EMBL Programme 2017-2021

Table of Contents

List of boxes

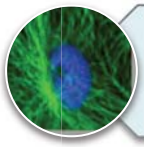
B. Research

C. Services

E. Technology Development, Technology Transfer and Interaction with Industry

A. Introduction & Summary

The European Molecular Biology Laboratory (EMBL) is a unique institution. It is an intergovernmental organisation and centre of excellence in scientific research and training set up by its member states to promote the molecular life sciences in Europe and beyond. We pursue a set of missions that, in combination, allow EMBL to complement and provide added value to the activities of member state scientific communities. EMBL's five missions are:



1. Forefront research: uncovering the molecular basis of life

2. Providing world-class research infrastructure and services

3. Training and inspiring the next generation of stellar scientists

4. Driving research, innovation and progress through technology development, interactions with industry and technology transfer

5. Taking a leading role in the integration of life science research in Europe

EMBL is also unique in the widespread use of fixed-term contracts; only around 11% of its staff members currently have open-ended contracts. This turnover system generates a constant supply of superbly trained, internationally networked research, technical and administrative staff members who leave EMBL to take up positions in its member and associate member states, enriching the national communities and promoting collaboration in Europe. In addition, turnover ensures that EMBL's activities are constantly undergoing renewal and updating on a scale that is not possible in other research-based institutions. In the past five years 25% of EMBL's faculty has turned over, which has led to the acquisition of new skills and expertise in many areas.

Figure A.1 EMBL has experienced 25% of faculty turnover in five years. This translates into a turnover of expertise, and constant renewal and rejuvenation of the Laboratory's scientific profile.

In addition to the continuous staff turnover, every five years we undertake a major strategic planning exercise to develop an EMBL Programme, a blueprint of planned activities that covers all five missions for the upcoming five-year period, in this case 2017-2021. The EMBL Programme is developed with input from all parts of the organisation, the scientific faculty and senior technical, service and administrative staff. It forms the basis for discussion of a budget, or Indicative Scheme, to align EMBL's future activities with the collective wishes of our member and associate member states and ultimately to decide on funding for them. As forefront biomedical research leads to new discoveries and rapid and sometimes disruptive change, this future plan is never predictable in detail. This means that the EMBL Programme serves as a broad guideline rather than a detailed directive of how EMBL will develop in the next five years. Our service portfolio and integration activities will be modified by demands coming from our member states. In our research, we rely on the scientific curiosity and creativity of our young faculty and therefore must allow them the freedom to pursue the topics that interest them most. The spirit of discovery together with breakthrough technology developments and the unique collaborative opportunities available at EMBL are the driving forces that ultimately determine the scientific challenges that EMBL will tackle in the future.

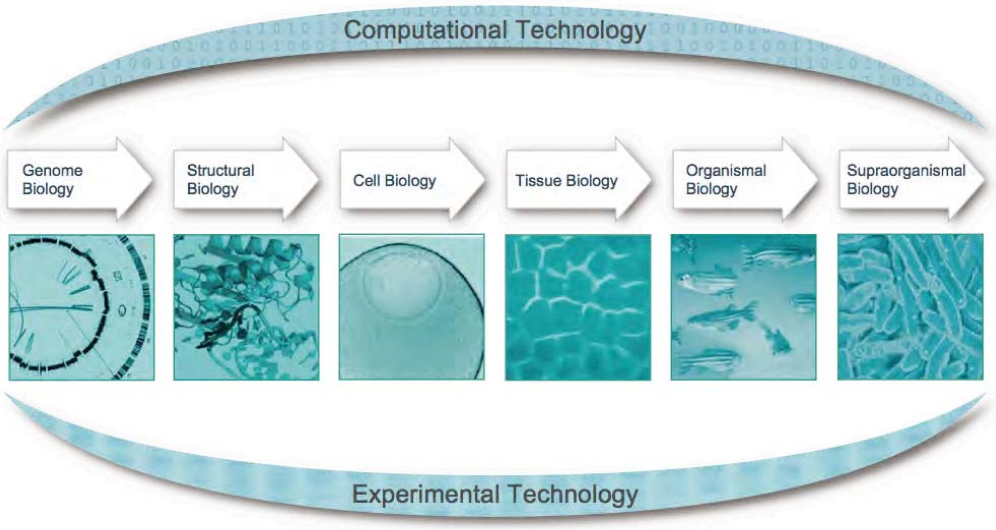
We are preparing the EMBL Programme 2017-2021 as we enter a new and remarkably exciting era of life science research. Molecular Biology is becoming a fully quantitative discipline across the whole spectrum of life's organisation, all the way from the genotype to phenotype, from molecules and complexes over cells and tissues to whole organisms, communities and ecosystems. This 'molecular revolution', is comparable to the digitisation of all sources of information in our society and it comes with equally vast opportunities and challenges. The digital data revolution already affects a broad public. Computational analysis has led to reasonably accurate predictive models of extremely complex systems ranging from weather forecasting through to consumer behaviour. Similarly, the molecular revolution, summarised in the title of our 2017 – 2021 Programme, 'Digital Biology', will allow us to understand and make predictions about proteins, protein

