D18–D25 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 23 November 2013*
doi:10.1093/nar/gkt1206

The European Bioinformatics Institute's data resources 2014

Catherine Brooksbank*, Mary Todd Bergman, Rolf Apweiler, Ewan Birney and Janet Thornton

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 17, 2013; Revised November 1, 2013; Accepted November 4, 2013

ABSTRACT

Molecular Biology has been at the heart of the 'big data' revolution from its very beginning, and the need for access to biological data is a common thread running from the 1965 publication of Dayhoff's 'Atlas of Protein Sequence and Structure' through the Human Genome Project in the late 1990s and early 2000s to today's population-scale sequencing initiatives. The European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk) is one of three organizations worldwide that provides free access to comprehensive, integrated molecular data sets. Here, we summarize the principles underpinning the development of these public resources and provide an overview of EMBL-EBI's database collection to complement the reviews of individual databases provided elsewhere in this issue.

INTRODUCTION

The molecular life sciences are becoming increasingly data-driven and reliant on open-access databases (1). This is as true of the applied sciences as it is of fundamental research: in the past year, we have witnessed announcements that the UK's National Health Service will invest in sequencing the genomes of up to 100 000 citizens (see http://www.gov.uk/government/speeches/strategy-for-uk-life-sciences-one-year-on and http://news.sciencemag.org/biology/2012/12/u.k.-unveils-plan-sequence-whole-genomes-100000-patients); the Faroe Islands are planning to sequence the genome of every citizen who wishes to have this information (see http://www.fargen.fo/en/), and large-scale metagenomics projects are helping us to map the global biodiversity of the oceans (2).

The European Bioinformatics Institute (EMBL-EBI), part of the European Molecular Biology Laboratory, makes these large-scale efforts possible. It helps scientists deposit their research data into public collections, produces value-added knowledge bases and makes its entire holdings accessible to all, thereby enabling millions of scientists worldwide to explore, analyse, interpret and derive new knowledge from decades of scientific endeavour.

Among its other roles (Appendix 1), EMBL-EBI has a mission to provide free and open access to biomolecular information, spanning scientific literature and the data supporting it: DNA and protein sequences; biomolecules and their structures, functions, reactions and interactions; and practical tools for analysis and discovery.

These offerings include personally identifiable genetic and phenotypic data resulting from biomedical research projects—an area of growing importance as healthcare systems embrace genomic medicine. Managing access to these data sets is a high-priority activity at EMBL-EBI.

EMBL-EBI's core resources are foundational members of international consortia, which share data globally and foster competitiveness among their members. Some of these collaborations have a long history [e.g. the International Nucleotide Sequence Database Collaboration (INSDC) (3), the worldwide Protein Data Bank (wwPDB) (4), UniProt (5) and Ensembl (6)]. Others, driven by EMBL-EBI, are more recent [e.g. IMEx (7) for protein interaction data; ProteomExchange (8) for protein identification data and COSMOS (9) for metabolomics data]. The Global Alliance (10)—a large-scale international effort to enable the secure sharing of genomic and clinical data—is the most recent of these. Each of these collaborations exemplifies the fundamental principles of EMBL-EBI service provision (Appendix 2).

DESIGNED TO BE USED

EMBL-EBI embraces user-centred design (UCD), an approach that focuses on the behaviour and needs of the people who will actually use the product. UCD has been successfully applied to design in many different domains, although its application to bioinformatics services (11) is

*To whom correspondence should be addressed. Tel: +44 1223 492525; Fax: +44 1224 494468; Email: cath@ebi.ac.uk

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

relatively recent. The case for using UCD for bioinformatics services is compelling: even major bioinformatics resources are known to suffer from usability problems (12), which prevent users from completing tasks (13).

By placing the user at the forefront of our minds as we design, test and implement our services, we create more useful and user-friendly resources. This approach has been used to completely redesign the EMBL-EBI website—a major project that has involved every team at EMBL-EBI. The redesign puts users at the centre of the process, providing an intuitive new interface to EMBL-EBI services. It aims for consistent functionality without stifling the individual data resource brands.

