tissue. In this context, the choice of mapping and quantification approaches is a major consideration (Fonseca et al., 2012).

Interpreting spatial variability in gene expression levels

Tumours typically consist of multiple clones, all of which can carry mutually exclusive genetic mutations and display highly divergent gene expression profiles. In the clinical context, an intervention targeting a mutation present in one clone may not have any effect upon other clones, allowing the tumour to remain present. We collaborated with Colin Watts and Simon Tavaré at the University of Cambridge to analyse a dataset in which at least four spatially independent sub-samples were taken from glioblastoma multiforme tumours from eight patients. We measured gene expression levels for each sub-sample to estimate the correlation between the different sub-samples of the tumour, and developed computational methods to estimate and test the structure of the sub-sample covariance matrix, accounting for correlation between genes with highly co-ordinated patterns of expression. Ignoring this correlation structure can lead to inaccuracies in the estimation of the sub-sample correlation matrix.

Modelling single-cell RNA-seq

Studying gene expression levels at the single-cell level is crucial in many biological contexts, such as for example in tumours, which might harbour multiple different sub-types. We collaborated with groups in the EMBL Developmental Biology Unit to systematically assess the quality and limitations of generating this type of data using high-throughput sequencing. We designed dilution experiments to study both the technical and biological variability present in single-cell RNA-seq data (Brennecke et al., in preparation). In parallel, we developed computational methods for quantifying the variability present in gene-expression profiles obtained from multiple cells taken from the same tissue (Kim and Marioni, 2013).

Future plans

We will work with our experimental collaborators to apply our methods to important biological questions. From a computational perspective, modelling single-cell transcriptomics data will increase in importance. Methods for storing, visualising, interpreting and analysing the data generated will be critical if we are to exploit the information to the fullest extent. We will also work on methods for analysing conventional next-generation sequencing data, building on previous work.

Selected publications

Goncalves, A. et al. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res* 22, 2376-2384.

Fonseca, N.A. et al. (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169-3177.

Kim J.K. and Marioni J.C. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* (in press).

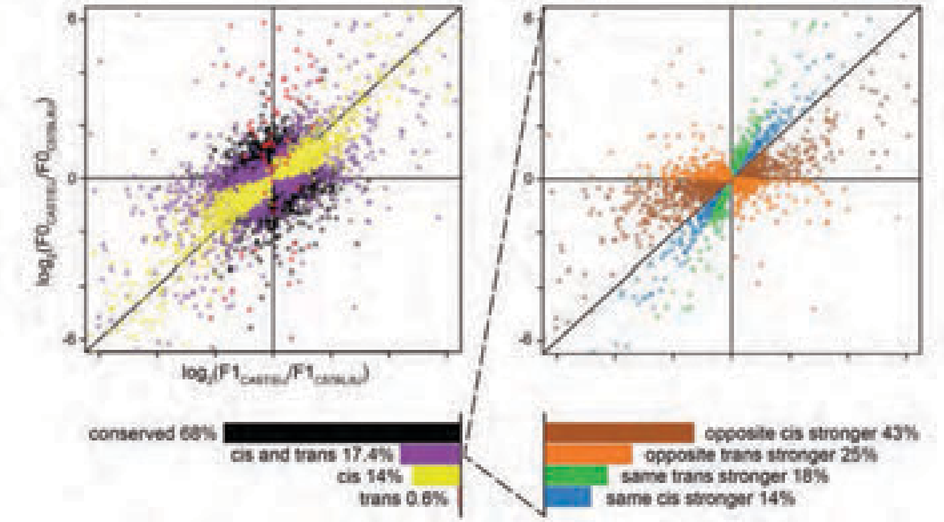


Figure. Classifying genes by their regulatory function. We used RNA-seq data generated from F0 mice and their F1 hybrids to classify genes into sets depending upon their regulatory mechanism. [Goncalves et al., Genome Research 2012.]

Rebholz group

Phenotypes and multilingual resources

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 Europe PubMed Central.

Research in the Rebholz group focuses on extracting facts from the literature. Our goal is to connect literature content automatically to other biomedical data resources and to evaluate the results. Our research targets centre on the extraction of information on the bioactivities of drugs, the identification of phenotype mentions and the analysis of the discourse structure in the scientific literature. Another focus is multilingual information retrieval with augmented multi-lingual terminological resources, for example in the CLEF-ER initiative.

Summary of progress

- Normalised the representation of concepts in the literature (using LexEBI, IeXML, Whatizit and CALBC), thereby contributing to the Europe PubMed Central resource and sparking success in the acquisition of MANTRA, a novel research project on the automatic generation of multi-lingual biomedical terminologies;
- Developed and evaluated a new machine-learning-based solution for the characterisation of scientific statements to qualify information-extraction results according to the quality of the author's statements;
- Carried forward efforts in the identification of phenotypes, reaching the evaluation of methods for the transformation of pre- into post-composed representations;
- Developed a library for biomedical knowledge manipulation (Brain) evaluated it for data integration using OWL representations of data and automatic reasoning techniques (work by Samuel Croset);
- Began to normalise and translate patient-specific descriptions in different languages by combining post-composed phenotype identification and multilingual terminological resources.

Our group develops solutions that link named entities in the literature (e.g., genes, proteins, diseases) to entries in a reference database: LexEBI, a terminological resource; IeXML, an annotation framework for documents; Whatizit, an information-extraction infrastructure; and CALBC, an evaluation infrastructure. LexEBI provides full coverage of domain knowledge for gene and protein names, chemical entities, diseases, species and ontological terms and is integrated with IeXML and Europe PubMed Central. CALBC harmonises annotations from automatic text mining solutions and comprises 889 282 Medline abstracts with over 12 million annotations.

Major achievements

Optimising the semantic linkage of literature to biomedical data resources is key to supporting intelligent full-text searches. The identification of phenotypes in text using Entity-Quality (EQ) representations is the latest contribution to this goal, as it helps link results from experimental biology to evidence from electronic patient records.

Standardisation of the scientific literature: the multilingual approach

The biomedical literature is in many different languages and does not use a consistent language across labelling systems. In addition, terminological resources are incomplete and are not well aligned with the literature resources. In the context of the EU-funded MANTRA project, we gathered the publicly available documents in which each English

Dietrich Rebholz-Schuhmann

PhD in immunology, University of Düsseldorf, 1989. Senior Scientist at GSF, Munich, 1995. Director of Healthcare IT, LION Bioscience AG, Heidelberg, 1998.

At EMBL-EBI from 2003 to 2012.

version is accompanied by a non-English-language version ('parallel corpora'): 120 638 patent documents, 1 593 546 Medline titles and 364 005 EMEA documents. We annotated and harmonised documents in English to generate a Silver Standard Corpus, and started to apply this same approach to non-English documents. We formulated a community challenge in which participants use these documents to produce English and non-English entity labels. The goal is to exploit these data to augment existing English and non-English terminologies for public use.

Network extraction from the scientific literature and evaluation against BioModels database

Metabolic and signalling pathways are often viewed separately, even though both are composed of interactions involving proteins and other chemical entities. Bringing together data from all available resources makes it possible to judge the functionality, complexity and completeness of any given network. The full integration of relevant information from the scientific literature remains a challenge. However, we processed the literature using rich linguistic features, such as full-text parsing, in order to extract biological interactions, and moved towards combining these with information from scientific databases to support the extension of knowledge about biological networks.

Knowledge discovery and novel text-mining solutions

Scientists use a well-established discourse structure in their published work, including familiar categories such as method, result and conclusion. We defined 11 such core scientific concepts (hypothesis, motivation, goal, object, background, method, experiment, model, observation, result, conclusion) and used previously trained machine-learning classifiers (SVM and CRF) to automatically recognise these concepts on a corpus of 265 full-text articles. We evaluated our automatic classifications against a manually annotated gold standard and demonstrated promising accuracies with the categories experiment, background and model (F1 scores of 76%, 62% and 53%, respectively). Interestingly, local sentence features (e.g., unigrams, bigrams, grammatical dependencies) were the most discriminative features while section headings, which encode the document structure, played an important role.

Development and use of phenotype ontological resources

The identification of phenotype mentions in the literature, case reports and patient records supports knowledge discovery in established resources. A major challenge is identification and semantic normalisation of phenotype mentions, as they vary in structure and form. We developed an approach to identifying complex phenotype mentions based on a combination of standard terminological resources and a context-sensitive, corpus-trained solution. We then demonstrated that this approach is robust and enables efficient identification of complex phenotype descriptions.

We developed an alternative approach, Entity-Quality (EQ) representations of phenotypes, using the pre-composed representation of phenotype labels in standard ontologies. This solution enables the automatic alignment of different phenotype ontologies. We applied it to a subset of mammalian and human phenotype ontology concepts and identified the correct EQ representation in over half of structure and process phenotypes for the mammalian. The approach was less successful with human phenotype ontologies.

Future plans

Dr Rebholz-Schuhmann now leads a research team at the University of Zurich's Institute of Computational Linguistics. His group's work at EMBL-EBI continues at a reduced scale in 2013 as PhD students Ying Yan, Chen Li, and Samuel Croset complete their work and research fellows Nigel Collier and Maria Liakata further their efforts in the text-mining domain.

Selected publications

Liakata, M., et al. (2012) Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28, 991-1000.

Campos, D., et al. (2012) Harmonization of gene/protein annotations: towards a gold standard MEDLINE. *Bioinformatics* 28, 1253-1261.

Oellrich, A., et al. (2012) Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PLoS One* 7, e38937.

Harrow, I., et al. (2013) Towards virtual knowledge broker services for semantic integration of life science literature and data sources. *Drug Discovery Today* (in press).

Saez-Rodriguez group

Systems biomedicine

Our group aims to achieve a functional understanding of signalling networks and their deregulation in disease and seeks to apply this knowledge to novel therapeutics.

Human cells are equipped with complex signalling networks that allow them to receive and process the information encoded in myriad extracellular stimuli. Understanding how these networks function is a compelling scientific challenge but also has practical applications, as alteration in the functioning of cellular networks underlies the development of diseases such as cancer and diabetes. Considerable effort has been devoted to identifying proteins that can be targeted to reverse this deregulation. However, their benefit is often unexpected. it is hard to assess their influence on the signalling network as a whole and thus their net effect on the behaviour of the diseased cell. Such a global understanding can only be achieved by a combination of experimental and computational analysis.

Our research is hypothesis-driven and tailored towards producing mathematical models that integrate diverse data sources. To this end, we collaborate closely with experimental groups. Our models integrate a range of data (from genomic to biochemical) with various sources of prior knowledge, with an emphasis on providing both predictive power of new experiments and insights into the functioning of the signalling network. We combine statistical methods with models describing the mechanisms of signal transduction, either as logical or physico-chemical systems. For this, we develop tools and integrate them with existing resources. We then use these models to better understand how signalling is altered in human disease and predict effective therapeutic targets.

Summary of progress

- Developed CellNOpt, an R/Cytoscape platform to model signalling networks using logic formalisms of different quantitative and time resolution;
- Co-organised (including the set up and hosting of the website) the seventh edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods);
- Established methods to analyse large drug screenings in cancer cell lines and predict drug efficacy;
- Developed DvD, a pipeline for drug repurposing using public repositories of gene expression data.

Major achievements

In 2012 we established CellNOpt, as an R/Cytoscape platform for modelling signalling networks using logic formalisms of different quantitative and time resolution. CellNOpt uses high-throughput biochemical data to generate models spanning simple Boolean logic models that coarsely describe signalling networks and continuous fuzzy-logic models, as well as differential equation systems describing details in the dynamics of the underlying biochemical processes.

Our group co-organised the seventh edition of DREAM (Dialogues in Reverse Engineering Assessment of Methods; www.the-dream-project.org), a community effort organised around challenges to advance the inference of mathematical models of cellular networks. In 2012 we co-ordinated a DREAM challenge to investigate ways to select the most informative experiments in order to identify the parameters in a network model. We also established various methods to analyse large drug screenings in cancer cell lines, particularly for the prediction of drug response from genetic features, in the context of the Wellcome Trust Sanger Institute/EMBL-EBI (ESPOD) postdoctoral programme and in collaboration with the groups of Garnett, McDermott and Stratton at the Sanger Institute, Benes at Massachusetts General Hospital and Wessels at the Netherlands Cancer Institute.

Julio Saez-Rodriguez

PhD University of Magdeburg, 2007. Postdoctoral work at Harvard Medical School and MIT.

At EMBL-EBI since 2010. Joint appointment, EMBL Genome Biology Unit.

We refined methods to infer new effects of already approved compounds from gene expression data, leading to new drug repurposing opportunities. We implemented these methods in collaboration with Sanofi and the Bork lab at EMBL Heidelberg in the Drug v Disease tool (DvD).

Future plans

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

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

Selected publications

Pacini, C., et al. (2013) DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 29, 132-134.

Iorio, F., et al. (2013) Transcriptional data: a new gateway to drug repositioning? *Drug Disc Today* (in press).

Terfve, C., et al. (2012) CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol* 6, 133.

Eduati, F., et al. (2012) Integrating literature-constrained and data-driven inference of signalling networks. *Bioinformatics* 28, 2311.

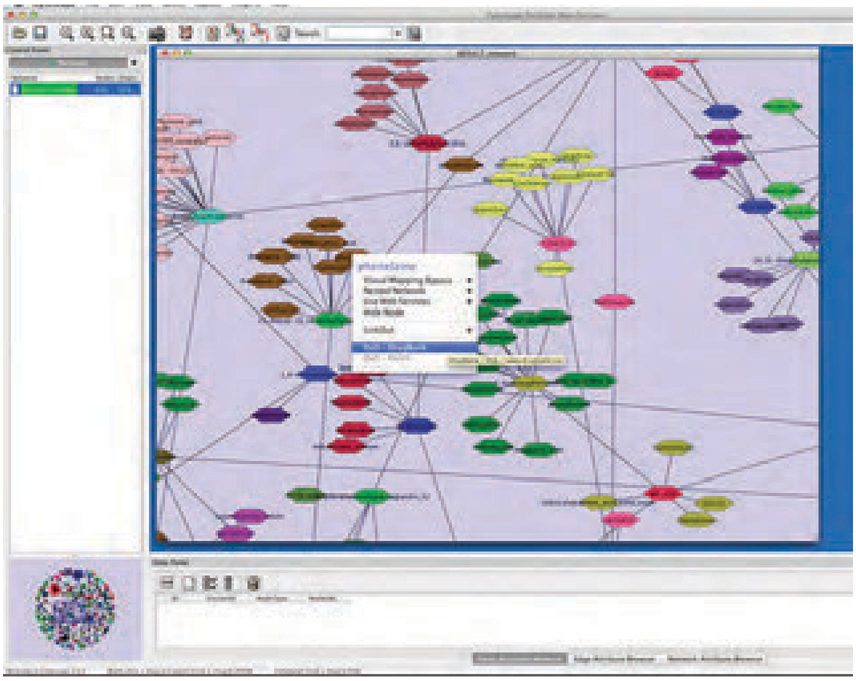


Figure. Screenshot in Cytoscape of DvD, a tool for drug repurposing using public repositories of gene expression data.

Thornton group

Proteins: structure, function and evolution

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

We explore the way enzymes perform catalysis by gathering relevant data from the literature and developing novel software tools, which allow us to characterise enzyme mechanisms and navigate the catalytic and substrate space. In parallel, we investigate the evolution of these enzymes to discover how they can develop new mechanisms and specificities. This involves integrating heterogeneous data with phylogenetic relationships within protein families, which are based on protein structure classification data derived by colleagues at University College London (UCL). The practical goal of this research is to improve the prediction of function from sequence and structure and to enable the design of new proteins or small molecules with novel functions.

We also explore sequence variation between individuals in different contexts and for different species. To understand more about the molecular basis of ageing in different organisms, we participate in a robust collaboration with experimental biologists at UCL. Our role is to analyse functional genomics data from flies, worms and mice and, by developing new software tools, relate these observations to effects on life span.

Summary of progress

- Developed a sophisticated, novel web tool to compare enzyme reactions and navigate through reaction (E.C.) space—used the tool to re-analyse enzyme reactions, focusing on the isomerases;
- Published our analysis of enzyme evolution based on a phylogenetic analysis involving sequences, structures and the chemistry of the reaction;
- Released a new web tool, FunTree, to analyse the evolution of protein domain families and their catalytic functions (Furnham et al., 2012a,b);
- Completed our analysis of human nsSNP data from the 1000 Genomes Project, developing a pipeline to map the observed mutations onto protein sequences and structures. Demonstrated a radical difference between the distribution of variants observed in the 1000 Genomes data and those in OMIM;
- Further developed our database of survival curves and used it to perform meta-analyses of the influence of strain, sex and bacterial infection over many independent experiments—this database and analysis tool will be publicly available in 2013;
- Developed an answer-set-programming-based approach to model the relationship between the insulin signalling pathway and longevity, enabling the combination of data from multiple experiments to test factors that affect longevity (Papatheodorou et al., 2012).

Major achievements

Enzymes are crucial to life, and understanding how they have evolved to perform the wide variety of reactions found across all kingdoms of life is fundamental to a broad range of biological studies, especially those leading to new therapeutics. Unravelling the evolution of novel enzyme function requires combining information on protein structure, sequence, phylogeny and chemistry (i.e., interacting small molecules and reaction mechanisms). We developed a protocol for integrating this wide range of data, which we applied to a relatively large number of families comprising some very diverse relatives. One observation was that some changes in function among relatives are more common than others, with most of the functionality observed in nature confined to relatively few families (Furnham et al., 2012a). This work provides a starting point to address questions about the most common paradigms that underlie the ability of an enzyme to evolve new catalytic functions. We began to explore ways this information can be used to predict the function of an enzyme that has yet to be experimentally characterised and to design new enzymes for industrial and medical purposes. The resource is freely available (FunTree; Furnham et al., 2012b).