complexes, supramolecular structures, cells, organisms and, increasingly, organismal communities in much the same way. These advances will ultimately lead to new ways to understand, model and predict human health and intervene in disease.

However, we are also aware that we are preparing the current Programme against the background of a period in which research and research infrastructure budgets at EMBL and in most of our member states have either stagnated or even decreased over an extended period, as a result of the economic crisis. This situation makes it even more important than usual that the EMBL Programme 2017 - 2021 provides added value to our member states.

In preparing the Programme we have responded to this challenge in two main ways. First, we have focused on EMBL's unique strengths – these lie in our interdisciplinary collaborative culture and our integrative and multi-scale approach to systematically understand how living systems arise from their molecular building blocks in terms of structure and function. Our expertise in areas like quantitative multi-scale imaging, integrative molecular data analysis and computational approaches to massive data handling, equips EMBL to drive progress in the new era of Digital Biology in a unique way.

Figure A.2 Digital Biology at EMBL 2017-2021. EMBL's research covers the whole spectrum of life's organisation from genomes over molecules and cells to organisms and beyond towards communities. Research at EMBL bridges these scales thanks to cutting-edge experimental technologies that visualise and quantify molecular processes, and computational technologies that integrate data across different scales of organisation.



Second, EMBL's main priority for the period 2017-2021 is to strengthen and expand our service portfolio for the benefit of users from the scientific communities in our member states. Recent technological progress in several areas of the life sciences, coupled with financial limitations in research budgets, have led to an ever-increasing demand from member state researchers that EMBL should increase its provision of access to interesting new technologies that are otherwise unavailable to them. Examples include the new ultra-high resolution electron microscopy methods, combinations of novel methods in

imaging technology, some of them developed at EMBL, and use for biological experiments of the European X-ray Free Electron Laser (European XFEL) that will be available in Hamburg from 2017. These data-intensive technologies all require new computational tools and infrastructures for data handling, storage and analysis to integrate them with the massive amounts of genomic and other data being produced by biomedical scientists. We therefore intend to ask the member states to provide us with the funding that will be required to meet the demands of their research communities for access to new and cutting-edge technologies and infrastructures in these areas.

Box A.1: EMBL's activities at a glance

EMBL is a leader in basic life science research

- **650 publications** per year on average (2012-2014) whose overall quality places us in the top 10 research institutions worldwide (Source: Scimago)
- **4 EMBL publications** among the top 100 highest cited papers ever
    - One of them in the top ten
- **75% of EMBL's** close to **2000 scientific publications** 2012-2014 have been published jointly with **more than 800 other institutions** worldwide
    - **65%** of EMBL publications are collaborations with **member state institutions**
- **21 ERC** (European Research Council) grant awards (status 2015)
    - ≈ 30% of EMBL research group leaders