EMBL-EBI's search engine (14) displays results in an organized manner, according to the central dogma of molecular biology (i.e. DNA makes RNA makes protein). This results in an uncluttered results 'dashboard' from which users can explore genes, protein sequences, gene expression, molecular structures and related scientific literature. The search allows easy comparison of key information for human, mouse, fly and other species.

Bioinformatics services on the EMBL-EBI website are displayed according to nine major themes (Figure 1; see also http://www.ebi.ac.uk/services), which were informed by user feedback. We have organized this review in the same way.

LITERATURE

Access to the scientific literature is a basic requirement for research. EMBL-EBI coordinates the development of Europe PubMed Central Europe (PMC) (15) in

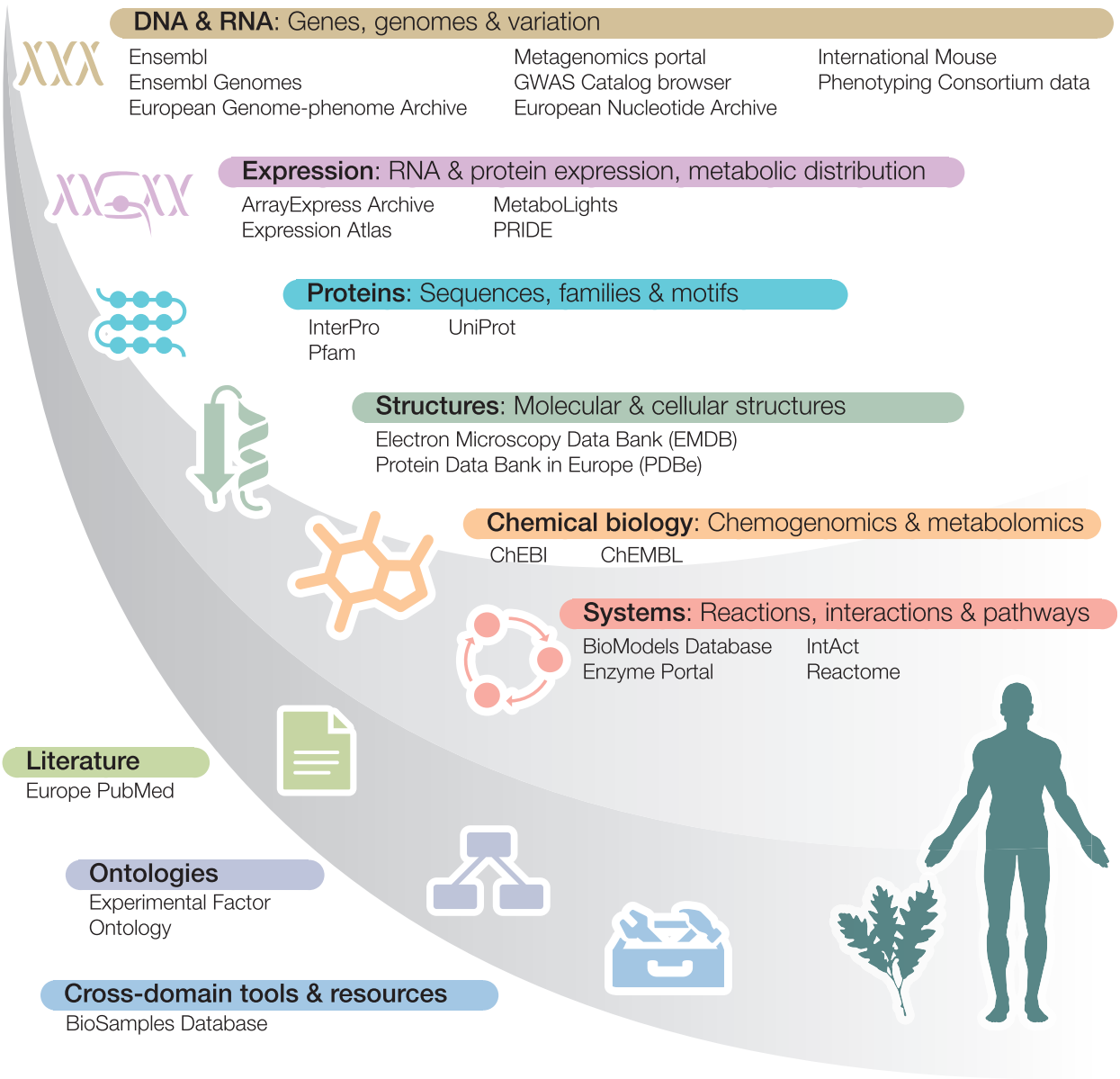

Figure 1. EMBL-EBI's core data resources. The figure summarizes the resources described in this review; a fuller summary of EMBL-EBI's data resources and tools, which links to each resource, is available at http://www.ebi.ac.uk/services.