A challenge in our ageing research has been the integration of incomplete knowledge on pathways with existing

Janet Thornton

PhD King's College & National Institute for Medical Research, London, 1973. Postdoctoral research, University of Oxford, NIMR & Birkbeck College, London. Lecturer, Birkbeck College 1983-1989. Professor of Biomolecular Structure, University College London since 1990. Bernal Professor at Birkbeck College, 1996-2002. Director of the Centre for Structural Biology at Birkbeck College and University College London, 1998-2001.

Director of EMBL-EBI since 2001.

experimental datasets and the relating of these to measured ageing phenotypes, including longevity. We developed a logic-based method that employs answer set programming and used it to infer the signalling effects of genetic perturbations, based on a model of the insulin signalling pathway (Papatheodorou et al., 2012). We applied our method to analyse the outputs from several RNA expression experiments for Drosophila mutants in the insulin pathway that alter lifespan. Our comparisons reveal that the transcriptional changes observed for each mutation usually provide negative feedback in the insulin signalling pathway, perhaps reflecting the need for homeostasis. We see opposite effects of transcriptional feedback on short- and long-lived flies. We also identify an S6K-mediated feedback in two long-lived mutants. Our methods are available on our website in NetEffects.

Future plans

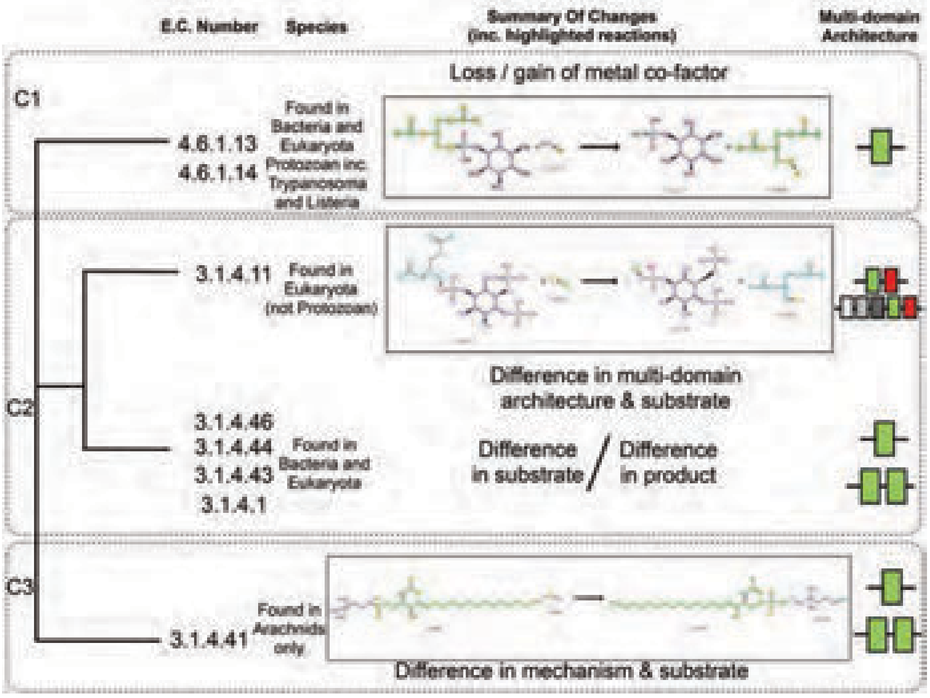
Our work on understanding enzymes and their mechanisms using structural and chemical information will include a study of how enzymes, their families and pathways have evolved. We will study sequence variation in different individuals, including humans, flies and bacteria, and explore how genetic variations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

Selected references

Furnham, N., et al. (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Comp Biol* 8, e1002403.

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

Papatheodorou, I., et al. (2012) Using answer set programming to integrate RNA expression with signalling pathway information to infer how mutations affect ageing. *PLoS One* 7, e50881.



E.C. Number
Species
Summary Of Changes (inc. highlighted reactions)
Multi-domain Architecture
Loss / gain of metal co-factor
C1
4.6.1.13
4.6.1.14
Found in Bacteria and Eukaryota Protozoan inc. Trypanosoma and Listeria
C2
3.1.4.11
Found in Eukaryota (not Protozoan)
Difference in multi-domain architecture & substrate
3.1.4.46
3.1.4.44
3.1.4.43
3.1.4.1
Found in Bacteria and Eukaryota
Difference in substrate / Difference in product
C3
3.1.4.41
Found in Arachnids only
Difference in mechanism & substrate

Figure. The evolution of novel enzyme functions: The phosphatidylinositol (PI) phosphodiesterase domain superfamily divides into three clades (C1 - C3) by phylogenetic analysis. Significant changes in function occur between clades (as well as within clades) as defined by the Enzyme Commission classification number with underlying changes in reaction chemistry as captured by atom-atom mapping of the reaction of small molecules as well as changes in multi-domain architecture.

The EMBL International PhD Programme at EMBL-EBI

Students mentored in the EMBL International PhD Programme receive advanced, interdisciplinary training in molecular biology and bioinformatics.

Theoretical and practical training underpin an independent, focused research project under the supervision of an EMBL-EBI faculty member and monitored by a Thesis Advisory Committee (TAC). The TAC comprises EMBL-EBI faculty, local academics and, where appropriate, industry partners. Our PhD students are awarded their certificates from the University of Cambridge.

Nenad Bartonicek

Computational studies on the biogenesis and function of small non-coding RNAs

Nenad worked in the Enright group, focusing on non-coding RNA. The function of miRNAs and piRNAs has only recently been investigated. miRNAs act as global silencers of genes, targeting several hundreds at a time. Some piRNAs specifically target genomic parasites (the retroviruses) while the function of others remains unknown. Nenad studied how these small non-coding RNAs can be detected, what drives their biogenesis and how RNA's genomic position is specific to their DNA counterparts. He also developed a web server, SylArray, that allows fast detection, analysis and visulaisation of small RNA signatures.

Joe Foster

Transomics: Integrating core 'omics' concepts

Joe's work in the Apweiler group focused on the challenges of dealing with high-throughput –omics data. Looking at 'omics' as a whole, Joe took inspiration from the successful aspects of particular 'omics' fields and applied them to others. He developed several methods for the evaluation of publicly available data from mass spectrometry derived data in the PRIDE database. From this, an open source R library was developed for the quality control of proteomics data. Secondly, a database of theoretical lipid species was created and a web application based on the protein equivalent (UniProt), developed. Statistical analysis of lipidomics data from human colorectal cancer was also part of Joe's work and he produced a machine classifier that predicts whether a sample is of tumour or normal origin, solely based upon quantitative lipid data.

José Afonso Guerra-Assunção

Evolutionary analysis of animal microRNAs

José Afonso Guerra-Assunção in the Enright group identified novel strategies to enable a coherent view of miRNA evolution and developed a tool for large-scale, species-independent miRNA mapping. His analysis of phylogenetic profiles uncovered interesting co-evolution between miRNA and proteins across more than 800 species. The ability to put miRNA biology into an evolutionary context is potentially very useful, and Afonso's tools are openly available to the community and designed to be easily updated.

Angela Goncalves

RNA sequencing for the study of gene expression regulation

Angela, in the Brazma group, developed a pipeline and methodology for expression qualification in RNA sequencing datasets. Her subsequent work on the evolution of gene expression regulation focused on the divergence of liver gene expression and of isoform usage between closely related mice strains. By comparing the expression of two parental strains to the allele-specific expression in their hybrid offspring, she found that divergent expression between closely related mammalian species results mostly from a combination of regulatory variants acting in the cis and trans, the majority of which act in opposite directions. These results suggest extensive occurrence of compensatory regulation of gene expression levels, an observation that has important implications for understanding how speciation occurs.

Gregory Jordan

Analysis of alignment error and sitewise constraint in mammalian comparative genomics

Gregory's work in the Goldman group explored methods for sequence alignment and the detectiion of positive selection, carrying out a series of simulation experiments and empirical studies in mammals and primates. Gregory’s experimental work showed that rodents have less detectable positive selection in protein-coding genes than primates and other mammals. He developed new methods for combining site-wise estimates across genes and protein-coding domains, identifying well-known and novel signals for positive selection in mammals. Gregory also used codon-based models to identify genes undergoing accelerated evolution in gorillas with plausible relationships to this primate’s unique phenotypic and behavioural characteristics.

Inigo Martincorena

Genome-scale strategies controlling the impact of deleterious mutations

Inigo, in the Luscombe group, focused on fundamental questions about how point mutation rate varies along a genome. He showed that the neutral point mutation rate varies by over an order of magnitude across 2659 *Escherichia coli* genes, and that this variation appears to have been evolutionarily optimised to reduéce the risk of deleterious mutations. His thesis introduces an evolutionary risk management system to explain how non-random mutation rates evolve in a genome. By examining genome-wide RNS-binding and RNA –sequencing data, he demonstrated that the splicing regulator hnRNPC binds to hundreds of intronic Alu sequences, repressing their accidental exonisation and avoiding the deleterious disruption of hundreds of human transcripts.

Pablo Moreno

Bioinformatic methods for species-specific metabolome inference

Pablo, in the Steinbeck group, investigated ways to identify the large number of metabolites that remain illusive to metabolomics. His methods for predicting organism-specific metabolomes through metabolism database integration, text-mining and chemical enumeration provide a starting point for the semi-automatic generation of species-specific metabolomes with new techniques that produce a rich organism-specific catalogue of small molecules. Among other important observations, his analyses reveal that we are still far from having a complete Human Metabolome catalogue, and that current sources of metabolome knowledge are highly complementary.

Anika Oellrich

Supporting disease candidate gene discovery based on phenotype mining

Anika, in the Rebholz group, worked on supporting the identification of disease-gene candidates by mining phenotype information from the Mammalian Genome Informatics (MGI) database, the Online Mendelian Inheritance in Man (OMIM) database and the scientific literature. She developed a pipeline that ranks mouse models for human genetic disorders and enables the identification of promising disease-gene candidates. She also generated mouse-specific disease profiles, and demonstrated their validity and evaluated her results against disease-gene reporting databases.

Petra Schwalie

Genome-wide analysis of alignment error and sitewise constraint in mammalian comparative genomics.

Petra, working in the Flicek group, used a genomics approach to explore transcriptional regulation in animals. She analysed the evolution of CTCF binding in vertebrates and demonstrated the role of retrotransposons in shaping the binding landscape of this key genome regulator. She showed a CTCF-independent role of the cohesin protein complex in tissue-specific transcriptional regulation of human cell lines and, finally, used in vivo approaches in the fruit fly brain to reveal both extensive RNA polymerase II binding differences among tissues and largely tissue-invariant use of the histone variant H2A.v.

Diva Tommei

Transcriptional characterization of glioma neural stem cells

Gliomas are the most lethal central nervous system cancers. Working in the Bertone group, Diva sought to identify differences between healthy brain and glioma stem cells, conducting an in-depth characterization of their transcriptomes comparing gene and isoform expression, molecular signature profiles and small non-coding RNAs. She identified signature differentially expressed genes including known glioma oncogenes and novel candidates, and built a glioblastoma pathway to visualise changes, which identified the up-regulation of inflammatory genes in the glioma lines, suggestive of an immune-evasion phenotype.

Support

Our support teams provide foundational services for teams throughout EMBL-EBI, assisting with training, web content and development, communications, public affairs, administration and, crucially, IT support.

In 2012 Lindsey Crosswell led the newly formed External Relations team, responsible for communications and advocacy, while Cath Brooksbank took on strategic management of the expanding user-training programme. External Services was split into two teams, with Rodrigo Lopez heading web production activities such as development and maintenance of web-based tools, and new team leader Brendan Vaughan preparing to steer the new web development team, responsible for the external-facing website.

Designed to be used

The redesign of the EMBL-EBI website was one of the biggest projects undertaken at the institute during 2012. The strategy focused on keeping users at the centre of the process in order to provide an intuitive interface that encourages people to explore our many bioinformatics services. The new website features consistent functionality while retaining individual resource branding. During the process all EMBL-EBI services moved to secure access (https), and the global search tool was integrated into several core data resources.

Feet on the ground and heads in the cloud

Petteri Jokinen's Systems and Networking team built a private cloud in 2012. The new ELIXIR Embassy Cloud enables external organisations to perform secure data analyses on large datasets using virtual machines. The team created several other virtualised environments, both in our London data centres and on the Genome Campus, allowing us to run web servers, databases and other services from virtual machines. These services are economical in terms of disk space and power usage, providing a clean, green and expandable basis for future operations.

A growing audience

Our training programme continued to evolve in response to demand from emerging research communities, and provided a curriculum covering the full spectrum of EMBL-EBI activities. Train online, the e-learning resource launched in 2011, served over 33 000 unique IP addresses in its first full year. Another challenge addressed by the Brooksbank team was the discoverability of biomedical science training courses. EMBL-EBI entered the realm of 'training informatics' through its involvement in EMTRAIN, which launched a new resource: on-course® for course seekers in June 2012.

The ENA and Functional Genomics teams expanded their training offerings significantly in 2012 to cope with demand from both large-scale sequencing centres and research groups incorporating next-generation sequencing into their portfolio of methodologies. New courses were offered in –omics-based data analysis for plant biologists and in combined experimental/computational metagenomics. Meanwhile, our Industry Programme expanded its geographical coverage, running workshops in the US and Japan in addition to a packed programme in Hinxton, Cambridge, UK.

The publication of the ENCODE papers sparked unprecedented media interest in EMBL-EBI from around the world, as Joint Associate Director, Dr Ewan Birney led the analysis and helped ensure that

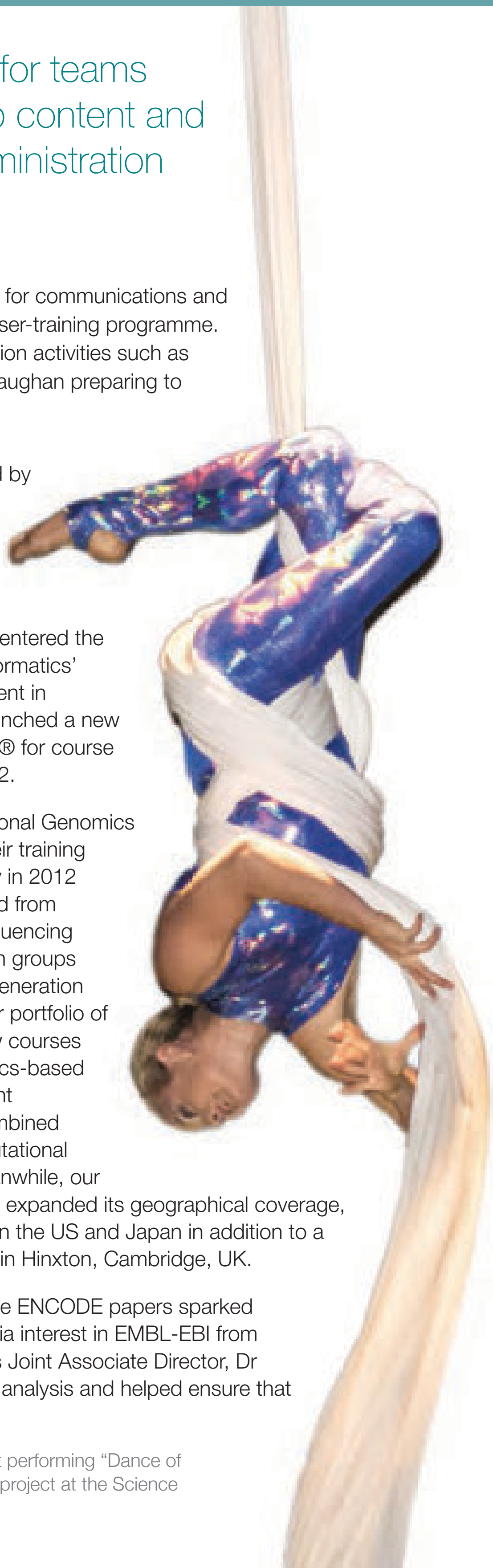


Figure 1. Aerial acrobat performing "Dance of DNA' for the ENCODE project at the Science Musium, London

Nature could deliver a publishing model on a scale appropriate to the project's overwhelming output. Media coverage of the story was universal: 12 935 news stories online, 457 of which were in high-tier publications and made front-page news in the Guardian and New York Times, with feature articles in magazines such as the Economist and countless discussions generated. External Relations acted as liaison and facilitator throughout.

Building bridges

As ELIXIR entered the final stage of its Preparatory Phase in 2012 a further five nations signed the ELIXIR Memorandum of Understanding, bringing the total number of signatories to 15. EMBL-EBI lent close support to the efforts of the ELIXIR Interiim Board, enabling the agreement of interim operating budgets for the ELIXIR Hub for 2012 and 2013 and the representation of ELIXIR within the EU institutions. A founding Director was recruited and in 2013 Dr Niklas Blomberg of AstraZeneca R&D Mölndal, Sweden will take up his post.

The BioMedbridges project is integrating biomedical science research infrastructures to ensure that all of the data in these diverse undertakings can be queried in a consistent way. The project began in January 2012 and is building technical infrastructure, shared standards, and secure access to personally identifiable data. Use cases span the integration of data for personalised medicine, and bridging multiple scales of data visualisation, from molecules to whole organisms.

Serving Life-science Information for the Next Generation (SLING) reached its successful conclusion in August 2012. An integrating activity funded by the European Commission under FP7, SLING sought to ensure that European scientists are optimally equipped to exploit biomolecular information. It delivered a training programme reaching more than 1000 delegates at 33 different host institutes throughout Europe, and brought together a community of trainers that continues as GOBLET, a global foundation for bioinformatics education and training.

Figure 2. Course attendee Diana Legzdina

Training

As part of EMBL-EBI's mission to serve the research community, the Training Team co-ordinates EMBL-EBI's training and scientific outreach activities.

We empower users and potential users of EMBL-EBI's data resources and tools; inform potential staff, students and visitors; and reach out to educators. Our training programme enables scientists at all career stages to make the most of Europe's core biological data resources and liaises with the wider bioinformatics training community. We engage with new user groups in academia and in industry; showcase careers; and provide coherence to EMBL-EBI's growing portfolio of external training.