EMBL is a pioneer in the development of technology and instrumentation

- Advanced imaging technologies, including light sheet microscopy (SPIM) and correlative light and electron microscopy techniques
- World-leading software for imaging, genomics and structural biology data analysis
- Innovative synchrotron beamline technology and automation

EMBL provides world-class services and cutting-edge infrastructures

- **11 million web requests** per day for EMBL-EBI data resources
- **2500 structural biology user visits** on average per year at the facilities in Hamburg and Grenoble
- **1000 Core Facility users** (350 of those from outside EMBL) gain access to cutting-edge technologies and learn about setting up similar equipment and facilities in their home institutions
- **500 visitors** per year on average access technologies, learn methods or carry out collaborative projects at EMBL

EMBL is a hotbed of innovation

- **17 EMBL spin-out** companies since 1997
- **46 patents** granted (2012-2014)
- **800 license & collaboration contracts** concluded (2012-2014)

EMBL trains Europe's next generation of leading researchers

- **230 PhD students & 250 postdoctoral fellows** steady state
    - 95% of PhD students successfully submit their thesis
    - 50 PhD graduations every year
- **6700 EMBL alumni**
    - Over 80% of EMBL alumni move on to one of the member states
    - 38% of them are now in senior positions
    - 76% in academia and 13% in industry

EMBL is a platform for training and scientific exchange

- **180 courses and conferences** organised across all EMBL sites (2012-2014)
- Close to **18,000 participants** (2012-2014)

EMBL is an instrument for European integration

- 21 member states, 2 associate member states, 4 prospect member states
- EMBL **coordinated 20** and **participated in 88 collaborative projects** funded by the European Commission Framework Programmes 2012-2014

EMBL serves as a role model of scientific organisations

- 9 EMBL partner institutions in 8 member states have implemented the successful EMBL model