collaboration with The University of Manchester (Mimas and NaCTeM) and the British Library. PMC is part of PMC International, which is coordinated by the US National Center for Biotechnology Information and includes PMC Canada. Launched in November 2012, Europe PMC is funded through a collective of European funders, coordinated by the Wellcome Trust.

Europe PMC combines the entire collection of PubMed abstracts, PMC full-text articles, patent abstracts (European, US and international), National Health Service (NHS) clinical guidelines, Agricola records and other record types, and builds innovative tools to help researchers explore every aspect of the literature. Because it is developed at EMBL-EBI, it is uniquely positioned to link abstracts and articles seamlessly to the underlying data. For example, Europe PMC is integrated with UniProt, the Protein Data Bank in Europe (PDBe) and the European Nucleotide Archive (ENA).

CROSS-DOMAIN TOOLS AND RESOURCES

The integration of different -omics data types requires the consistent application of metadata to data sets derived from the same sample. Sample metadata are managed within EMBL-EBI's BioSamples database (16), which provides links to assays for specific samples (including reference samples such as cell lines) and accepts direct submissions. In 2013, the BioSamples database launched a new user interface, application programming interface (API) and submission-accessioning service.

ONTOLOGIES

Biologists and bioinformaticians look to ontologies and other types of controlled vocabularies as a means of standardizing the way data are described, queried and analysed. Subtle differences in the use of terms and phrases can hamper communication among scientists, and can make automated data exchange prohibitively difficult. Ontologies address these issues, making information in databases more readily human- and computer-readable.

The Gene Ontology (GO) (17) 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 EMBL-EBI, the GO editors play a key role in managing the distributed task of developing and improving GO, whereas the UniProt GO annotation (GOA) program adds high-quality GOAs to proteins in the UniProt Knowledgebase (UniProtKB) (18). Recent enhancements include expanding the functionality of GO's direct ontology submission tool, TermGenie, which now includes a 'free form' input for experienced users; integration of the GO and ChEBI ontologies (19); and improvements to the electronic GOA pipeline.

Another cross-cutting tool is the Experimental Factor Ontology (EFO) (20), which began its life as a practical means of categorizing gene expression data sets. In the past year, it has broadened its application considerably to support annotation of genome-wide association studies (GWAS) and the integration of genomic and disease data.

DNA AND RNA

Nucleotide sequence data are a central reference point onto which many other types of information can be built.

The public record of nucleotide sequence data

The ENA (21) manages the staggering volumes of data generated by next-generation sequencing. The ENA team has developed CRAM, an openly accessible software toolkit and file format for compressing sequence data, leveraging the specific data properties of DNA sequence (22). Officially launched in November 2012, CRAM is a community-led endeavour that is being incorporated into existing tools and pipelines so that researchers can save on local storage space. It also has the advantage of keeping the public archives to a manageable size.

Reference genomes

Ensembl (6), produced jointly by EMBL-EBI and the Wellcome Trust Sanger Institute, enables and advances genome science by providing high-quality integrated annotation on vertebrate genomes within a consistent and accessible infrastructure. Ensembl's new features include the Variant Effect Predictor (23), which predicts the effects of variant positions and alleles on overlapping transcripts and regulatory regions. Ensembl features the genomes of 75 vertebrate species [the mountain gorilla (24) being a notable recent addition]. It also houses the substantial data sets produced by the ENCODE project (25).

Ensembl Genomes (26), launched in 2009 to expand EMBL-EBI's taxonomic coverage of reference genomes, added a significant number of new species to its database in the past year. It now includes the genomes of biting midges, butterflies (27), barley (28), wheat (29) and >6000 bacterial species.