Summary of progress

- Managed the successful reorganisation of the Outreach and Training team into two new teams: Training and External Relations;
- Actively involved more than 150 members of personnel in 232 events, reaching an audience of >9770 people in 35 countries on six continents;
- Supported our colleagues to create new courses in Train online, EMBL-EBI's free, web-based training resource;
- Supported a group of bioinformaticians from Australia to develop and deliver their own next-generation sequencing data-analysis courses based on the EMBL-EBI model;
- Worked to develop a pan-European framework for continuing professional development for the biomedical sciences (the LifeTrain initiative);
- Contributed to the development and launch of on-course®, the EMTRAIN course portal, and continued to update its catalogue of short courses;
- Contributed to the formation of GOBLET, a new global foundation for bioinformatics educators and trainers;
- Launched the EMBL-EBI ambassador scheme to support our personnel to raise awareness of EMBL-EBI to the scientific community;
- Worked with the External Services Team to redevelop EMBL-EBI's training web pages and contribute more broadly to the EMBL-EBI website redesign project.
- Developed a new system for recording and reporting on EMBL-EBI training activities.

Major achievements

The reorganisation of EMBL-EBI's outreach and training teams into Training (led by Cath Brooksbank) and External Relations (led by Lindsey Crosswell) reflects the success of our engagement with a wide range of different audiences, and enables EMBL-EBI's training, outreach and communications to develop further in the future. We transitioned smoothly into the new structure.

The user base of Train online, EMBL-EBI's free, web-based training resource, has swelled significantly in its first full year since launch. In 2012 we had 33 000 unique users (based on unique IP addresses, which may represent all users at a single institute). More than half of our users are return visitors. We worked with our colleagues in the service teams to create new content for Train online. All our major data resources now have quick tours; ENA, IntAct and Reactome have long courses; and we have a new course format based on videos and tutorials from our face-to-face courses.

In 2012 we ran a large number of events and exhibitions, including training at EMBL-EBI, off-site training throughout the world, conference exhibitions, careers fairs, workshops and other events. New courses included one on –omics-based data analysis for plant biologists and a combined experimental/computational metagenomics course at the Advanced Training Centre in Heidelberg, both of which were well received and will be repeated in 2013. We trained medical geneticists and aquaculturists in the Faroe Islands,

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.

biotechnologists in Japan, nutrition scientists in London, molecular biologists in Thessaloniki, plant scientists in Norwich and many more. We are indebted to our trainers – in other EMBL-EBI teams and from further afield. We are also extremely grateful to the hosts of our external events, who put a huge amount of effort into ensuring that these run smoothly and meet the needs of their local trainees.

We co-ordinated several public engagement events, including an evening of lectures and discussions on the science of sporting success shortly before the 2012 Olympics and a learning lab for science teachers on making sense of biological data. Our interactions with the Young Rewired State were quite memorable: four young coders joined us for a day and a half, during which they wrote a game based on genetic variation. Their achievements were extremely impressive.

We are a partner in EMTRAIN, an IMI project to establish a pan-European platform for professional development covering the whole life cycle of medicines research. A large part of our work in 2012 was on the LifeTrain initiative, which is working towards a mutually recognised framework for continuing professional development in the biomedical sciences. We worked with course providers, representatives from the ESFRI biomedical science research infrastructures, professional bodies, employers and others to agree on a set of principles that will enable course seekers to find high-quality training, course providers to gain meaningful recognition for developing high-quality training and employers to find and develop staff. We contributed to the launch of on-course®, EMTRAIN's comprehensive online course catalogue, which now contains more than 1100 short courses.

The EMBL-EBI ambassador scheme was launched in 2012, providing a mechanism for any member of EMBL-EBI personnel to propose their ideas for raising awareness of EMBL-EBI to the scientific community. The strongest ideas are supported through coaching, provision of materials and a small financial contribution from the Training Team. Our 'train the trainer' project, which we ran with BioPlatforms Australia, supported trainers from Australia to develop and deliver their own training on next-generation sequencing data analysis, resulting in well-received courses in Sydney and Melbourne, with further courses planned in Perth and Adelaide.

Future plans

We anticipate that Train online will become an increasingly important part of our training portfolio; over the coming year we will continue to add new content, including video-based courses modelled on some of our most in-demand face-to-face training.

We look forward to working with our many external collaborators to continue delivering both established and new hands-on courses to our users. We are especially excited about working with the ESFRI biomedical science research infrastructures to deliver new courses ranging from cheminformatics for screening (with EU-Openscreen) to marine metagenomics (with EMBRC). Setting these new endeavours in the context of the LifeTrain framework will help us to serve both our industrial and our academic users better, and underlines our passion for excellence in training.

Selected publications

Pavelin, K., et al. (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comp Biol* 8, e1002554.

Klech, H., et al. (2012) European initiative towards quality standards in education and training for discovery, development and use of medicines. *Eur J Pharm Sci* 45, 515-520.

Figure. Learning lab

Industry programme

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

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

Summary of progress

- Organised regular quarterly strategy meetings in Hinxton and one in Heidelberg;
- Ran eight workshops on topics prioritised by partners, including one conducted in the US and a scientific retreat co-sponsored by Wellcome Trust Scientific Conferences;
- Organised an information workshop and hands-on training for small and medium-sized enterprises (SMEs) in Barcelona, Spain;
- Organised a symposium in Tokyo on information services in biotechnology for representatives of the Japanese pharmaceutical industry and their collaborators.

Major achievements

A major remit of our programme is to foster precompetitive projects. Our knowledge-exchange workshops provide members with opportunities to identify and document shared needs they consider to be pre-competitive. In 2012, EMBL-EBI staff worked with several industry partners to extend the Minimal Information About a Bioactive Entity standard (MIABE; Orchard, S., et al., 2012). As members of the Pistoia Alliance and active participants in IMI projects, EMBL-EBI is a key member in several public–private partnerships (see list) and in 2012 the Industry Programme continued to play a supporting role.

Our quarterly strategy meetings were well attended, and provided opportunities for members to learn first hand about emerging developments at EMBL-EBI and to prioritise future activities. One of these meetings was held in Heidelberg and featured a talk by EMBL Director General, Prof. Iain Mattaj, while another incorporated a workshop on the informatics needs for the agri-food industry. The scientific retreat we co-organised with the Wellcome Trust Scientific Conferences, was well received, as were our eight knowledge-exchange workshops covering topics prioritised by industry members (see Table). We ran one training workshop on Bioconductor and R specifically for our industry partners. Reaching out to member companies with travel limitations, we organised our first event in the US: the 1000 Genomes Project workshop, hosted by Novartis in Cambridge, MA.

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

SMEs are major drivers of the economy. Our programme encourages individuals from SMEs to take advantage of EMBL-EBI training, services and support and organises an annual workshop showcasing freely available tools and information resources that can add value to their business processes immediately. The 2012 SME workshop was held at the Barcelona Biomedical Research Park in collaboration with BioCat, the Spanish National Bioinformatics Institute and co-host UPF Research Programme on Biomedical Informatics. Workshop topics were selected following consultation with a regional focus group and included both EMBL-EBI resources and bioinformatics services provided by Spanish institutes.

We reached out to our established network of pharmaceutical industry contacts in Japan, which include programme member Astellas Pharma Inc., by organising a symposium at the British Embassy in Tokyo for representatives of the Japanese pharmaceutical industry and their collaborators.

Dominic Clark

PhD in Medical Informatics, University of Wales, 1988. Imperial Cancer Research Fund, 1987–1995. UK Bioinformatics Manager, GlaxoWellcome R&D Ltd., 1995–1999. Vice President, Informatics, Pharmagene, 1999–2001. Managing Consultant, Sagentia Ltd., 2001-2009.

At EMBL-EBI since 2006 (secondment 2006-2009).

This symposium was hosted and partly funded by The British Foreign and Commonwealth Office (FCO), and featured presentations by the European Patent Office, the Japanese Patent Office, DDBJ and EMBL-EBI staff with a focus on dissemination of information relating to patented sequences and compounds.

Future plans

We see our interactions with industry partners growing even stronger as the flood of data continues to rise and industry's need 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 2013, EMBL-EBI will continue its participation in IMI projects and will seek out opportunities for intercontinental knowledge-exchange workshops.

The completion in 2013 of the new South Building, which will house the ELIXIR Technical Hub, will present opportunities for working with larger companies in new ways, for example through secondments. We will continue to organise the popular SME forum events jointly with regional industry organisations affiliated with the Council of European BioRegions (CEBR), and to conduct them both in the new building and at host institutes throughout Europe.

Selected publications

Orchard, S., et al. (2012) Shouldn't enantiomeric purity be included in the 'minimum information about a bioactive entity? Response from the MIABE group. *Nat Rev Drug Disc* 11, 730.

Hardy, B., et al. (2012) Food for thought: a toxicology ontology roadmap. *ALTEX* 29, 129-137.

Hardy, B., et al. (2012) Toxicology ontology perspectives. *ALTEX* 29, 139-56.

Innovative Medicines Initiative projects

- eTOX is developing innovative in silico strategies and novel software tools to better predict the toxicological profiles of small molecules in early stages of the drug development pipeline.
- EMTRAIN is a platform for education and training covering the whole life cycle of medicines research, from basic science through clinical development to pharmacovigilance.
- DDMoRe, the Drug Disease Model Resources consortium, is developing a public drug and disease model library.
- EHR4CR is designing a scalable and cost-effective approach to interoperability between electronic health record systems and clinical research.
- EU-AIMS is a large-scale drug-discovery collaboration that brings together academic and industrial R&D with patient organisations to develop and assess novel treatment approaches for autism.
- EMIF: Aims to develop a common information framework of patient-level data that will link up and facilitate access to diverse medical and research data sources.

Industry workshops

- Using electronic health records (EHRs) for translational bioinformatics
- Chemogenomics
- 1000 Genomes Project
- R & Bioconductor training workshop
- Metabolomics
- Antibody informatics
- Systems biology for toxicology pathways
- 1000 Genomes and NGS data analysis (held at Novartis in Cambridge, US)
- Open source software for systems, pathways, interactions and networks (scientific retreat co-funded by Wellcome Trust Scientific Conferences)
- Pre-clinical safety data (held at EMBL Heidelberg)

EMBL-EBI Industry Programme members

Astellas Pharma Inc.

AstraZeneca

Bayer Pharma AG

Boehringer Ingelheim

Eli Lilly and Company

F. Hoffmann-La Roche

Galderma

GlaxoSmithKline

Johnson & Johnson Pharma. R&D

Merck Serono S.A.

Nestlé Research Centre

Novartis Pharma AG

Novo Nordisk

Pfizer Inc

Sanofi-Aventis R&D

Syngenta

UCB

Unilever

External relations

As a European Intergovernmental scientific organisation, EMBL-EBI has a broad reach throughout Europe and internationally. The role of the External Relations team is to communicate effectively with EMBL-EBI's diverse audiences, translating complex scientific ideas to make them easily comprehensible and disseminating them through a wide range of media and communications channels.

Our team manages EMBL-EBI's relationships at a political and policy level with a broad range of stakeholders throughout the world. Through engagement with representatives of the EU, scientific institutions, ministries, funding bodies and policy makers in the EMBL member states and beyond, we provide support for EMBL- EBI's research, its services and the collaborative projects it co-ordinates.

Summary of progress

- Played a key role in the design, architecture, testing and co-ordination of the new EMBL-EBI website, managing the overall content of the global site;
- Managed extensive television and media coverage for EMBL-EBI on the publication of the ENCODE papers;
- Helped secure signature of five nations to the ELIXIR Memorandum of Understanding, bringing the total number of signatories to 15;
- Undertook a comprehensive programme of advocacy with EU to influence funding allocation under Horizon 2020;
- Generated 32 news stories, 12 of which were press releases; coverage in media outlets: 1471 online articles about EMBL-EBI (12 935 about ENCODE).

Major achievements

During 2012 an extensive internal reshaping of teams took place at EMBL-EBI to enable our dynamic, fast-growing Institute to respond efficiently to changing needs and priorities in service provision. These changes resulted in a significant portion of the Institute's outreach activities, namely its communications, printed publications, web content and social networking activity, being moved to be managed by the External Relations team. These activities are entirely complimentary to the political engagement activities undertaken within the team. Additionally, a new staff member joined the External Relations team, having responsibility for graphic design for both web and print. Bringing design activities in house resulted in great efficiencies in time and cost, and ensures consistent high-quality branding and design.

Throughout the year, the focus of political engagement and lobbying activity has remained firmly on ELIXIR in the last of year of its Preparatory Phase and has brought about important progress. A further five member states have signed the ELIXIR Memorandum of Understanding, bringing the total number of signatories to 15. Our team supported the Interim Board countries in their ELIXIR Node submissions.

External Relations showcased ELIXIR at scientific conferences in Denmark, Switzerland, Spain, France and Ireland in 2012. In parallel, we undertook considerable lobbying efforts with EU Institutions, both on behalf of ELIXIR and alongside representatives of the other ESFRI biological and medical science (BMS) projects. This included the preparation of a joint positioning paper for the ESFRI BMS Research

Lindsey Crosswell

BA Hons , London University. BP plc, Government and Public Affairs Manager, 1995–2003. Head of External Relations, Chatham House, Royal Institute of International Affairs 2000–2003 (secondment), Director of Development, Oundle School 2004–2008.

At EMBL-EBI since 2011.

Infrastructures and direct meetings with key opinion formers within the EU, with the aim of increasing the budget allocation for ELIXIR under Horizon 2020.

The team facilitated a number of high-level visits and delegations to EMBL- EBI throughout the year, sharing the Institute's work with interested parties and furthering their engagement with EMBL, ELIXIR member states and potential future members. These visits included scientific delegations from Russia and India, representatives of funding bodies, senior Government officials and numerous journalists and film crews from Europe and beyond.

The publication of the ENCODE papers sparked unprecedented media interest in EMBL-EBI from around the world, as Joint Associate Director, Dr Ewan Birney led the analysis and helped ensure that Nature could deliver a publishing model on a scale appropriate to the project's overwhelming output. In addition to drafting consortium press materials and keeping communication channels open with all 30 consortium press offices, our team worked with the Science Museum in London to produce an exhibition about ENCODE for a large-scale press conference and public engagement activity. Coverage of the ENCODE publications was universal (12 935 news stories online; 457 of these in high-tier publications), with front-page features in the Guardian and New York Times, feature articles in magazines such as the Economist and countless discussions generated on news websites. Our team acted as liaison and facilitator throughout the process.

Future plans

In May 2013 the External Relations Team will welcome Dr Niklas Blomberg as the first ELIXIR Director. We will work alongside Niklas in lending comprehensive support to ELIXIR in the first year of its implementation phase. These efforts will include helping to progress the European Consortium Agreement through the process of ratification by the individual member states, leading to the Agreement becoming effective upon receipt of the first five national signatures. We will support the process of Node submissions by the ELIXIR member states, with the first Nodes coming into effect upon the signature of the European Consortium Agreement. We also anticipate continuing to engage potential future signatories to the ELIXIR Memorandum of Understanding and encouraging Node submissions from new ELIXIR member states.

In the autumn we will celebrate the official opening of the EMBL-EBI South building, which will house ELIXIR, with a formal ceremony to which we will invite representatives from our many stakeholders across Europe. We anticipate that the first calls under the Horizon 2020 funding programme will be issued in late 2013 and we will lead the preparation of detailed responses on behalf of EMBL-EBI with the aim of maximising funding.

We will continue to reach out to the media, and to help our colleagues engage journalists by conveying the high-impact and inspiring nature of their work. To maintain regular contact with our many audiences, we will continue to have a strong and evolving presence on various social media platforms. We will produce printed materials regularly for target audiences, for example our Guide to Bioinformatics Resources and the Annual Scientific Report, but will move to a more digital focus. We will put considerable effort into maintaining the content of the global EMBL-EBi website and helping our colleagues to create appealing, concise content for their visitors.

Selected publication

Crosswell, L.C. and Thornton, J.M. (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol* 30, 241-242.

External services

Our team manages the EMBL-EBI web infrastructure, delivers platforms for web service development and provides robust, secure frameworks for deploying public bioinformatics services.

We also develop and maintain the global EBI search engine, the job dispatcher framework and corresponding SOAP/REST web services for programmatic access. We use DRUPAL for our web development and facilitate access to project management services such as document management (Alfresco), project documentation and tracking (Confluence and JIRA), source code control (CVS, SVN and GIT) and user support (RT).

Summary of progress

- Developed a new EMBL-EBI website;
- Enabled secure access to all EMBL-EBI services;
- Handled a sustained increase in the usage of EBI services;
- Integrated the global EBI Search and EBI web services within several data resources;
- Unified our web portal technology with DRUPAL.

Major achievements

During 2012 we focused on the development of a new global EMBL-EBI website, to be launched in early 2013. This carefully orchestrated effort involved all teams and groups. We used the open-source web content management system DRUPAL and carried out extensive training and testing for both the website and a new Intranet. We drew on user experience design (UXD) techniques throughout the project, and this proved invaluable for creating a unified approach and facilitating collaboration throughout the institute. Usability testing has become a routine component of all web projects.

Our new intranet, launched in 2012, provides up-to-date internal resources and administrative information for staff at all levels. We also manage project- and document-management tools and have made Confluence, Jira and Alfresco available throughout the institute. Confluence currently has more than 60 work ‘spaces’, Jira tracks more than 200 projects and Alfresco has more than 20 sites and some 300 registered users in the system. We worked with the Systems and Networking team to deploy single sign-on, and engaged in European efforts to allow federated authentication: amongst these HAKKA, an ELIXIR-driven collaboration with CSC Finland.

Secure access to services

In 2012 we completed the implementation of secure access for all sites under the ebi.ac.uk domain. Users can access EMBL-EBI services using unencrypted URLs (http) as well as encrypted/secure ones (https). We began the process of determining which activities should be served using secure, encrypted access as the default.