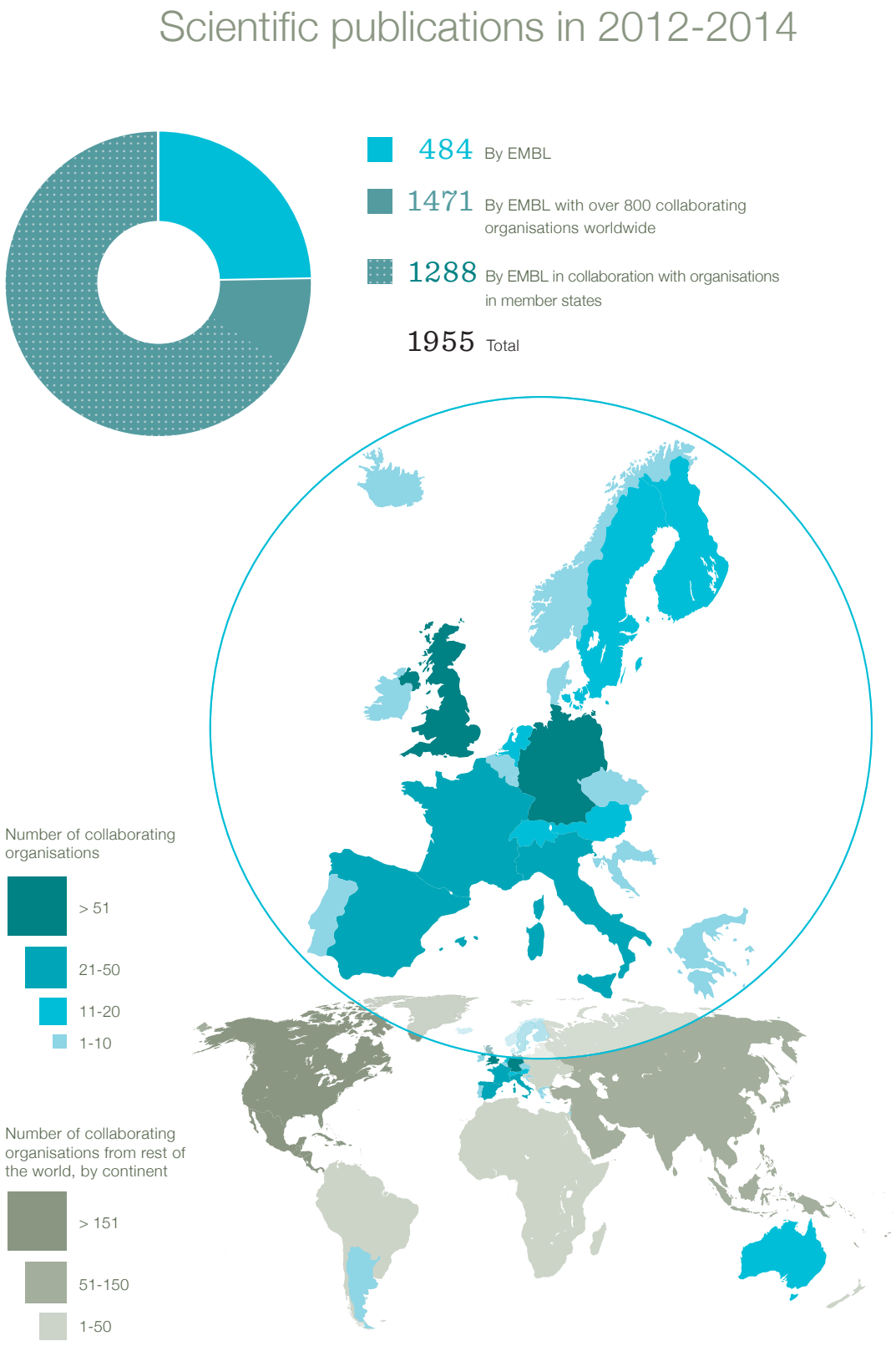


Figure A3 EMBL's scientific publications 2012-2014. 65% of the close to 2000 papers published resulted from collaborations with institutes in the EMBL member states.

1. EMBL's place in Europe

At a time in which Europe and the way its states work together as a Union in a range of different fields – political, economical, social and not least scientific - is frequently questioned, EMBL serves as a success story that demonstrates the remarkable progress that can be achieved when European countries work together. EMBL was founded 40 years ago to create a European centre of excellence for molecular biology. Today this mission has been more than accomplished: from an organisation of a few hundred biologists EMBL has grown into an international venture with over 1800 staff (representing over 80 nationalities) across five sites in Europe. Our member states have increased from 10 founding to 21 current member states and two non-European associate members. EMBL has developed into Europe's hub for the life sciences (Box A.1), with a set of missions that goes well beyond research and is designed to add maximal value to our member states. EMBL's five missions, and our goal to excel in all of them, are key to the success of the laboratory. The missions produce synergy, benefitting from each other to create a multiplier effect. A major advance in technology development, for example, will not only drive forward our research programme but also benefit the user communities of our scientific services, lead to interaction with industry and inspire new training courses.

Figure A.4 EMBL's 21 member states and 2 associate member states (dark green) and 4 prospect member states (light green).

This comprehensive spectrum of activities, which both supplement and foster related activities in the national research communities, allows EMBL to serve member countries that face different national challenges in the life sciences. We do so through a system that is not based on the notion of juste retour but that instead allows us to tailor our activity to countries' needs. While some countries gain most from EMBL membership through access to research infrastructure and technologies not available in their national institutions, others benefit more from the specialised training programmes offered by EMBL, by attracting large numbers of our alumni back into their national systems or by setting up an EMBL partnership institution. This responsive way of operating ensures that EMBL

provides maximal value to each of its members and the success of the system is illustrated by the continued interest of countries within and beyond Europe to join EMBL.

The continued support of our member states, in particular during periods of financial constraint in many European countries, is another testimony to EMBL's value. We are very grateful for this continued support. It is essential that the sustainable funding and long-term commitment of our member states allows EMBL to pursue groundbreaking projects, invest in large-scale infrastructures, attract the very best scientists and thereby ensure that EMBL maintains its leading position in the life sciences and can provide maximum benefit back to Europe and its member states in return.