Ensembl Genomes also provides the underlying architecture for several new community portals. Understanding the basis of crop diseases was the driver behind the launch of PhytoPath (http://www.phytopathdb.org), a new portal for plant pathogen data. EMBL-EBI's involvement in the transPLANT project has spawned a new integrative portal for plant genomics data (http://www.transplantdb.eu). Ensembl Genomes has also made metabolic data for >4000 bacterial genomes available through the Microme portal (http://www.microme.eu).

Linking genotype to phenotype

EMBL-EBI's philosophy is to make its data openly available to the research community, but where personally identifiable data are involved, it is important that we honour the consent agreements under which patients provide data, which nearly always exclude the use of genetic data to identify individuals. The European Genome-phenome Archive (EGA) is EMBL-EBI's service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA

contains exclusive data collected from individuals whose consent authorizes data release only for specific research use or to *bona fide* researchers. The EGA provides the necessary security required to control access, maintain patient confidentiality and provide access for those researchers and clinicians who are authorized to view the data. In all cases, data access decisions are made by the appropriate data access-granting organization (DAO) and not by the EGA. An independent Ethics Committee audits the EGA protocols and infrastructure.

Resequencing projects are providing vast amounts of data that link genotype to phenotype, with the ultimate goal of establishing the connections between genetic variation and disease. Two resources reviewed in this NAR database issue provide access to these data, one focusing on GWAS, the other on knockout mice.

EMBL-EBI and the US National Human Genome Research Institute (NHGRI) jointly develop the Catalog of Published GWAS (30). The catalogue is a publicly available manually curated collection of published GWAS with a distinctive and dynamic visualization tool that enables users to click on single-nucleotide polymorphism (SNP)–trait associations mapped to chromosomal locations. Each association is annotated with terms from the EFO (see 'Cross-domain tools and resources' above) to help the user identify SNPs associated with a specific phenotype.

The International Mouse Phenotyping Consortium (IMPC) is building the first comprehensive functional catalogue of a mammalian genome (31). To do this, it is creating a knockout mouse strain for every known protein-coding gene—20 000 mouse strains in total—using a rigorously standardized set of phenotyping protocols. These strains will be made available in public repositories, and data pertaining to each will be made publicly available in near real time, along with open tools for their analysis. Project data will be delivered through a service (http://www.mousephenotype.org) managed by the MPI2 consortium (EMBL-EBI, the Wellcome Trust Sanger Institute, MRC Harwell). Users will be able to search by term, gene, tissue or disease, so they may identify associations between phenotype, gene and protocol swiftly. The results are displayed using the same principles and underlying architecture as EMBL-EBI's global search. The service is expected to launch in early 2014.

Metagenomics

While the projects described above are accumulating an ever greater depth of knowledge about the genomes of long-studied organisms, another approach—metagenomics (32)—increases 'breadth' of knowledge by presupposing nothing about the identity of the organisms present in a sample. EMBL-EBI's ENA and InterPro teams have created an integrated resource—the Metagenomics Portal (http://www.ebi.ac.uk/metagenomics/)—that allows researchers to submit, archive and analyse genomic information from environments containing many species. New functionality is being added regularly in response to user demand.

EXPRESSION

The combination of transcriptomics, proteomics and metabolomics data can provide a powerful basis for deriving a system-based understanding of biological systems. To facilitate such integration, EMBL-EBI is working towards the integration of the ArrayExpress Archive (33), Expression Atlas (33), Proteomics Identifications Database (PRIDE) (34) and MetaboLights (35), EMBL-EBI's newly launched metabolomics database.

EMBL-EBI is developing a Baseline Expression Atlas, which uses high-throughput sequencing-based expression data to report 'absolute' gene expression levels, rather than relative levels. Concurrently, significant improvements have been made to the ArrayExpress archive interface, and the resource has accepted its millionth assay.

PRIDE (34) is a public resource for mass spectrometry-based protein expression data. In 2013, PRIDE achieved the successful and stable implementation of the ProteomeXchange data workflow (http://www.proteomexchange.org). As a result, data depositions more than tripled in number of submissions and in volume.