Increased usage of EBI services

We are happy to report that the migration and establishment of new services in the London data centres is now a ‘business as usual’ endeavour. In 2012 our sequence-analysis tools saw marked growth in usage. Programmatic access to our resources using web services proved very popular, with 80% of all job requests (for services such as BLAST, FASTA and InterProScan) using the SOAP and REST programmatic interfaces. Consumers of these tools include uniprot.org, ensemblgenomes.org, Interpro and many labs based in Europe and the US. They are also heavily used by tools such as Blast2GO, BlastStation, STRAP, T-Coffee, CCP4, Geneious and GMU-metagenomics. The average number of jobs per month was four million (up from three million in 2011). The number of datasets available for sequence searching reached over 5000, including many species, strains and assemblies from Ensembl Genomes.

Continuous integration of the EBI Search engine

Our team is responsible for the EBI Search engine web services, which are used in the European Nucleotide Archive (ENA)/EMBL-Bank, ENA’s Sequence Read Archive, Ensembl Genomes, the Enzyme Portal and the Locus Reference Genomic sequences resource (LRG). InterPro, Pombase and MetaboLights will begin using EBI search services in 2013.

Rodrigo Lopez

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

At EMBL-EBI since 1995.

We are also helping to integrate our search services into third-party portals, as with JDispatcher tools. We develop this system in agile style and in 2012 began implementing new features such as predictive text, auto-suggestion and faceted navigation of results. Speed is of the essence, and we explored ever-faster ways of indexing EMBL-EBI data and keeping mean response times under 500ms.

Web portals

UXD methodologies are an integral part of our work process, and continue to prove beneficial to both the organisation and scientific users. This helps us simplify design by removing unnecessary features, optimising functionality, streamlining development efforts and integrating business-process-driven goals. UXD has been an essential communication tool and a driving force behind the new global EMBL-EBI website.

Our team is responsible for the maintenance of more than 30 project portals in addition to the global website. In 2012 we welcomed the Bioinformatics Training Network, EMBL-EBI's Train online, the EMBL Staff Association, BioMedBridges, HipSci, RDF and the EBI Intranet to our growing portfolio. Our Help Desk handled more than 1350 tickets, offering assistance for programmatic access and data acquisition; technical help for users experiencing problems with services; consultation on best practices; and training on various resources. Internally, we handled over 4000 requests regarding all aspects of web development and production.

Outreach and training

In 2012 we participated in 16 different events, including ISMB in California, US; seven SLING dissemination events in Europe; Patenting and Industry programme seminars in Tokyo, Japan; Hands-on training workshops at Harvard University in Boston, US; and on-site training courses.

Future plans

The deployment of a new, user-driven, global website is a major undertaking, and in late 2012 we changed the composition of our team to reflect the importance of this activity: Brendan Vaughan was appointed as new Team Leader for Web Development and will take up this role in 2013. Rodrigo Lopez continues as Team Leader for Web Production. Both teams will review their working practices to ensure that EMBL-EBI remains at the vanguard of web development. Our platform-oriented approach already provides the fundamental building blocks for ELIXIR, and we will continue to develop web infrastructure solutions that allow our services to interweave their data for optimum discoverability and utility. In 2013 we will work to ensure the EBI Search provides sophisticated results from a simple-to-use interface. We aim to make it easy for users to explore biological data at EMBL-EBI in novel and truly meaningful ways.

Selected publications 2012

Li, W., et al. (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics* 28, 1650-1651.

Clarke, L., et al. (2012) The 1000 Genomes Project: data management and community access. *Nat Methods* 9, 459-462.

Robinson, J., et al. (2012) The IMGT/HLA database. *Nucleic Acids Res* 41, 1–6.

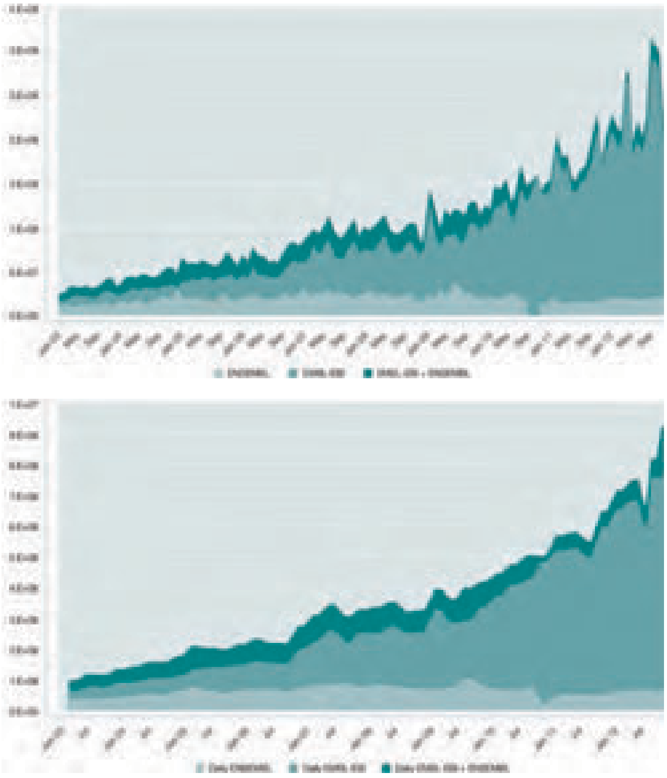


Figure. Usage statistics and supporting the use of EBI Services. (a) Jobs per month since 2005 showing Web browser usage, programmatic access using SOAP/REST web services and total for tools such as Blast, Fasta, ClustalOmega and InterProScan. (b) Quarterly requests per day for www.ebi.ac.uk and www.ensembl.org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

Systems and networking

Our team manages EMBL-EBI's IT infrastructure, which includes compute and database servers, storage, desktop systems and networking.

We also provide database administration (Oracle, MySQL, etc), manage the campus Internet connection and support EMBL-EBI staff in their daily computer-based activities. We work closely with all project groups, maintaining and planning their specific infrastructures.

Major achievements

Clouds and virtualisation

We built the private, VMware-based ELIXIR Embassy Cloud for hosting external organisations. It allows users to run Linux or Windows virtual machines in a secure virtual data centre of their own, and provides an excellent network connection to EMBL-EBI services. The ELIXIR Embassy Cloud allows direct mounting of large datasets such as the 1000 Genomes archive and provides high bandwidth and low latency communication with our online services. Communication from Embassy to service is over our internal networks, rather than the public Internet, and allows user organisations to establish a VPN back to their 'home' organisation for secure communication.

We also built several internal virtualised environments for running web servers, databases (Oracle and MySQL) and other services. These virtualised services, in both the London data centres and on the Genome Campus, increase our flexibility and resilience, and have already saved a lot of physical space and electricity. We expect this trend to continue.

Networking

The networking team completed major changes in the Genome Campus network's backbone so that it supports more traffic and enables more flexible routing. We also implemented single sign on (Shibboleth) and authentication (Eduroam).

Storage

We introduced flash-based storage to several key databases, improving their speed significantly. We continued growing our storage infrastructure: the largest single file system, used in the Sequence Retrieval Archive, is now approximately three petabytes. Substantial effort was put into retiring old storage, migrating to new devices and upgrading existing systems. We also evaluated several new storage systems (including object storage) as a foundation for future needs.

Computing

We started moving to a new configuration management platform. The migration of the in-house built system is ongoing but the new infrastructure is ready and a majority of machines are using the new method. We undertook a major effort to improve statistics collection and reporting. We also upgraded the EBI cluster to run Red Hat 6.3.

Databases

We implemented a virtual database solution, deployed an implementation of a 'columnar' database to Hinxton (Cambridge) and London data centres, completed migration of Oracle databases from Solaris to Linux and completed the upgrade of Oracle databases to 11g. We began testing and implementing flash memory array storage for some i/o intensive databases. We also worked on the testing, development and preparation for the migration of Oracle and MySQL in VMWARE environment.

Petteri Jokinen

MSc in Computer Science 1990, Helsinki University.

At EMBL-EBI since 1996.

Data centre capacity framework and London move

We are working closely with EMBL-EBI Administration to create a Data Centre Equipment Framework for the procurement of IT equipment under OJEU regulation. The Framework is operational and managed by the Systems team. We also co-ordinated the migration of EMBL-EBI public services from the Genome Campus Data Centre to the London Data Centres, ensuring clear planning and communication between EMBL-EBI Services groups in order to achieve completion and guarantee continuity of service provision.

Figure. Blades in the Genome Campus data centre.

Administration

The EMBL-EBI Administration Team aims to provide a timely and efficient administrative support network for those working at EMBL-EBI.

Our activities span budgetary, financial and purchasing matters; human resources; grants and external funding management; facilities management (including health and safety) as well as pre- and post-doc programmes. We co-ordinate and integrate administrative activities throughout EMBL-EBI in order to facilitate interactions with the wider scientific community through, for example, organising meetings and courses and arranging travel for our extremely mobile staff.

Summary of progress

- Organised and delivered the administrative aspect of the UK's Large Facilities Capital Fund Programme, which focuses on both construction of a new building on site and equipment/data facilities over the next eight years;
- Continued the development of the budgetary process;
- Continued efforts to attract high-quality staff through targeted recruitment and advertising, and improved their induction into EMBL-EBI;
- Contributed to the implementation of the EMBL grants management database and to the development of new reporting software;
- Contributed to the development and implementation of the new intranet;
- Continued to develop and sustain our Health & Safety practices and procedures.

Major achievements

The Large Facilities Capital Funding Programme consists of two projects: the construction of a new building on the Genome Campus (estimated cost: €28 million) and the acquisition, over the next eight years, of equipment and space in a commercially run data centre (estimated cost: €60 million). The Programme is designed to help meet the growing demand for EMBL-EBI services and, in the context of ELIXIR, support life science research and its translation to medicine and the environment, the bio-industries and society. These funds were committed by the UK government in December 2011 and routed via the Biotechnology and Biomedical Research Council (BBSRC). The new build on campus is on track to open in September 2013. The first Suppliers' Framework Agreement has been established and the first 'mini-competitions' for equipment launched by the Systems and Networking Group.

We have been refining our financial and budgetary process following the introduction of cash budgets, as well as contributing to the development of new EMBL financial reporting tools which will be introduced in 2013.

Our Grants Office has contributed to the development of the EMBL Converis Grants Database Management System which went live in December 2012. The Grants Database enables more information to be captured, analysed and reported on, which should assist applications, grant-holders and the review of overall success rates and statistics. For the first time there will be one central electronic repository of grant related documents.

The Human Resources (HR) team gained a staff member this year, providing an opportunity to work on improving processes and developing internal documentation to ensure we provide accurate and consistent advice to staff and supervisors.

A monthly newcomers meeting for joiners has been introduced, and this is backed up with improved employment information on the intranet.

All processes are kept under review and examples of improvements include the internal database of emergency contact details for all personnel based at Hinxton, Cambridge and the development of a smarter, more intuitive, clearer approval process for opening new posts.

A new EBI intranet was introduced in November. Work undertaken by External Services to explore usability and the introduction of the Drupal environment (and Systems for

Mark Green

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

At EMBL-EBI since 2003.

Shibboleth) has provided a set of pages that are far easier to navigate. Administration played its part by helping to ensure the content is available and easy to find. There is much work still to be done over the coming year to fully optimise the new intranet.

We started work early in the year to create an Alumni Working Group (as promised in last year's report) and a programme of activities for UK-based alumni is being planned by the External Relations team.

The co-operation agreement between the Wellcome Trust Sanger Institute and EMBL-EBI was signed during 2012 and this is, amongst other things, paving the way to help manage some complex transfers and appointments of personnel between our respective institutes.

We continue to participate in campus initiatives such as the 'Sex in Science' seminar series, which addresses important issues such as work–life balance and the 'leaky career pipeline' for women scientists.

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

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

Future plans

We will continue developing longer-term strategic financial plans taking account of EMBL, external and LFCF funding. We will develop processes and procedures to facilitate the establishment of an ELIXIR Hub at EMBL-EBI. We will embed the new EMBL Converis Grants Database Management System and help implement the new EMBL Business Warehouse/Objects software, facilitating analysis and reporting of financial and personnel data. We will continue to add and revise pages for the new intranet, actively seeking feedback from our end users wherever possible. We will develop the quality of guidance provided to Group and Team Leaders. We will continue to develop arrangements for newcomers and overhaul our exit procedures for departing staff. We will continue to maintain good interactions between a wide diversity of stakeholders such as the BBSRC, the Wellcome Trust and the NIH. In respect of LFCF funding we will, together with other colleagues in EMBL-EBI, migrate staff to the new building and start the process of reviewing requirements for Data Centre space acquisition for the period 2015-2019.

Figure. Ground breaking ceremony for the new South Building, which will house the ELIXIR Technical Hub.

Funding and resource allocation

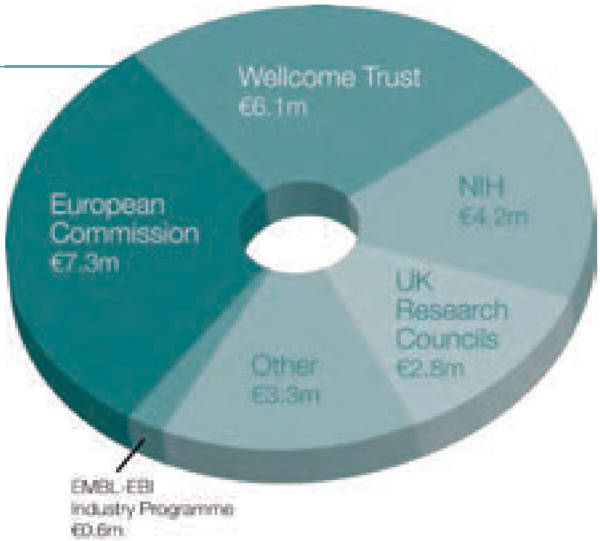
Despite another year of austerity in many countries, our funding and staff numbers remained stable in 2012, with small increases in both internal EMBL and external grant funding. Yet we continued to handle an exponential rise in biological data as well as unprecedented growth in the use of our resources.

The figures reported here exclude overheads on internal funds. External funding represents available funds in 2012.

Sources of funding

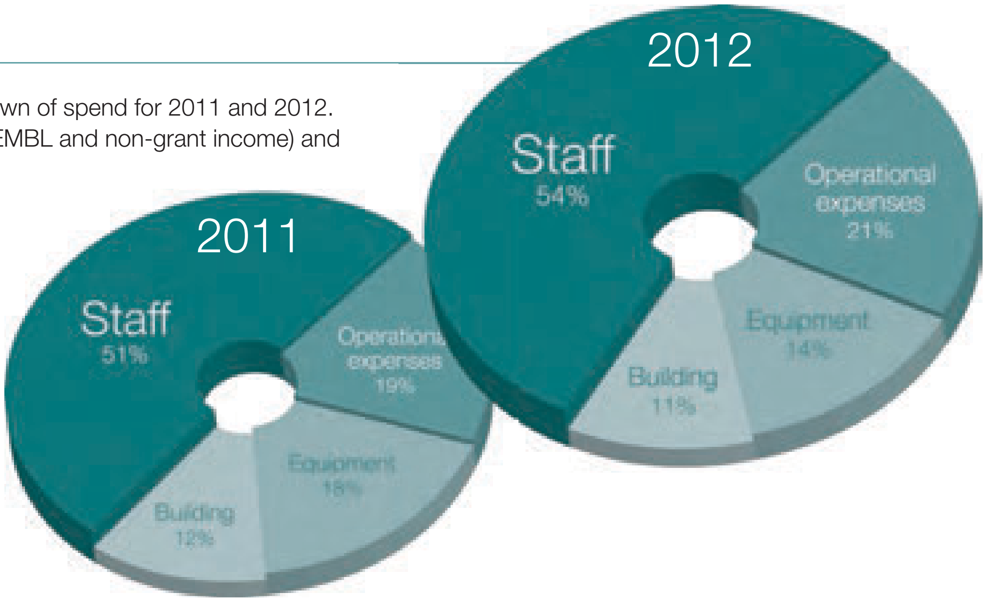
EMBL-EBI is primarily funded by EMBL member states.

Our major sources of external funding include the European Commission (€7.3 million), the Wellcome Trust (€6.1 million), the US National Institutes of Health (€4.2 million), the UK Research Councils (€2.8 million) and the EBI Industry Programme (€0.6 million). We also benefit from a large number of grants from various other sources (total, €3.3 million).



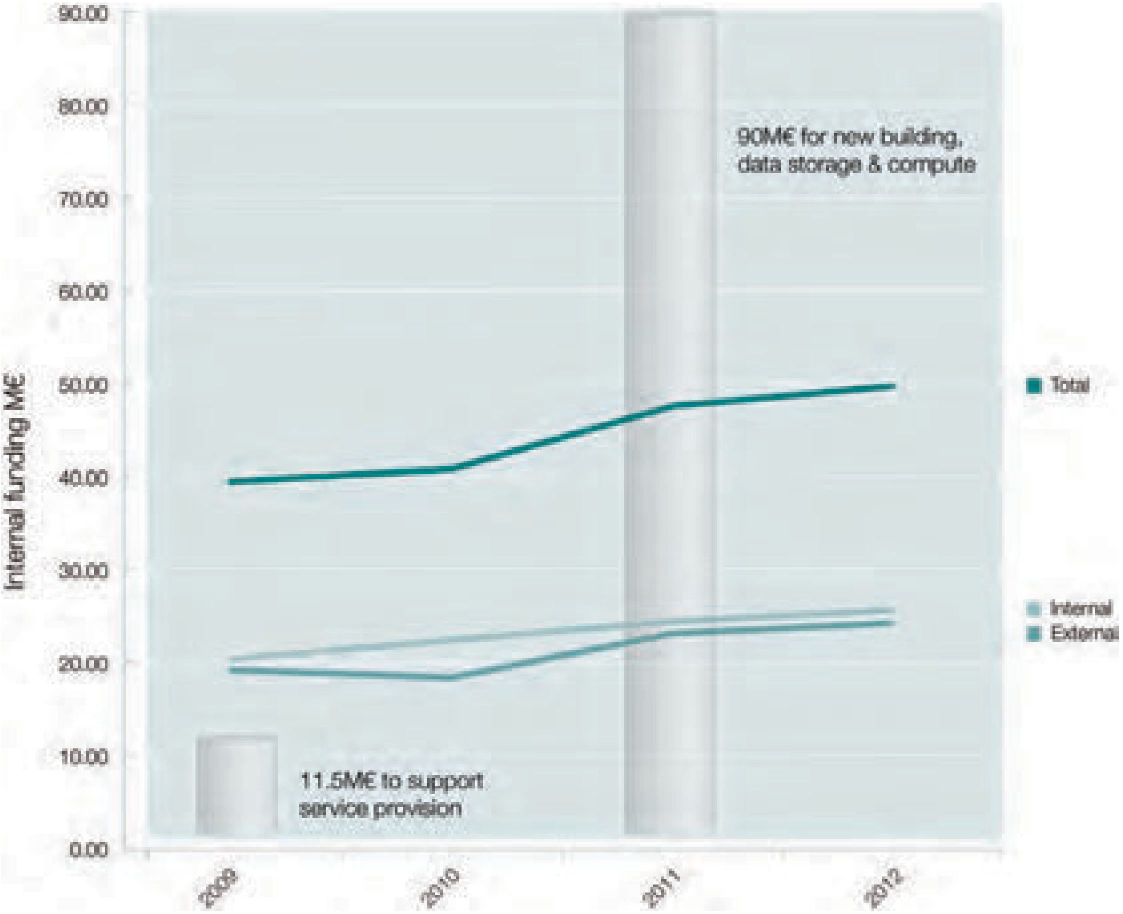
Spending

This figure shows the breakdown of spend for 2011 and 2012. Spending from both internal (EMBL and non-grant income) and external sources combined.