2. Highlights of EMBL strategy 2017-2021

This section highlights those aspects of EMBL's five missions where we are expecting to see change, including potential changes in investment, in the course of the next Indicative Scheme. The large majority of EMBL's activities will build on our unique strengths and utilise our turnover system to pioneer new directions in support of EMBL's five missions. In the later chapters of this document we provide a report of the considerable progress made over the first three years of the current Programme and outline the future plans for all of EMBL's activities for the period 2017-2021 in more detail.

2.1 Forefront life science research: uncovering the molecular basis of life

EMBL is one of the world's top research institutions in the molecular life sciences. Excellence is the guiding principle for all our research. We achieve excellence by applying strict criteria to recruiting only the best young scientists, by mentoring and training them to realise their full potential and by subjecting them to regular, stringent review by internationally leading experts under the supervision of EMBL's Scientific Advisory Committee.

Benefitting our central mission to understand how life arises from molecular building blocks, EMBL's research portfolio comprehensively covers our mission, ranging from elucidating the structures and functions of individual molecules to an understanding of how complex molecular systems organise cells, organs, tissues, organisms and organismal communities. Underpinning EMBL's ability to work across all scales of biological organisation is the interdisciplinary composition of our research base. Only just over half of EMBL's researchers are biologists, the other half is a mix of chemists, physicists, medical doctors, computer scientists, engineers and mathematicians. EMBL actively promotes interdisciplinary collaboration among its researchers to leverage synergies between their diverse skill sets.

Digital Biology

We have chosen ‘Digital Biology’ as the theme for the EMBL Programme 2017-2021 because it unites our increasing ability to quantitatively assess biological processes at the molecular level with the ensuing opportunities and challenges in data analysis and integration. EMBL has unique strengths in both of these areas. Our scientists have been at the forefront of developing, advancing and applying the experimental approaches that underpin the molecular revolution by producing quantitative data. This is complemented by world-leading computational expertise in the life sciences, a field that EMBL entered early and that we have considerably strengthened and expanded across the whole Laboratory as part of the current EMBL Programme ‘Information Biology’. This unique combination of experimental and computational skills forms the foundation on which EMBL will drive the molecular revolution forward in the era of Digital Biology.

A major result of recent advances in molecular measurement and computational technologies is that human biology has become more and more accessible to researchers at EMBL. As a consequence, molecular and cellular research on human systems has developed from a small activity of a few research groups to a major cross-cutting research theme in our Programme 2017-2021. We welcome this change, in particular because it allows us to contribute more and more directly to advancing medicine while staying focused on our basic research mandate. Medicine and molecular biology are on converging paths with diagnosis increasingly moving from the phenotypic description of symptoms towards the molecular characterisation of patients and disease states. EMBL is not itself a translational research institute, but its focus on molecular and cellular technology, coupled with its experience in data management and innovative integrative analysis, make it an obvious player to help bridge the worlds of basic, translational and clinical research in the future. In addition, the training available at EMBL is exactly what is required to provide the knowledge that will support medical treatments and practitioners as medicine becomes increasingly molecular and digitised.

Within the overall context of Digital Biology we have identified five broad research themes on which EMBL will focus over the next indicative scheme. These are detailed in Section B.2 of this Programme and have been developed to complement rather than duplicate national research strengths in our member states. We would like to highlight the following proposed new research activities:

Tissue Biology and Disease Modelling

Following an in-principle decision by EMBL Council in 2014, EMBL will create a new outstation for tissue biology and disease modelling in Barcelona, Spain. The new outstation will leverage expertise in the areas of multicellular tissue/organ-level imaging, image-driven modeling across multiple scales and experimental genetics on model species and manipulatable human systems to understand phenotypic variability at the level of tissues and organs. Research will concentrate on the ‘mesoscopic’ scale, that links molecular and cellular biology to larger-scale physiology, and will thereby complement research activities in other EMBL Units and enrich our imaging technology portfolio. The new outstation will also operate a service facility that provides access to mesoscopic imaging technologies and services in image-driven modelling to the external scientific community.