The final database making up this 'expression and distribution trinity' is the newly launched MetaboLights (35)—the first general purpose open-access database for metabolomics and its derived information. MetaboLights includes a reference layer with information about individual metabolites, their chemistry, spectroscopy and biological roles, connected with a study archive into which researchers deposit primary and metadata on metabolomics studies.

PROTEINS

Protein sequence provides another 'information hub' for the molecular biologist, onto which experimentally validated information about the behaviour and localization of proteins can be hung, and from which hypotheses about structure and function may be generated.

UniProt (5), the unified resource of protein sequence and functional information, is maintained by EMBL-EBI in collaboration with the Swiss Institute of Bioinformatics and Universities of Georgetown and Delaware. UniProt is closely integrated with Ensembl (6) and Ensembl Genomes (26), and has generated new reference proteome sets to match their genes in the reference genomes. UniProt prioritizes the manual annotation of experimental data for human and other reference proteomes in collaboration with other worldwide resources, ensuring the highest quality knowledge is available to researchers. UniProt's automatic annotation exploits the results of manual annotation, resulting in a widening of taxonomic and annotation depth.

EMBL-EBI's protein resources have embraced UCD (see 'Designed to be used'). The UniProt development team has designed new interfaces that enhance user interaction with the website and facilitate access to data. The UniProt content team has extended the databases to accommodate a rapidly growing volume of data and to incorporate variation and proteomics data.

InterPro (36), EMBL-EBI's database of protein families, domains and motifs, has completely re-implemented its back-end to optimize user query functionality. A new user interface has been developed and tested with users, and the results have informed the global EMBL-EBI website redesign process.

Pfam (37), a database of hidden Markov models and alignments describing conserved protein families and domains, is one of InterPro's 11 member databases, and is in the process of migrating from the Wellcome Trust Sanger Institute to EMBL-EBI. Pfam's latest release adds real-time searches of DNA sequences for matches to Pfam models, representative proteome sequence sets to provide non-redundant views of alignments and annotations to disease.

STRUCTURES

Understanding molecular structure is crucial to understanding function. PDBe (38,39), the European arm of the worldwide Protein Data Bank collaboration (wwPDB), provides sophisticated tools for analysing structures, several of which were improved significantly in the past year. These include tools that enhance the analysis of nuclear magnetic resonance entries and many improvements to EMDB (38), the European resource for electron microscopy-based models. EMDB now has a new search service and an interactive viewer for electron tomograms. PDBe is increasingly integrated with other types of information, including sequence data [through the SIFTS service (40)] and the GO (17) (through a new module in the PDBeXplore tool).

SYSTEMS

The genes and gene products encoded by genomes do not act in isolation but do so in coordinated systems, often containing protein, small molecule and oligonucleotide or oligosaccharide components. EMBL-EBI's molecular systems resources enable researchers to build a holistic view of life at the molecular level, building up from enzymes and their mechanisms, through protein—protein interactions and networks, to pathways and quantitative models.

The Enzyme Portal (41,42), launched in February 2012, combines high-quality data from 10 previously isolated databases and organizes information about each enzyme in such a way that the user can flip from information about a single enzyme function to resolved structures, reactions and pathways, substrates and products, relevance to disease and relevant publications. Users can also search the Enzyme Portal by protein sequence.

IntAct (43), EMBL-EBI's database of molecular interactions, is now closely integrated with the MINT database at the University of Rome, and is serving as the curation platform for eight global partner organizations based in Canada, India, Ireland, Italy, Singapore, UK and the USA through the IMEx Consortium (7).

EMBL-EBI, the Ontario Institute of Cancer Research and New York University Medical Center jointly develop the Reactome (44,45) database of curated human pathways. Reactome's website (http://www.reactome.org) has been completely redeveloped, and now features a modular pathway browser, a comprehensive set of web services and integrated molecular-interaction, structural and expression data.