Growth of funding

This figure shows the growth of our total funds over the past four years, from both external sources and EMBL member states (n.b. the EMBL indicative scheme begins in 2012). The figures for 2012 are not yet finalised and we have not included centrally met costs and overheads.



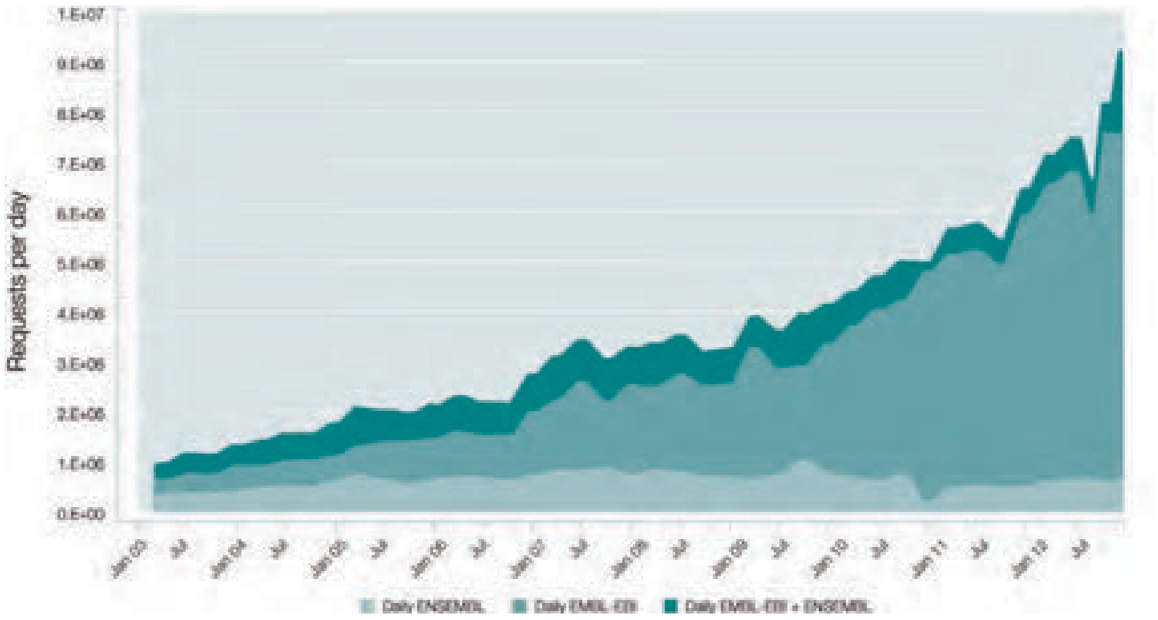
Capital investment

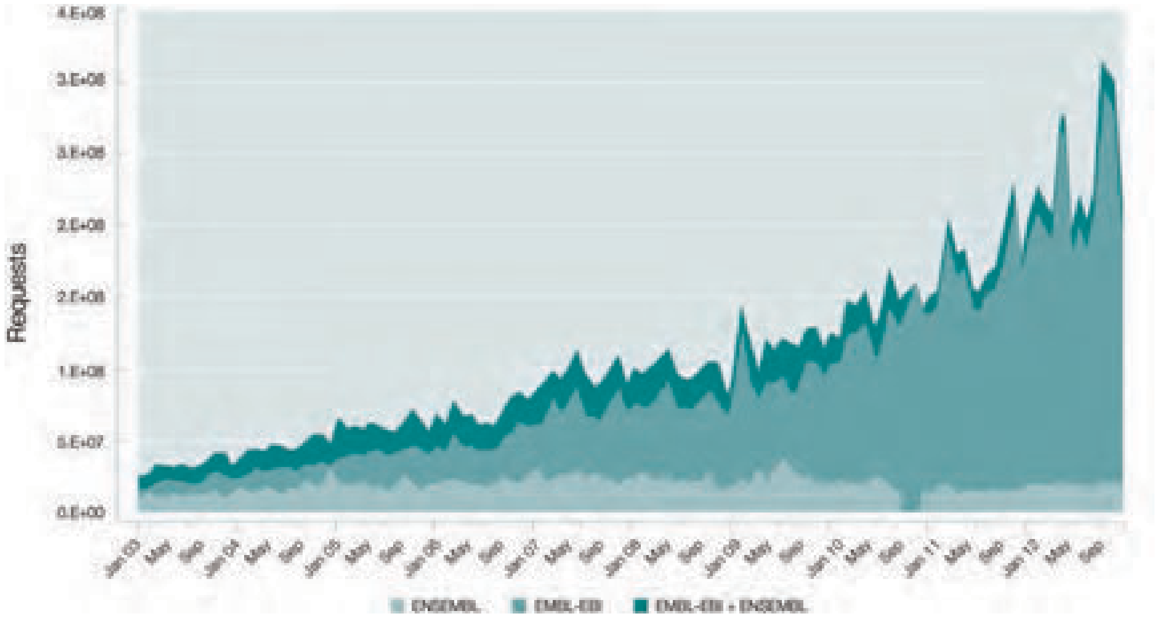
In 2009 we received £10 million (€11.5 million) from the UK government's Large Facilities Capital Fund to enable data service provision, including acquisition of space and equipment. In 2011, the Large Facilities Capital Fund committed a further £75 million (€90 million) to support the acquisition of high-security data centre space and equipment, and to fund the construction of a new EMBL-EBI building on campus, which will house the ELIXIR hub. Some of these funds were spent in 2012 as the new South Building, which will house the ELIXIR Hub and other activities, made rapid progress. Most of these funds will be spent over the next six years to continue supporting high-security data centre space and equipment. The building is scheduled to open in autumn 2013.

Growth of core resources

In 2012 there were on average 7.0 million requests on our services per day, not including Ensembl (compare to 5.3 million in 2011). Including Ensembl, the number of daily requests was 9.2 million on average, compared to 5.7 million in 2011.

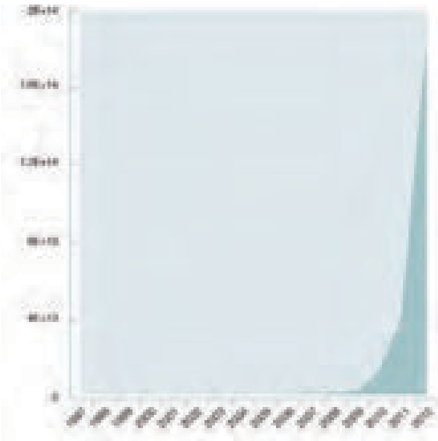
We have seen steady growth in usage and in the number of computers accessing our services. The number of unique IPs, or web addresses, accessing our website (11 000 per day) grew by 16.2% during 2012 (compared to 6.1% growth in 2011); this figure is based on cumulative IP counts for both years. We approach these figures with caution, as an IP address could represent a single person or an entire organisation, so the figures represent a minimum number of users.



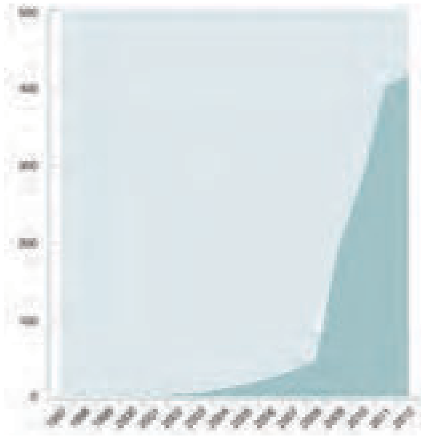


In 2012 all of our core data resources grew substantially. This growth, as in past years, reflects improvements in technology that allow scientists to generate ever more data at ever lower costs. The nucleotide sequence databases, for example have a doubling time of less than one year, which means that more than 50% of our sequence data has been in the archive for less than a year. We expect this growth to continue in future years.

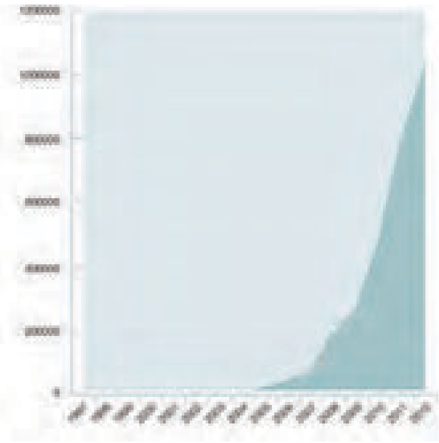
Nucleotide seq.



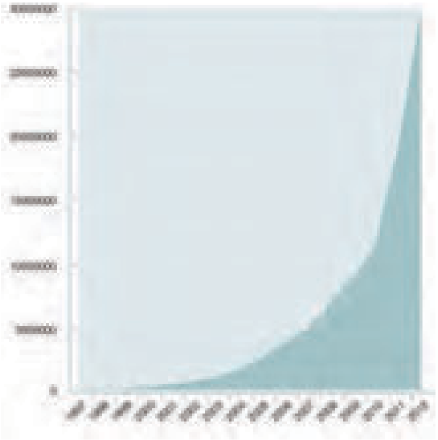
Genomes



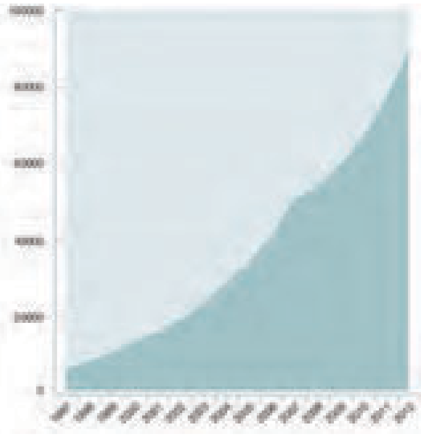
Array Express



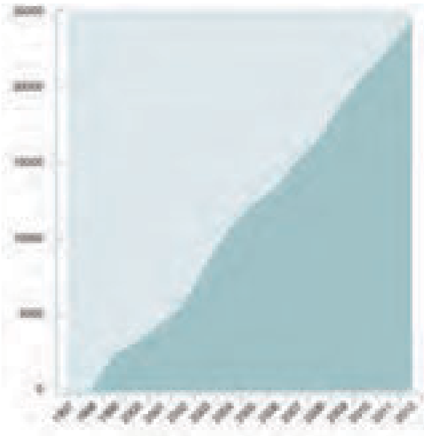
UniProtKB



PDBe



InterPro

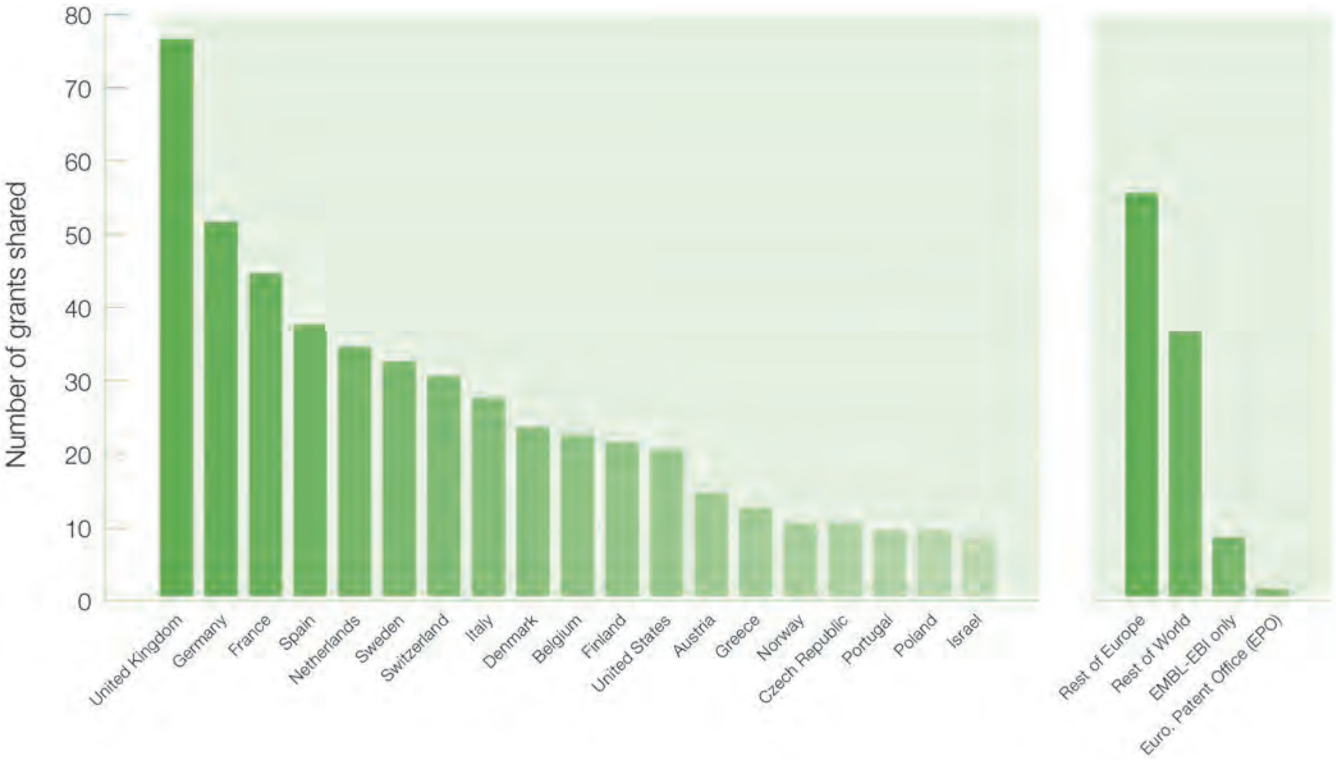


Collaborations

EMBL-EBI is a highly collaborative institute: almost all of our resources are funded through collaborative agreements, and upwards of 90% of our scholarly publications are authored with colleagues at other institutes throughout the world.

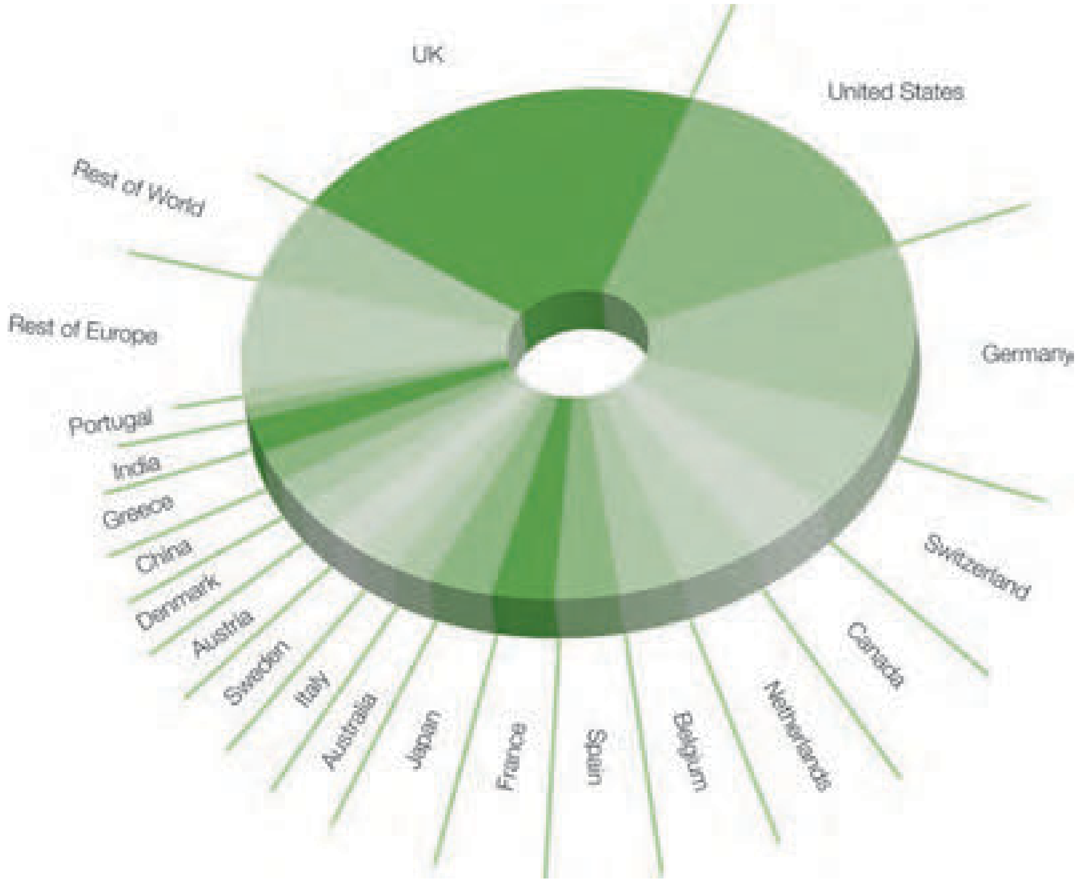
Shared funding

In 2012, EMBL-EBI shared grants with researchers and institutes in 58 countries throughout the world, most notably in the larger European countries but also with colleagues in unexpected places like Burkina Faso. Of 143 grants received, only 8 were exclusively for EMBL-EBI. These figures are potentially underestimated, as all partners are not always listed on grants.