Submissions to the BioModels database (46), EMBL-EBI's database of computational models of biological processes, have more than doubled since 2011. A new 'top-down' approach to building quantitative models has been implemented. Rather than building up from the mechanistic details of a specific process, the new approach uses pathways from data resources such as Kyoto Encyclopedia of Genes and Genomes and Reactome (44) as starting points. The BioModels database is also one of several EMBL-EBI databases that are actively involved in exposing their data to the semantic web: there is now a resource description framework (RDF) representation of the models in the database, and users have access to powerful queries using SPARQL.

CHEMICAL BIOLOGY

Chemical biology has been a major growth area for EMBL-EBI in the past decade. Major drivers for this growth have included the emergence and maturation of computational systems biology (47) and public investment in computational approaches to drug discovery (48,49).

ChEMBL (50), EMBL-EBI's database of drugs and bioactive entities, now has a unified chemistry resource lookup and registration system called UniChem (51). The ChEMBL team supports the neglected-disease community, and now provides one-stop access to all data from the Medicines for Malaria Venture's open-access MalariaBox (52) and other open-access malaria research efforts, including new high-value malaria and tuberculosis data sets. A version of ChEMBL has been built using only open-source software, and this has been made available as a virtual machine.

ChEBI, a resource for reference chemical structures, nomenclature and ontological classification, now offers a standalone tool for classifying compounds that resemble natural products. Several thousand natural products have been added to the database, and a new version of the SENECA tool, which helps to elucidate structures for natural products, has been implemented. ChEBI has also introduced new tools for searching (OntoQuery) and analysis (BiNChE) of information contained in the ChEBI Ontology.

USER TRAINING

It is essential that our users can access EMBL-EBI's data efficiently and get the most out of their own datasets when comparing them with the public record. To that end, EMBL-EBI provides an extensive user training programme (http://www.ebi.ac.uk/training), coordinated and funded centrally, but with input from all the resource

teams. In turn, as training activities offer a unique interface between service developers and users, they are invaluable in the evolution of existing resources and the creation of new ones. EMBL-EBI's diversifying community of users is reflected in its user training offering. The programme, courses and materials are created in response to user demand, and cover the full spectrum of EMBL-EBI's activities.

Face-to-face training courses reach ~4000 users a year—a small fraction of EMBL-EBI's user community. Train online (http://www.ebi.ac.uk/training/online/), EMBL-EBI's eLearning resource launched in 2011, supplies on-demand instruction to users the world over, in the form of free short courses designed for bench-based biologists. Quick tours provide an overview of each of the core data resources and show users where to go for more information. Introductory courses explain some of the important concepts behind the bioinformatics resources and introduce key subject areas, such as functional genomics. Walk-through courses provide a more in-depth exploration of a resource, structured as tutorials with use cases, guided examples and quizzes. Finally, video courses are based on some of our most popular face-to-face training courses. They provide video lectures and accompanying course materials.

Simply being able to discover relevant training courses is challenging for many life scientists, and EMBL-EBI's training team has entered the realm of 'training informatics' through its involvement in the EMTRAIN project, which launched on-course® (http://www.on-course.eu)—a new resource for course seekers in the biomedical sciences—in June 2012 (53).

CONCLUDING REMARKS

The foundations of EMBL-EBI's data collection are comprehensive archival collections of biomolecular information, and our expanding and diversifying user base demands that we serve a growing number of specialized communities. To support research in the applied sciences, we must provide access to data from medical sequencing projects, with appropriately consented access to the data. EMBL-EBI is dedicated to remaining user-focused and developing interfaces and training opportunities that enable discovery in all areas of molecular biology. Our continuous interaction with our users is the driver that enables us to feed back to our developers, which, in turn, helps our services to remain relevant to our users' needs.

ACKNOWLEDGEMENTS

The EMBL-EBI is indebted to the support of its funders: EMBL's member states, the European Commission, the Wellcome Trust, the UK Research Councils, the US National Institutes of Health and our industry partners. The authors are also indebted to hundreds of thousands of scientists who have submitted data and annotation to the shared data collections. The authors would like to thank the many colleagues who provided input to this manuscript.

FUNDING

Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Marx,V. (2013) Biology: the big challenges of big data. *Nature*, **498**, 255–260.
2. Karsenti,E., Acinas,S.G., Bork,P., Bowler,C., De Vargas,C., Raes,J., Sullivan,M., Arendt,D., Benzoni,F., Claverie,J.M. *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biol.*, **9**, e1001177.
3. Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2013) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **41**, D21–D24.
4. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
5. The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
6. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
7. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S., Cesareni,G. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
8. Vizcaíno,J.A., Côté,R.G., Csordas,A., Dianes,J.A., Fabregat,A., Foster,J.M., Griss,J., Alpi,E., Birim,M., Contell,J. *et al.* (2013) The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.*, **41**, D1063–D1069.
9. Salek,R.M., Haug,K. and Steinbeck,C. (2013) Dissemination of metabolomics results: role of MetaboLights and COSMOS. *Gigascience*, **2**, 8.
10. Hayden,E.C. (2013) Geneticists push for global data-sharing. *Nature*, **498**, 16–17.
11. Pavelin,K., Cham,J.A., De Matos,P., Brooksbank,C., Cameron,G. and Steinbeck,C. (2012) Bioinformatics meets user-centred design: a perspective. *PLoS Comp. Biol.*, **8**, e1002554.
12. Javahery,H., Seffah,A. and Radhakrishnan,T. (2004) Beyond power: making bioinformatics tools user-centered. *Commun. ACM*, **47**, 58–63.
13. Bolchini,D., Finkelstein,A., Perrone,V. and Nagl,S. (2009) Better bioinformatics through usability analysis. *Bioinformatics*, **25**, 406–412.
14. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
15. McEntyre,J.R., Ananiadou,S., Andrews,S., Black,W.J., Boulderstone,R., Buttery,P., Chaplin,D., Chevuru,S., Cobley,N., Coleman,L.A. *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.*, **39**, D58–D65.
16. Gostev,M., Faulconbridge,A., Brandizi,M., Fernandez-Banet,J., Sarkans,U., Brazma,A. and Parkinson,H. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.*, **40**, D64–D70.
17. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
18. The Gene Ontology Consortium. (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.