Collaborative publications

In 2012, EMBL-EBI scientists produced 223 scholarly papers, most of which were produced in collaboration with other institutes. Our collaborations are truly global in scope, and in terms of papers published our most productive partnerships have been with partner institutes in the UK, US, Germany, Switzerland, Canada and the Netherlands.



Note: Publication affliation figures are based on an export from the Scopus service, which provides affiliation data for papers.

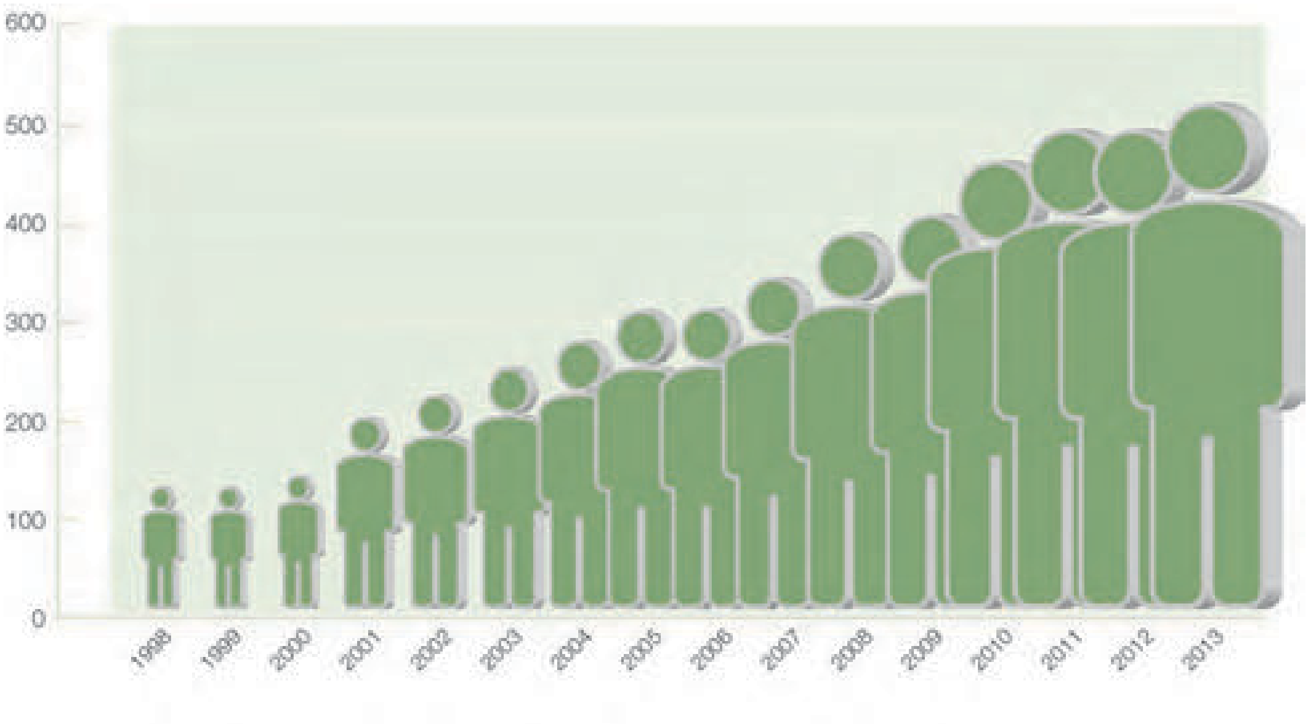
Staff growth

Our organisational structure reflects our mission: services, research, training and industry support, with overarching internal support. We had 499 members of staff in 2012, reflecting no change from 2011.

Major changes in our organisation in 2012 included the retirement of Graham Cameron as Associate Director and the joint appointment of Rolf Apweiler and Ewan Birney as new Associate Directors, responsible for strategic oversight of services. Three of our long-standing Research Group Leaders, Nicolas Le Novère, Nicholas Luscombe and Dietrich Rebholz-Schuhmann, said goodbye to EMBL, and they will be sorely missed.

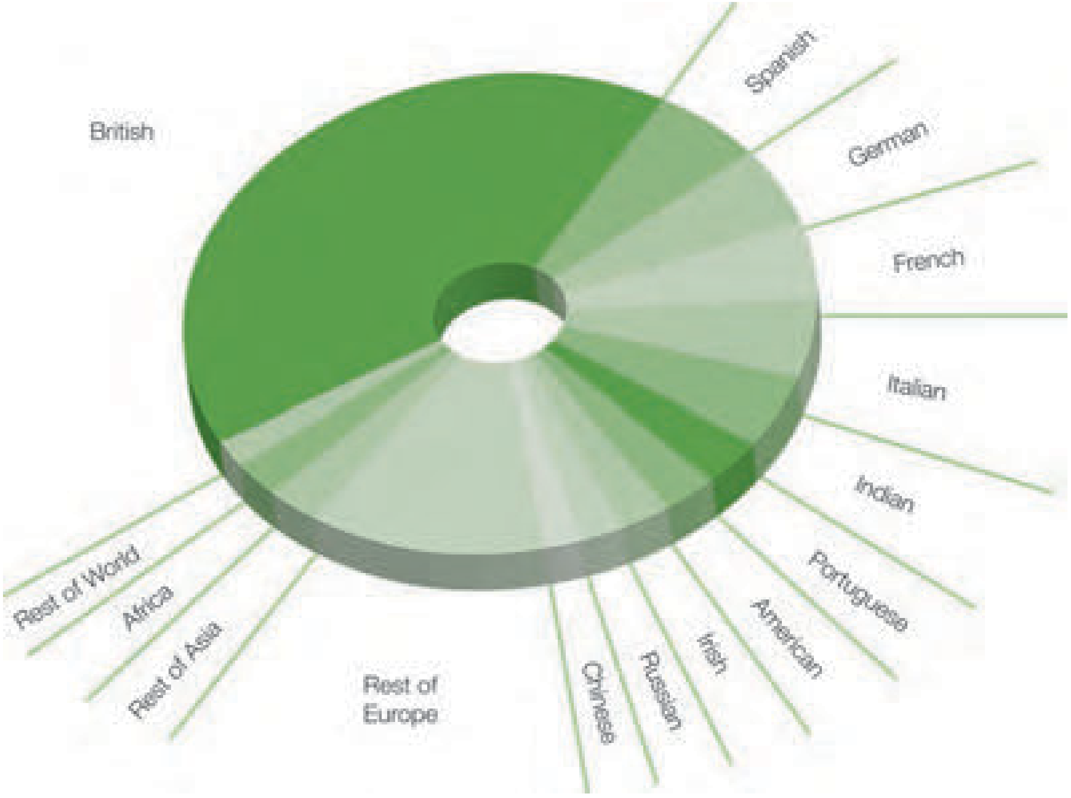
Newly appointed Group and Team Leaders who took up their posts late in 2012 include Alex Bateman, who leads Protein Resources; Justin Paschall, responsible for the newly formed Variation Team; and Oliver Stegle, Research Group Leader. Incoming Research Group Leaders in 2013 include Pedro Beltrao and Sarah Teichmann. We wish them all a warm welcome.

Growth of staff



Staff nationalities in 2012

As a European organisation we are proud to report that our personnel in 2012 represented 53 nationalities (48 in 2011). We welcomed a steady stream of visitors, 58 of whom stayed with us for longer than one month (compare to 47 last year).



Scientific advisory commitees

EMBL Scientific Advisory Committee

Council establishes a Scientific Advisory Committee (SAC), which shall give advice to Council, in particular with regard to proposals from the Director General on the realisation of the programme of the Laboratory.

- Siv Andersson, Sweden (2007-2012)
- Naama Barkai, Israel (2011-2013)
- Konrad Basler, Switzerland (2007-2012)
- Jan Drenth, Netherlands (1979-1984)
- Denis Duboule, Geneva, Switzerland (2011-2013)
- Roderic Guigo, Spain (2012-2014)
- Daniel Louvard, France (2012-2014)
- Ron Milligan, United States (2012-2014)
- Andrew Murray, United States (2008-2010, 2011-2013)
- Andrea Musacchio, Germany (2011-2013)
- Helen Saibil, United Kingdom (2007-2012)
- Sandra Schmid, United States (2008-2010, 2011-2013)
- Titia K. Sixma, Netherlands (2007-2012)
- Michael Snyder, United States (2011-2013)
- Alfonso Valencia, Spain (2007-2012)
- Jean Weissenbach, France (2012-2014)

Bioinformatics Advisory Committee

EMBL-EBI has an established guidance structure in the form of a Bioinformatics Advisory Committee (BAC). The Committee gives advice to the institute with regard to scientific strategy, future directions and proposals on the realisation of its programme.

The BAC is composed of distinguished scientists appointed in their own right, not as representatives of member states. Membership of the Committee is drawn from experts in the field of bioinformatics and other relevant scientific disciplines. Some are also members of EMBL's Scientific Advisory Committee and data resource advisory boards.

2012 Membership

- Philippe Sanseau of GlaxoSmithKline, United Kingdom
- Roderic Guigo of the Centre de Regulacio Genomica, Spain
- Tim Hubbard of the Wellcome Trust Sanger Institute, United Kingdom
- Olli Kallioniemi of VTT Medical Biotechnology, Finland
- Jonathan Knowles of the Roche Group, Switzerland
- Matthias Uhlén of the Royal Institute of Technology (KTH), Sweden

Scientific Advisory Committees - Services

BioModels Scientific Advisory Committee

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

Cheminformatics: ChEMBL and ChEBI Advisory Committee

- Steve Bryant, NIH, United States
- Edgar Jacoby, Novartis, Switzerland
- Andrew Leach, GlaxoSmithKline Plc, United Kingdom (Chair)
- Tudor Oprea, University of New Mexico, United States
- Alfonso Valencia, CNIO, Spain
- Peter Willett, University of Sheffield, United Kingdom

European Nucleotide Archive Scientific Advisory Board

- Mark Blaxter, University of Edinburgh, United Kingdom
- Antoine Danchin, CNRS, Institut Pasteur, France
- Roderic Guigo, Centre de Regulació Genomica, Spain
- Tim Hubbard, Wellcome Trust Sanger Institute, United Kingdom (Chair)
- Jim Ostell, National Centre for Biotechnology Information, United States
- Babis Savakis, University of Crete & IMBB-FORTH, Greece
- Martin Vingron, Max-Planck Institute for Molecular Genetics, Germany
- Jean Weissenbach, Genoscope, France
- Patrick Wincker, Genoscope, France

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

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

Ensembl Genomes Scientific Advisory Board

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

The Gene Ontology Scientific Advisory Board

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

InterPro/Pfam Scientific Advisory Board

- Philip Bourne, University of California, San Diego, United States
- Michael Galperin, National Center for Biotechnology Information, United States
- Erik Sonnhammer, Stockholm University, Sweden (Chair)
- Alfonso Valencia, Structural Computational Biology Group, CNIO, Spain

Literature Services

- Gianni Cesareni, University of Rome, Italy
- Wolfram Horstmann, Bodleian Library, Oxford, United Kingdom
- Tim Hubbard, Wellcome Trust Sanger Institute, United Kingdom (Chair)
- Larry Hunter, University of Colorado Health Sciences Center, United States
- Mark Patterson, eLife, United Kingdom
- Carola Tilgmann, Lund University, Sweden

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

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

Worldwide Protein Data Bank (wwPDB) Advisory Committee

- Stephen K. Burley, Eli Lilly, United States (Chair)
- Jianpeng Ding, Shanghai Institutes for Biological Sciences, China
- Wayne Hendrickson, Columbia University, United States
- Genji Kurisu, Institute for Protein Research, Osaka University, Japan
- Gaetano Montelione, Rutgers University, United States
- Keiichi Namba, Osaka University, Japan
- Michael G. Rossmann, Purdue University, United States
- Helen Saibil, Birkbeck College London, United Kingdom
- Titia Sixma, Netherlands Cancer Institute, the Netherlands
- Soichi Wakatsuki, High Energy Accelerator Research Organisation (KEK), Japan
- Cynthia Wolberger, Johns Hopkins School of Medicine, United States
- Edward N. Baker, University of Auckland, New Zealand (Ex Officio)
- R. Andrew Byrd, National Institutes of Health, United States (Ex Officio)

EMDataBank Advisory Committee

- Joachim Frank, Columbia University, United States (Chair)
- Achilleas Frangakis, Goethe University Frankfurt, Germany
- Richard Henderson, MRC Laboratory of Molecular Biology, United Kingdom
- Maryanne Martone, University of California, San Diego, United States
- Michael Rossmann, Purdue University, United States
- Andrej Sali, University of California, San Francisco, United States
- Paula Flicker, National Institute of General Medical Sciences, United States (Observer)

Reactome Scientific Advisory Committee

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

The Universal Protein Resource (UniProt) Scientific Advisory Committee

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

Major database collaborations

ARRAYEXPRESS

- Dana Farber Cancer Institute, United States
- DDBJ Omics Archive, DNA Databank of Japan, Japan
- Gene Expression Omnibus, NCBI, United States
- Functional Genomics Data Society
- Penn Center for Bioinformatics, University of Pennsylvania School of Medicine, United States
- Stanford Microarray Database, Stanford University, United States

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, United States

ChEBI

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

ChEMBL

- BindingDB, University of California, San Diego, United States
- CanSAR, Institute of Cancer Research, United Kingdom
- PubChem, NCBI, National Institutes of Health, United States

THE EUROPEAN NUCLEOTIDE ARCHIVE

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

- National Center for Biotechnology Information, Bethesda, USA (GenBank, Trace Archive and Sequence Read Archive)
- National Institute of Genetics, Mishima, Japan (DNA DataBank of Japan, Trace Archive and Sequence Read Archive)

Other ENA collaborations:

- Catalogue of Life
- Genomics Standards Consortium

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, United States
- Broad Institute, United States
- DOE Joint Genome Institute, Walnut Creek, United States
- Ensembl at the Wellcome Trust Sanger Institute, United Kingdom
- Genome Browser at the University of California, Santa Cruz, United States
- Map Viewer at the National Center for Biotechnology Information, United States
- Mouse Genome Informatics at the Jackson Laboratory, United States
- Rat Genome Database at the Medical College of Wisconsin, United States
- The Roslin Institute, Midlothian, United Kingdom

ENSEMBL GENOMES

- Gramene at Cold Spring Harbor Laboratory, United States
- PomBase with University College London and the University of Cambridge, United Kingdom
- PhytoPath with Rothamsted Research, United Kingdom
- VectorBase: a collaboration with University of Notre Dame, USA; Harvard University, USA; Institute of Molecular Biology and Biochemistry, Greece; University of New Mexico, USA; and Imperial College London, United Kingdom
- Microme, a European collaboration with 14 partners
- transPLANT, a European project with 11 partners
- WormBase, a collaboration with the California Institute of Technology and Washington University, USA; Ontario Institute for Cancer Research, CA; Wellcome Trust Sanger Institute and Oxford University, United Kingdom

THE GENE ONTOLOGY CONSORTIUM

- Agbase, Mississippi State University, United States
- The Arabidopsis Information Resource, Carnegie Institution of Washington, United States
- Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, United States
- British Heart Foundation, University College London, United Kingdom
- Candida Genome Database, Stanford University, United States
- DictyBase at Northwestern University, United States
- EcoliWiki
- FlyBase at the University of Cambridge, United Kingdom
- GeneDB S. pombe and GeneDB for protozoa at the Wellcome Trust Sanger Institute, United Kingdom
- Gramene at Cornell University, United States
- Institute for Genome Sciences, University of Maryland, United States
- The J. Craig Venter Institute, United States
- Mouse Genome Informatics, The Jackson Laboratory, United States
- Muscle TRAIT, University of Padua, Italy
- Plant-Association Microbe Gene Ontology, Virginia Polytechnic Institute and State University, United States
- Rat Genome Database at the Medical College of Wisconsin, United States
- Reactome at Cold Spring Harbor Laboratory, United States
- Saccharomyces Genome Database, Stanford University, United States
- WormBase at California Institute of Technology, United States
- The Zebrafish Information Network at the University of Oregon, United States

THE IMEX CONSORTIUM

- Centro Nacional de Biotecnologia, Spain
- DIP at the University of California, Los Angeles, United States
- MINT at University Tor Vergata, Italy
- MIPS at the National Research Centre for Environment and Health, Germany
- Neuroproteomics platform of National Neurosciences Facility, Australia
- Shanghai Institutes for Biological Sciences, Shanghai, China

INTERPRO

- CATH-Gene3D at University College London, United Kingdom
- HAMAP at the Swiss Institute of Bioinformatics, Switzerland
- InterPro at EMBL-EBI, United Kingdom
- PANTHER at University of Southern California, Los Angeles, United States
- Pfam at the Wellcome Trust Sanger Institute, United Kingdom
- PIRSF at the Protein Information Resource, Georgetown University Medical Centre, United States
- PRINTS at the University of Manchester, United Kingdom
- ProDom at INRA and CNRS, France
- PROSITE at the Swiss Institute of Bioinformatics, Switzerland
- SCOP at the Laboratory of Molecular Biology, University of Cambridge, United Kingdom
- SMART at EMBL, Heidelberg, Germany
- SUPERFAMILY at the University of Bristol, United Kingdom
- TIGRFAMs at The Institute of Genome Research, United States

THE PROTEIN DATABANK IN EUROPE

PDBe is a partner in the World Wide Protein Data Bank (wwPDB). Other partners include:

- BioMagResBank, University of Wisconsin, United States
- PDBj at Osaka University, Japan
- Research Collaboratory for Structural Bioinformatics, United States

PRIDE

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

Europe PubMed Central

- Europe PubMed Central is part of PubMed Central International. Other database partners include:
- PubMed Central, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, United States
- PubMed Central Canada

REACTOME

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

UNIPROT - THE UNIFIED PROTEIN RESOURCE

- UniProt at EMBL-EBI is part of the UniProt Consortium. Other partners include:
- UniProt at the Protein Information Resource, Georgetown University Medical Centre, United States
- UniProt at the Protein Information Resource, University of Delaware, United States
- UniProt at the Swiss Institute of Bioinformatics, Switzerland

Publications in 2012

In 2012, EMBL-EBI scientists produced approximately 224 scholarly papers, most of which were produced in collaboration with other institutes. Our collaborations are truly global in scope, and in terms of papers published our most productive partnerships have been with partner institutes in the UK, US, Germany, Switzerland, Canada and the Netherlands.

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

Adams, D., Altucci L., Antonarakis S. E., et al. (2012) BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol 30, 224-6.

Adams, R.R., Tsorman, N., Stratford, K., et al. (2012) The Input Signal Step Function (ISSF), a standard method to encode input signals in SBML models with software support, applied to circadian clock models. J Biol Rhythms 27, 328-332.

Adamusiak, T., Parkinson H., Muilu J., et al. (2012) Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search and exchange of phenotype and genotype information. Hum Mutat 33, 873.

Albers, C. A., Paul, D. S., Schulze H., et al. (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet 44, 435–439.

Alcantara, R., Axelsen K. B., Morgat A., et al. (2012) Rhea–a manually curated resource of biochemical reactions. Nucleic Acids Res 40, D754–60.

Alcántara, R., Onwubiko J., Cao, H., et al. (2012) The EBI enzyme portal. Nucleic Acids Res 41, D773–D780.

Allan, C., Burel J. M., Moore J., et al. OMERO: flexible, model-driven data management for experimental biology. Nat Methods 9, 245-253.

Altenhoff, A. M., Studer R. A., Robinson-Rechavi M. and Dessimoz C. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comp Biol 8, e1002514.

Amid, C., Birney E., Bower L., et al. (2012) Major submissions tool developments at the European nucleotide archive. Nucleic Acids Res 40, D43–D47.

Andrade, A. Q., Blondé W., Hastings J. and Schulz S. (2012) Process attributes in bio-ontologies. BMC Bioinform 13, 217.

Andrade, A. Q., Kreuzthaler M., Hastings J., Krestyaninova M. and Schulz S. (2012) Requirements for semantic biobanks. Studies in Health Technology and Informatics 180, 569–573.

Auer, S., Dalamagas T., Parkinson H., et al. (2012) Diachronic linked data: towards long-term preservation of structured interrelated information. Proceedings of the First International Workshop on Open Data, New York, US.

International Arabidopsis Informatics Consortium. (2012) Taking the next step: Building an Arabidopsis Information Portal. Plant Cell 24, 2248–2256.

Ballester, P. J. (2012) Machine learning scoring functions based on random forest and support vector regression. Lecture Notes in Computer Science 7632 LNBI, 14-25.

Ballester, P. J. Shape Recognition Methods and Systems for Searching Molecular Databases. U.S. Patent No. 8,244,483. 14 Aug. 2012.

Ballester, P. J., Mangold M., Howard N. I., et al. (2012) Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. J Royal Soc Interface 9, 3196-3127.

Berman, H. M., Kleywegt, G. J., Nakamura, H. and Markley J. L. (2012) The future of the protein data bank. Biopolymers; http://dx.doi.org/10.1002/bip.22132.

Berman, H. M., Kleywegt, G. J., Nakamura, H. and Markley J. L. (2012) The protein data bank at 40: reflecting on the past to prepare for the future. Structure 20, 391-396.

Birney, E. (2012) The making of ENCODE: Lessons for big-data projects. Nature 489, 49–51.

Boraska, V., Jerončić A., Colonna V., et al. (2012) Genome-wide meta-analysis of common variant differences between men and women. Hum Mol Genet 21, 4805-15.

Brandizi, M., Kurbatova N., Sarkans U. and Rocca-Serra P. (2012) Graph2tab, a library to convert experimental workflow graphs into tabular formats. Bioinformatics. 28, 1665-1667.

Brazma, A., Cerans K., Ruklisa D., Schlitt T. and Viksna J. (2012) HSM: a hybrid ystem based approach for modelling intracellular networks. Gene.

Brenchley, R., Spannagl M., Pfeifer M., et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491, 705–710.

Buchel, F., Wrzodek C., Mittag F., et al. (2012) Qualitative translation of relations from BioPAX to SBML qual. Bioinformatics (Oxford, England). 28, 2648-2653.

Burge, S., Kelly E., Lonsdale D., et al. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database 2012.

Caldas, J., Gehlenborg N., Kettunen E., Faisal A., Rönty M., Nicholson A. G., et al. (2012) Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. Bioinformatics 28, 246–253.

Camilloni, C., Simone De A., Vranken, W. F. and Vendruscolo M. (2012) Determination of secondary structure populations in disordered States of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51, 2224-2231.

Campos, D., Matos S., Lewin I., Oliveira J. L. and Rebholz-Schuhmann D. (2012) Harmonization of gene/protein annotations: towards a gold standard MEDLINE. Bioinformatics. 28, 1253–1261.

Chan, J., Kishore R., Sternberg P. and Van Auken K. (2012) The Gene Ontology: enhancements for 2011. Nucleic Acids Res 40, D559–D564.

Chandras, C., Zouberakis M., Salimova E., et al. (2012) CreZOO–the European virtual repository of Cre and other targeted conditional driver strains. Database (Oxford). 2012, bas029.

Chen, C-K. K., Mungall C. J., Gkoutos G. V., et al. (2012) MouseFinder: candidate disease genes from mouse phenotype data. Hum Mutat 33, 858-866.

Cheng, C., Alexander R., Min R., et al. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res 22, 1658–1667.

Chepelev, L. L., Hastings J., Ennis M., Steinbeck C. and Dumontier M. (2012) Self-organizing ontology of biochemically relevant small molecules. BMC Bioinform 13, 3.

Clare, S., John, V., Walker, A. W., et al. (2013). Enhanced Susceptibility to Citrobacter rodentium Infection in MicroRNA-155-Deficient Mice. Infection and immunity 81, 723-732.

Clarke, L., Zheng-Bradley X., Smith R., et al. (2012) The 1000 Genomes Project: data management and community access. Nat Methods 9, 459-462.

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

Conrad, T., Cavalli F MG., Holz H., et al. (2012) The MOF chromobarrel domain controls genome-wide H4K16 acetylation and spreading of the MSL complex. Developmental Cell 22, 610–624.

Conrad, T., Cavalli, F.M.G., Vaquerizas. J. M., et al. (2012) Drosophila Dosage Compensation Involves Enhanced Pol II Recruitment to Male X-Linked Promoters. Science 337, 742–746.

Côté, R. G., Griss J., Dianes J. A., , et al. (2012) The PRoteomics IDEntification (PRIDE) Converter 2 Framework: An Improved Suite of Tools to Facilitate Data Submission to the PRIDE Database and the ProteomeXchange Consortium. Mol Cell Proteomics 11, 1682–1689.

Croset, S., Hoehndorf R. and Rebholz-Schuhmann D. (2012) Integration of the Anatomical Therapeutic Chemical Classification System and DrugBank using OWL and text mining. GI Working Group on Ontologies in Biomedicine and the Life Sciences (OBML).

Crosswell, L. C. and Thornton J. M. (2012) ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol. 30, 241–242.

Csordas, A., Ovelleiro, D., Wang, R., et al. (2012) PRIDE: Quality control in a proteomics data repository. Database 2012.

Dalby, A. R., Emam, I. and Franke, R. (2012) Analysis of Gene Expression Data from Non-Small Cell Lung Carcinoma Cell Lines Reveals Distinct Sub-Classes from Those Identified at the Phenotype Level. PLoS One 7(11), e50253.

Dasmahapatra, K. K., Walters J. R., Briscoe A. D., et al. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487, 94–98.

Davis, M. P. A., Abreu-Goodger C., van Dongen S., et al. (2012) Large-Scale Identification of MicroRNA Targets in Murine Dgcr8-Deficient Embryonic Stem Cell Lines. PLoS One 7, e41762.

de Bono, B. and Hunter P. (2012) Integrating knowledge representation and quantitative modelling in physiology. Biotechnol J 7, 958–972.

de Bono, B., Grenon P. and Sammut S. J. (2012) ApiNATOMY: A novel toolkit for visualizing multiscale anatomy schematics with phenotype-related information. Hum Mutat 33, 837-848.

De Matos, P., Adams N., Hastings J., Moreno P. and Steinbeck C. (2012) A Database for Chemical Proteomics: ChEBI. Methods Mol Biol 803, 273–296.

Dessimoz, C., Gabaldón T., Roos D. S., et al. (2012) Toward community standards in the quest for orthologs. Bioinformatics 28, 900–904.

Dessimoz, C., Gabaldón, T., Roos, D. S., Sonnhammer, E. L. and Herrero, J. (2012) Toward community standards in the quest for orthologs. Bioinformatics 28, 900-904.

Deutsch, E. W., Chambers M., Neumann S., et al. (2012) TraML: a standard format for exchange of selected reaction monitoring transition lists. Mol Cell Proteomics 11, 4.

Dimmer, E. C., Huntley R. P., Alam-Faruque Y., et al. (2012) The UniProt-GO annotation database in 2011. Nucleic Acids Res 40, D565–D570.

Dong, X., Greven M. C., Kundaje A., et al. (2012) Modeling gene expression using chromatin features in various cellular contexts. Genome Biol 13, R53.

Doreleijers, J. F., Vranken W. F., Schulte C., et al. (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. Nucleic Acids Res 40, D519-D524.

Dumousseau, M., Rodriguez N., Juty N. and Novere N. L. (2012) MELTING, a flexible platform to predict the melting temperatures of nucleic acids. BMC Bioinform 13, 101.

Eduati, F., de Las Rivas J., Di Camillo B., Toffolo G. and Saez-Rodriguez J. (2012) Integrating literature-constrained and data-driven inference of signalling networks. Bioinformatics (Oxford, England). 28, 2311–2317.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Engström, P. G., Tommei D., Stricker S. H., et al. (2012) Digital transcriptome profiling of normal and glioblastoma-derived neural stem cells identifies genes associated with patient survival. Genome Med 4, 76.

Faure, A. J., Schmidt D., Watt S., et al. (2012) Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Res 22, 2163-75.

Fernández-Suárez, X. M. (2012) Analyzing genomic data: understanding the genome. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2, 116-137.

Fidalgo, S., Ivanov, D. K. and Wood, S. H. (2012) Serotonin: from top to bottom. Biogerontology 1-25.

Flicek, P., Amode M. R., Barrell D., et al. (2012) Ensembl 2012. Nucleic Acids Res 40, D84–D90.

Fonseca, N. A., Rung J., Brazma A. and Marioni J. C. (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28, 3169–3177.

Fuellen, G., Dengjel J., Hoeflich A., et al. (2012) Systems Biology and Bioinformatics in Aging Research: A Workshop Report. Rejuvenation Research.

Furnham, N., De Beer, T. A. and Thornton J. M. (2012) Current challenges in genome annotation through structural biology and bioinformatics. Current opinion in structural biology 22, 594-601.

Furnham, N., Laskowski R. A. and Thornton J. M. (2012) Abstracting knowledge from the protein data bank. Biopolymers. 99, 183-188.

Furnham, N., Sillitoe I., Holliday G. L., et al. (2012) Exploring the Evolution of Novel Enzyme Functions within Structurally Defined Protein Superfamilies. PLoS Comp Biol 8, e1002403.

Furnham, N., Sillitoe I., Holliday G. L., et al. (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. Nucleic Acids Res 40, D776-D782.

Garnett, M. J., Edelman E. J., Heidorn S. J., et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575.

Gaulton, A., Bellis L. J., Bento, A. P., et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40, D1100-D1107.

Goldman, N. and Yang Z. (2012) Foreword. In: Cannarozzi, G. M. and Schneider A., Eds. Codon Evolution. Oxford University Press, USA, pp. ix-x.

Goncalves, A., Leigh-Brown S., Thybert D., et al. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res 22, 2376-84.

Gonzalez-Galarza, F. F., Lawless C., Hubbard S. J., et al. (2012) A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis. OMICS 16, 431-442.

Gore, S., Velankar S. and Kleywegt G. J. Implementing an X-ray validation pipeline for the Protein Data Bank. Acta Crystallogr (Section D: Biological crystallography) 68, 478-483.

Gostev, M., Faulconbridge A., Brandizi M., et al. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. Nucleic Acids Res 40, D64-D70.

Griss, J., Reisinger F., Hermjakob H. and Vizcaíno J. A. (2012) jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. Proteomics 12, 795–798.

Groenen, M. A. M., Archibald A. L., Uenishi H., et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393-8.

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

Gurha, P., Abreu-Goodger C., Wang T., et al. (2012) Targeted Deletion of MicroRNA-22 Promotes Stress-Induced Cardiac Dilation and Contractile Dysfunction: Clinical Perspective. Circulation 125, 2751–2761.

Hardy, B., Apic G., Carthew P., et al. (2012) Food for thought. A toxicology ontology roadmap. ALTEX. 29, 129–137.

Harrow, I., Filsell W., Woollard P., et al. (2012) Towards virtual know ledge broker services for semantic integration of life science literature and data sources. Drug Discovery Today.

Hastings, J., De Matos P., Dekker A., et al. (2012) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res 41, D456–D463.

Hastings, J., Josephs Z. and Steinbeck C. (2012) Accessing and using chemical property databases. Methods in Molecular Biology (Clifton, NJ) 929, 193.

Hastings, J., Magka D., Batchelor C., et al. (2012) Structure-based classification and ontology in chemistry. J Cheminform 4, 8.

Haudek-Prinz, V., Klepeisz P., Slany A., et al. (2012) Proteome signatures of inflammatory activated primary human peripheral blood mononuclear cells. J Proteomics.

Haug, K., Salek R. M., Conesa P., et al. (2012) MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 41, D781–D786.

Henderson, R., Sali A., Baker M. L., et al. (2012) Outcome of the first electron microscopy validation task force meeting. Structure 20, 205-214.

Hendrickx, P. M., Gutmanas A. and Kleywegt G. J. (2012) Vivaldi: visualisation and validation of biomacromolecular NMR structures from the PDB. Proteins; http://dx.doi.org/10.1002/prot.24213.

Hersey, A., Senger S. and Overington J. P. (2012) Open data for drug discovery: learning from the biological community. Future Med Chem 4, 1865-1867.

Hitz, M-P. P., Lemieux-Perreault L-P. P., Marshall C., et al. (2012) Rare copy number variants contribute to congenital left-sided heart disease. PLoS Genet 8, e1002903.

Hoffman, M. M., Ernst J., Wilder S. P., et al. (2012) Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 41, 827-841.

Hoffman, M. M., Ernst J., Wilder S. P., Kundaje A., Harris R. S., Libbrecht M., et al. (2012) Integrative annotation of chromatin elements from ENCODE data. Genome Res.

Holliday, G. L., Andreini C., Fischer J. D., et al. (2012) MACiE: exploring the diversity of biochemical reactions. Nucleic Acids Res 40, D783-D789.

Howe, K., Davis P., Paulini M., et al. (2012) WormBase: Annotating many nematode genomes. Worm 1, 15–21.

Hunter, C. I., Mitchell A., Jones P., et al. (2012) Metagenomic analysis: the challenge of the data bonanza. Brief Bioinform 13, 743-746.

Hunter, S., Jones P., Mitchell A., et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40, D306–D312.

Iorio, F., Rittman T., Ge H., Menden M. and Saez-Rodriguez J. (2012) Transcriptional data: a new gateway to drug repositioning? Drug Discovery Today.

Iqbal, Z., Caccamo M., Turner I., et al. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44, 226-32.

Jayaseelan, K V., Moreno P., Truszkowski A., et al. (2012) Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinform 13, 106.

Jones, A. M. E., Aebersold R., Daube M., et al. (2012) The HUPO initiative on Model Organism Proteomes, iMOP. Proteomics 12, 340-345.

Jones, A. R., Eisenacher M., Mayer G., et al. (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics 11, M111.014381.

Jones, F. C., Grabherr M. G., Chan Y. F., et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61.

Jordan, G. and Goldman N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29, 1125-1139.

Junion, G., Spivakov M., Girardot C., et al. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148, 473–486.

Jupe, S., Akkerman J. W., Soranzo N. and Ouwehand W. H. (2012) Reactome: a curated knowledgebase of biological pathways: megakaryocytes and platelets. J Thromb Haemost; doi: 10.1111/j.1538-7836.2012.04930.x

Juty, N., Le Novere, N. and Laibe, C. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res 40(1):D580-6, D580-D583.

Kahramanoglou, C., Prieto A. I., Khedkar S., et al. (2012) Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. Nature Comm 3, 886.

Kapushesky, M., Adamusiak T., Burdett T., et al. (2012) Gene Expression Atlas update – a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res 40, D1077-D1081.

Kerrien, S., Aranda B., Breuza L., et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40, D841–D846.

Kersey, P. J., Staines D. M., Lawson D., et al. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res 40, D91–D97.

Kinsinger, C. R., Apffel J., Gorman J., et al. (2012) Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles. Proteomics 12, 11-20.

Kirouac, D. C., Saez-Rodriguez J., Swantek J., et al. (2012) Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Systems Biology 6, 29.