19. Hill,D.P., Adams,N., Bada,M., Batchelor,C., Berardini,T.Z., Dietze,H., Drabkin,H.J., Ennis,M., Foulger,R.E., Harris,M.A. *et al.* (2013) Dovetailing biology and chemistry: integrating the gene ontology with the ChEBI chemical ontology. *BMC Genomics*, **14**, 513.
20. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
21. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European nucleotide archive. *Nucleic Acids Res.*, **41**, D30–D35.
22. Hsi-Yang Fritz,M., Leinonen,R., Cochrane,G. and Birney,E. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, **21**, 734–740.
23. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics*, **26**, 2069–2070.
24. Scally,A., Dutheil,J.Y., Hillier,L.W., Jordan,G.E., Goodhead,I., Herrero,J., Hobolth,A., Lappalainen,T., Mailund,T., Marques-Bonet,T. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.
25. The ENCODE Consortium, Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
26. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
27. The Heliconius Genome Consortium. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.
28. The International Barley Genome Sequencing Consortium, Mayer,K.F., Waugh,R., Brown,J.W., Schulman,A., Langridge,P., Platzer,M., Fincher,G.B., Muehlbauer,G.J., Sato,K. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
29. Brenchley,R., Spannagl,M., Pfeifer,M., Barker,G.L., D'Amore,R., Allen,A.M., McKenzie,N., Kramer,M., Kerhornou,A., Bolser,D. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.
30. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Xu,M., Flicek,P., Manolio,T. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-Trait associations. *Nucleic Acids Res.*, [epub ahead of print].
31. Koscielny,G., Yaikhom,G., Iyer,V., Meehan,T.F., Morgan,H., Atienza-Herrero,J., Blake,A., Chen,C.-K., Easty,R., Di Fenza,A. *et al.* (2014) The International Mouse Phenotyping Consortium (IMPC) web portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.*, [epub ahead of print].
32. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.
33. Rustici,G., Kolesnikov,N., Brandizi,M., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Ison,J., Keays,M. *et al.* (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41**, D987–D990.
34. Vizcaíno,J.A., Côté,R.G., Csordas,A., Dianes,J.A., Fabregat,A., Foster,J.M., Griss,J., Alpi,E., Birim,M., Contell,J. *et al.* (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.*, **41**, D1063–D1069.
35. Haug,K., Salek,R.M., Conesa,P., Hastings,J., de Matos,P., Rijnbeek,M., Mahendraker,T., Williams,M., Neumann,S., Rocca-Serra,P. *et al.* (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.
36. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
37. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
38. Gutmanas,A., Oldfield,T.J., Patwardhan,A., Sen,S., Velankar,S. and Kleywegt,G.J. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. *Acta Crystallogr. D Biol. Crystallogr.*, **69**, 710–721.
39. Velankar,S., Alhroub,Y., Best,C., Caboche,S., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Golovin,A., Gore,S.P. *et al.* (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **40**, D445–D452.
40. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfield,T.J., O'Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
41. de Matos,P., Cham,J.A., Cao,H., Alcántara,R., Rowland,F., Lopez,R. and Steinbeck,C. (2013) The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. *BMC Bioinformatics*, **14**, 103.
42. Alcántara,R., Onwubiko,J., Cao,H., Matos,P.D., Cham,J.A., Jacobsen,J., Holliday,G.L., Fischer,J.D., Rahman,S.A., Jassal,B. *et al.* (2013) The EBI enzyme portal. *Nucleic Acids Res.*, **41**, D773–D780.
43. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
44. D'Eustachio,P. (2013) Pathway databases: making chemical and biological sense of the genomic data flood. *Chem. Biol.*, **20**, 629–635.
45. Haw,R. and Stein,L. (2012) Using the Reactome database. *Curr. Protoc. Bioinform.*, 38:8.7.1-8.7.23.
46. Chelliah,V., Laibe,C. and Le Novère,N. (2013) BioModels Database: a repository of mathematical models of biological processes. *Methods Mol. Biol.*, **1021**, 189–199.
47. Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
48. Årdal,C. and Røttingen,J.A. (2012) Open source drug discovery in practice: a case study. *PLoS Negl. Trop. Dis.*, **6**, e1827.
49. Williams,A.J., Wilbanks,J. and Ekins,S. (2012) Why open drug discovery needs four simple rules for licensing data and models. *PLoS Comp. Biol.*, **8**, e1002706.
50. Willighagen,E.L., Waagmeester,A., Spjuth,O., Ansell,P., Williams,A.J., Tkachenko,V., Hastings,J., Chen,B. and Wild,D.J. (2013) The ChEMBL database as linked open data. *J. Cheminform.*, **5**, 23.
51. Chambers,J., Davies,M., Gaulton,A., Hersey,A., Velankar,S., Petryszak,R., Hastings,J., Bellis,L., McGlinchey,S. and Overington,J.P. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, **5**, 3.
52. Spangenberg,T., Burrows,J.N., Kowalczyk,P., McDonald,S., Wells,T.N. and Willis,P. (2013) The open access malaria box: a drug discovery catalyst for neglected diseases. *PloS One*, **8**, e62906.
53. Payton,A., Janko,C., Renn,O., Hardman,M. and EMTRAIN Consortium. (2013) on-course® portal: a tool for in-service training and career development for biomedical scientists. *Drug Discov. Today*, **18**, 803–806.

APPENDIX 1

EMBL-EBI's mission

- To provide freely available data and bioinformatics services to all facets of the scientific community.
- To contribute to the advancement of biology through basic investigator-driven research.
- To provide advanced bioinformatics training to scientists at all levels.
- To help disseminate cutting-edge technologies to industry.
- To coordinate biological data provision throughout Europe.

APPENDIX 2

EMBL-EBI's principles of service provision

Open: Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends on research consent agreements.

Compatible: EMBL-EBI is a world leader in the development of global bioinformatics standards, which are key to data sharing.

Comprehensive: Thanks to our many data-sharing agreements, EMBL-EBI resources are comprehensive and up-to-date. We work with publishers to ensure that biological data must be placed in a public repository and cross-referenced in the relevant publication.

Portable: All of our data and many of our software systems can be downloaded and installed locally.

High quality: Our databases are enhanced through annotation: highly qualified biologists add value to databases by incorporating features of genes or proteins from other sources, and automated annotation is subjected to rigorous quality control.