Klech, H., Brooksbank C., Price S., et al. (2012) European initiative towards quality standards in education and training for discovery, development and use of medicines. Eur J Pharma Sci 45, 520-520.

Koh, G. C. K. W., Porras P., Aranda B., Hermjakob H. and Orchard S. E. (2012) Analyzing Protein–Protein Interaction Networks. J Proteome Res 11, 2014–2031.

König, J., Zarnack K., Luscombe N. M. and Ule J. (2012) Protein–RNA interactions: new genomic technologies and perspectives. Nat Rev Genet 13, 77–83.

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

Kruger, F. A., Rostom R. and Overington J. P. (2012) Mapping small molecule binding data to structural domains. BMC Bioinform 13, S11.

Lahti, L., Torrente A., Elo L. L., Brazma A. and Rung J. (2012) Fully scalable online-preprocessing algorithm for short oligonucleotide microarray atlases. arXiv; preprint arXiv:1212.5932.

Lam, K C., Mühlpfordt F., Vaquerizas J. M., et al. (2012) The NSL Complex Regulates Housekeeping Genes in Drosophila. PLoS Genetics 8, e1002736.

Lam, M. P. Y., Vivanco F., Scholten A., et al. (2012) HUPO 2011: The New Cardiovascular Initiative-Integrating Proteomics and Cardiovascular Biology in Health and Disease. Proteomics 12, 749–751.

Lee, B. K., Bhinge A. A., Battenhouse A., et al. (2012) Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. Genome Res 22, 9–24.

Lewin, I., Kafkas S. and Rebholz-Schuhmann D. (2012) Centroids: Gold Standards with Distributional Variations. Proceedings of the International Conference on Language Resources and Evaluation.

Lewis, S., Csordas A., Killcoyne S., et al. (2012) Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. BMC Bioinform. 13, 324.

Lewis, S., CSORDAS A., Killcoyne S., Hermjakob H., Hoopmann M. R., Moritz R. L., et al. (2012) Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. BMC bioinformatics. 13, 324.

Li, L., Stefan M. I. and Le Novère N. (2012) Calcium Input Frequency, Duration and Amplitude Differentially Modulate the Relative Activation of Calcineurin and CaMKII. PloS one. 7, e43810.

Li, W., McWilliam H., GOUJON M., Cowley A., Lopez R. and Pearson W. R. (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. Bioinformatics. 28, 1650–1651.

Li, W., McWilliam H., Goujon M., et al. (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. Bioinformatics. 28, 1650–1651.

Liakata, M., Kim, J. H., Saha, S., et al. (2012) Three Hybrid Classifiers for the Detection of Emotions in Suicide Notes. Biomedical Informatics Insights 5, 175.

Liakata, M., Saha S., Dobnik S., Batchelor C. and Rebholz-Schuhmann D. (2012) Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28, 991–1000.

Lopes, M. C., Joyce C., Ritchie G. R. S., et al. (2012) A Combined Functional Annotation Score for Non-Synonymous Variants. Hum Hered 73, 47-51.

Löytynoja, A. (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. (Anisimova, M., Ed.) Evolutionary Genomics: Statistical and Computational Methods. 1, 203-235.

Löytynoja, A. (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. (Anisimova, M., Ed.). Evolutionary Genomics: Statistical and Computational Methods 1, 203-235.

Löytynoja, A., Vilella A. J. and Goldman N. (2012) Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 28, 1684-1691.

Lu, D., Davis M. P. A., Abreu-Goodger C., et al. (2012) MiR-25 Regulates Wwp2 and Fbxw7 and Promotes Reprogramming of Mouse Fibroblast Cells to iPSCs. PLoS One 7, e40938.

Ludtke, S. J., Lawson C. L., Kleywegt, G. J., et al. (2012) The 2010 cryo-em modeling challenge. Biopolymers 97, 651-654.

Luo, H., Löytynoja A. and Moran M. A. (2012) Genome content of uncultivated marine Roseobacters in the surface ocean. Environ Microbiol 14, 41-51.

MacArthur, D. G., Balasubramanian S., Frankish A., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823-8.

MacNamara, A., Terfve C., Henriques D., Bernabé B P. and Saez-Rodriguez J. (2012) State-time spectrum of signal transduction logic models. Phys Biol 9, 045003.

Mallon, A-M. M., Iyer V., Melvin D., et al. (2012) Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. Mamm Genome 23, 641-52.

Manakov, S. A., Morton A., Enright A. J. and Grant S. G. N. (2012) A neuronal transcriptome response involving stress pathways is buffered by neuronal microRNAs. Front Neurosci 6, 156.

Martincorena, I., Seshasayee A. S. N. and Luscombe N. M. (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 485, 95–98.

Massingham, T. and Goldman N. (2012) All your base: a fast and accurate probabilistic approach to base calling. Genome Biol 13, R13.

Massingham, T. and Goldman N. (2012) Error-correcting properties of the SOLiD Exact Call Chemistry. BMC Bioinform 13, 145.

Mattioni, M., Cohen U. and Novère N. L. (2012) Neuronvisio: A Graphical User Interface with 3D Capabilities for NEURON. Frontiers in neuroinformatics. 6, 20.

Melas, I. N., Mitsos A., Messinis D. E., et al. (2012) Construction of large signaling pathways using an adaptive perturbation approach with phosphoproteomic data. Molecular BioSystems 8, 1571–1584.

Menden, M. P., Iorio, F., Garnett, M., et al. (2012) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. arXiv preprint arXiv:1212.0504.

Metzakopian, E., Lin W., Salmon-Divon M., Dvinge H., et al. (2012) Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors. Development 139, 2625–2634.

Metzakopian, E., Lin W., Salmon-Divon M., et al. (2012) Genome-wide characterization of Foxa2 targets reveals upregulation of floor plate genes and repression of ventrolateral genes in midbrain dopaminergic progenitors. Development. 139, 2625–2634.

Milacic, M., Haw R., Rothfels K., et al. (2012) Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. Cancers. 4, 1180–1211.

Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., et al. (2012) Cryo-electron microscopy – a primer for the non-microscopist. FEBS J 280, 28–45.

Mitsos, A., Melas I. N., Morris M. K., et al. (2012) Non Linear Programming (NLP) Formulation for Quantitative Modeling of Protein Signal Transduction Pathways. PLoS One 7, e50085.

Murchison, E. P., Schulz-Trieglaff O. B., Ning Z., Alexandrov L. B., Bauer M. J., Fu B., et al. (2012) Genome sequencing and analysis of the tasmanian devil and its transmissible cancer. Cell 148, 780-791.

Nelson, A. C., Pillay N., Henderson S., et al. (2012) An integrated functional genomics approach identifies the regulatory network directed by brachyury (T) in chordoma. J Pathol 228, 274-85.

Newman, J., Bolton E. E., Muller-Dieckmann J., et al. (2012) On the need for an international effort to capture, share and use crystallization screening data. Acta Crystallogr (Section F: Structural biology and crystallization communications) 68, 253-258.

Nyrönen, T. H., Laitinen J., Tourunen O., et al. (2012) Delivering ICT infrastructure for biomedical research. 2012 Joint 10th Working IEEE/IFIP Conference on Software Architecture, WICSA 2012 and the 6th European Conference on Software Architecture, ECSA 2012. Helsinki, 20–24 August. Code92923, 37-44.

Oellrich, A., Grabmüller C. and Rebholz-Schuhmann D. (2012) Automatically transforming pre-to post-composed phenotypes: EQ-lising HPO and MP. GI Working group on ontologies in biomedicine and life sciences(OBML).

Oellrich, A., Hoehndorf R., Gkoutos G. V. and Rebholz-Schuhmann D. (2012) Improving Disease Gene Prioritization by Comparing the Semantic Similarity of Phenotypes in Mice with Those of Human Diseases. PLoS One 7, e38937.

Orchard, S. (2012) Molecular interaction databases. Proteomics 12, 1656–1662.

Orchard, S., Al-Lazikani B., Bryant S., et al. (2012) Shouldn't enantiomeric purity be included in the'minimum information about a bioactive entity? Response from the MIABE group. Nat Rev Drug Disc 11, 730–730.

Orchard, S., Albar J. P., Deutsch E. W., et al. (2012) From Proteomics Data Representation to Public Data Flow: A Report on the HUPO-PSI Workshop September 2011, Geneva, Switzerland. Proteomics 12, 351–355.

Orchard, S., Binz P. A., Borchers C., et al. (2012) Ten Years of Standardizing Proteomic Data: A Report on the HUPO-PSI Spring Workshop. Proteomics 12, 2767–2772.

Orchard, S., Kerrien S., Abbani S., et al. (2012) Corrigendum: Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9, 626–626.

Orchard, S., Kerrien S., Abbani S., et al. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9, 345–350.

Overington, J. P. (2013). Chemogenomics, Introduction to Cheminformatics: Data management, manipulation and properties (in press).

Papatheodorou, I., Ziehm M., Wieser D., Alic N., et al. (2012) Using answer set programming to integrate RNA expression with signalling pathway information to infer how mutations affect ageing. PLoS One 7, e50881.

Parts, L., Hedman, A.A.K., Keildson S., et al. (2012) Extent, Causes, and Consequences of Small RNA Expression Variation in Human Adipose Tissue. PLoS Genetics. 8, e1002704.

Patwardhan, A., Carazo J. M., Carragher B., et al. (2012) Data management challenges in three-dimensional EM. Nat Struct Mol Biol 19, 1203-1207.

Pavelin, K., Cham J. A., De Matos P., et al. (2012) Bioinformatics meets user-centred design: a perspective. PLoS Comp Biol 8, e1002554.

Perez-Riverol, Y., Audain E., Millan A., et al. (2012) Isoelectric point optimization using peptide descriptors and support vector machines. J Proteomics 75, 2269-2274.

Pestian, J. P. (2012) Introductory editorial. Biomedical Informatics Insights. 5, 1.

Radom, M., Rybarczyk A., Kottman R., et al. (2012) Poseidon: An information retrieval and extraction system for metagenomic marine science. Ecological Informatics.

Rebholz-Schuhmann, D., Oellrich A. and Hoehndorf R. (2012) Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet/

Reimand, J., Aun A., Vilo J., Vaquerizas J., Sedman J. and Luscombe N. (2012) m: Explorer: multinomial regression models reveal positive and negative regulators of longevity in yeast quiescence. Genome Biol 13, R55.

Reisinger, F., Krishna R., Ghali F., et al. (2012) jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. Proteomics 12, 790–794.

Reynolds, N., Latos P., Hynes-Allen A., et al. (2012) NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. Cell Stem Cell 10, 583–594.

Rung, J. and Brazma A. (2012) Reuse of public genome-wide gene expression data. Nat Rev Genet 14, 89–99.

Rustici, G., Kolesnikov, N., Brandizi, M., et al. (2013). ArrayExpress update—trends in database growth and links to data analysis tools. Nucleic Acids Res, 41(D1), D987-D990.

Sacco, F., Gherardini P F., Paoluzi S., et al. (2012) Mapping the human phosphatome on growth pathways. Mol Sys Biol 8, 1.

Sahakyan, Aleksandr B., Cavalli Andrea, et al. (2012) Protein Structure Validation Using Side-Chain Chemical Shifts. J Phys Chem B 116, 4754–4759.

Salazar, G. A., Garcia, L. J., Jones P., et al. (2012) MyDas, an Extensible Java DAS Server. PLoS One. 7, e44180.

San Mauro, D., Gower D. J., Cotton J. A., et al. (2012) Experimental design in phylogenetics: testing predictions from expected information. Systematic Biology. 61, 661-674.

Sansone, S-A., Rocca-Serra P., Field D., et al. (2012) Toward interoperable bioscience data. Nat Genet 44, 121–126.

Scally, A., Dutheil J. Y., Hillier LD. W., et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483, 169-75.

Schmidt, D., Schwalie P. C., Wilson M. D., et al. (2012) Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell 148, 335-48.

Schneider, M. V., Walter P., Blatter M. C., et al. (2012) Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. Brief Bioinform 13, 383–389.

Schroder, K., Irvine K. M., Taylor M. S., et al. (2012) Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. Proc Nat Acad Sci U.S.A. 109, E944-E953.

Schulz, M. H., Zerbino D. R., Vingron M. and Birney E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.

Selcuklu, S. D., Donoghue M. T. A., Rehmet K., et al. (2012) MicroRNA-9 inhibition of cell proliferation and identification of novel miR-9 targets by transcriptome profiling in breast cancer cells. J Biol Chem 287, 29516–29528.

Serra-Musach, J., Aguilar H., Iorio F., et al. (2012) Cancer develops, progresses and responds to therapies through restricted perturbation of the protein-protein interaction network. Integrative Biol 4, 1038–1048.

Sillitoe, I., Cuff A. L., Dessailly B. H., et al. (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 41, D490–D498.

Silva, Sousa Da A. W. and Vranken W. F. ACPYPE - AnteChamber PYthon Parser interface. BMC Res Notes 5, http://dx.doi.org/10.1186/1756-0500-5-367.

Sipos, B., Massingham T., Stütz A. M. and Goldman N. (2012) An improved protocol for sequencing of repetitive genomic regions and structural variations using mutagenesis and Next Generation Sequencing. PLoS One 7, e43359.

Škunca, N., Altenhoff A. and Dessimoz C. (2012) Quality of computationally inferred Gene Ontology annotations. PLoS Comp Biol 8, e1002533.

Spivakov, M., Akhtar, J., Kheradpour, P., et al. (2012) Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol 13, R49.

Stefan, M. I., Marshall D. P. and Le Novère N. (2012) Structural Analysis and Stochastic Modelling Suggest a Mechanism for Calmodulin Trapping by CaMKII. PLoS One 7, e29406.

Steinbeck, C., Conesa P., Haug K., et al. (2012) MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics 8, 757–760.

Taboureau, O., Hersey, A., Audouze, K., et al. (2012) Toxicogenomics Investigation Under the eTOX Project. J Pharmacogenom Pharmacoproteomics S7, 2153–0645.

Terfve, C. and Saez-Rodriguez, J. (2012) Modeling Signaling Networks Using High-throughput Phospho-proteomics. Advances in experimental medicine and biology. Adv Sys Biol 736, 19–57.

Terfve, C., Cokelaer, T., MacNamara, A., et al. (2012) CellNOptR: a flexible toolkit to train protein signaling net- works to data using multiple logic formalisms. BMC Systems Biol 6, in 13.

Uhlenhaut, H. N., Barish, G. D., Yu, R. T., et al. (2012) Insights into Negative Regulation by the Glucocorticoid Receptor from Genome-wide Profiling of Inflammatory Cistromes. Molecular Cell.

UniProt Consortium and others (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40, D71–D75.

Van Dongen, S. and Enright A. J. (2012) Metric distances derived from cosine similarity and Pearson and Spearman correlations. arXiv, preprint arXiv:1208.3145.

Van Iersel, M. P., Villeger A. C., Czauderna T., et al. (2012) Software support for SBGN maps: SBGN-ML and LibSBGN. Bioinformatics (Oxford, England) 28, 2016-2021.

Velankar, S., Alhroub Y., Best C., et al. (2012) PDBe: Protein Data Bank in Europe. Nucleic Acids Res 40, D445-D452.

Videla, S., Guziolowski C., Eduati F., et al. (2012) Revisiting the Training of Logic Models ofProtein Signaling Networks with a Formal Approach based on Answer Set Programming. Lecture Notes in Computer Science. 1–22.

Vizcaíno, J.A., Côté, R.G., Csordas, A., et al. (2013). The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 41(D1), D1063–D1069.

Wang, J., Zhuang J., Iyer S., et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22, 1798–1812.

Wang, R., Fabregat, A., Ríos, D., et al. (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. Nat Biotechnol 30, 135-137.

Ware, J. S., Walsh R., Cunningham F., Birney E. and Cook S. A. (2012) Paralogous annotation of disease-causing variants in long QT syndrome genes. Hum Mutat 33, 1188-1191.

Wassenaar, T. A., van Dijk M., Loureiro-Ferreira N., et al. (2012) WeNMR: Structural Biology on the Grid. J Grid Computing; http://dx.doi.org/10.1007/s10723-012-9246-z.

Wegner, J. K., Sterling A., Guha R., et al. (2012) Cheminformatics. Communications of the ACM 55, 65–75.

Wein, S. P., Côté R. G., Dumousseau M., et al. (2012) Improvements in the protein identifier cross-reference service. Nucleic Acids Res 40, W276–W280.

Wood, V., Harris M. A., McDowall M. D., et al. (2012) PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res 40, D695–D699.

Xue, V., Burdett T., Lukk M., Taylor J., Brazma A. and Parkinson H. MageComet–web application for harmonizing existing large-scale experiment descriptions. Bioinformatics (Oxford, England) 28, 1402-1403.

Xue, Y., Chen Y., Ayub Q., et al. (2012) Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. Am J Hum Genet 91, 1022-1032.

Ying Y., Kafkas S., Conroy M. and Rebholz-Schuhmann D. Towards Generating a Corpus Annotated for Prokaryote-Drug Relations. Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012).

Yip, K. Y., Cheng C., Bhardwaj N., et al. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 13, R48.

Yook, K., Harris T. W., Bieri T., et al. (2012) WormBase 2012: more genomes, more data, new website. Nucleic Acids Res 40, D735–D741.

Zdrazil, B., Pinto M., Vasanthanathan P., et al. (2012) Annotating human P-glycoprotein bioassay data. Mol Inform 31, 599-609.

Zhai, W., Nielsen R., Goldman N. and Yang Z. (2012) Looking for Darwin in genomic sequences –- validity and success of statistical methods. Mol Biol Evol 29, 2889-2893.

Zhang, X., Kendrick K. M., Zhou H., Zhan Y. and Feng J. (2012) A Computational Study on Altered Theta-Gamma Coupling during Learning and Phase Coding. PLoS One 7, e36472.

Zheng-Bradley, X. and Flicek P. (2012) Maps for the world of genomic medicine: The 2011 CSHL Personal Genomes Meeting. Hum Mutat 33, 1016-1019.

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

comms@ebi.ac.uk
www.ebi.ac.uk

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

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