Neurobiology

Over the next Indicative Scheme, EMBL plans to take advantage of considerable upcoming turnover to refocus the Mouse Biology Unit in Monterotondo on neuroscience and the role of epigenetics in neurobiology. This will necessitate some ‘re-tooling’ of the outstation and its support facilities and, to create sustainable critical mass, the current 6 groups and 2 teams in Monterotondo should be gradually increased to 10 groups/teams. Research at the interface of neurobiology and epigenetics is important to advance our understanding of the development of neural cell types and the role of genomic plasticity and genetic variation in neurodegenerative and neurological disorders as well as in behavioural adaptation. Recent recruitment has already begun to strengthen both neurobiological and epigenetic work at EMBL’s Mouse Biology Unit. Due to EMBL’s investigator-driven approach to research and very stringent standards of recruitment, it is difficult to make exact, detailed predictions about future research directions at this stage. We will aim to recruit group leaders who can operate at the interface between epigenetics and neuroscience and who can make good use of the computational, genomics, structural biology and imaging strengths available in other EMBL Units and thereby provide unique contributions to the fields of neurobiology and epigenetics.

2.2 Providing world-class research infrastructure and services

EMBL operates unique, world-leading research infrastructures that enable European researchers to achieve breakthroughs across many scientific disciplines. Building on our scientific strengths, our experience in service provision and aiming to address current pressing needs among the scientific community in our member states, EMBL proposes the following future plans for its service mission:

Bioinformatics services

The European Bioinformatics Institute (EMBL-EBI) is the world’s most comprehensive source of biological and biomolecular data. Its website receives over 11 million requests per day (compared to 5 million in 2010) and serves a broad and growing community of commercial and academic researchers throughout Europe and the world. To cope with life science data production, EMBL-EBI's total disk storage has had to increase five-fold since 2010. The trend towards Digital Biology, new technologies and resulting new data types, translates into new challenges regarding data processing, storage, and analysis. We expect genetic variation data, single-cell analysis and the growing need to provide reference biological datasets for use in medicine to have a major impact on our Bioinformatics services over the next indicative scheme. Moreover, dealing with the rapidly increasing wealth of image data will become a priority over the next years. EMBL is an obvious place to lead the development of image data repositories and standards that allow effective data sharing and the integration of images with other data types in Europe, because it combines expertise in developing data standards and resources with in-depth understanding of imaging technologies and the needs of imaging communities.

All of these developments have implications for IT infrastructure and staff growth at EMBL-EBI.

Over the coming decades the challenge of providing effective bioinformatics infrastructure will continue to grow. To meet this challenge, EMBL led the development of ELIXIR, a distributed European infrastructure for biological data that is now governed by an independent international collaboration agreement. EMBL-EBI hosts the central, coordinating ELIXIR hub and many of the EMBL-EBI services funded by the EMBL member states will be integrated with other ELIXIR services developed in national nodes. In this way, EMBL will work closely together with ELIXIR to share the data provision tasks as efficiently as possible. It is possible that ELIXIR will commission additional activities at EMBL-EBI to help drive aspects of this integration but it is clear that ELIXIR will not provide funding for any of the ongoing EMBL-EBI core activities and responsibilities.

IT infrastructure and services

EMBL has particularly high demands for IT infrastructure and service support. The biggest IT requirement (roughly 80%) stems from EMBL-EBI's bioinformatics services for millions of users and its function as the largest European centre for the storage, usage and distribution of digital biomolecular data, as described in the preceding section. At EMBL Heidelberg, extensive activities in genomics, imaging and computational biology also place considerable demands on IT infrastructure. The two sites between them currently manage over 60 Petabytes of data and make use of over 60,000 CPU cores. By 2021, EMBL's IT infrastructure and staff will need to handle Exabytes of data and, even allowing for expected developments in processing capacity per core, the number of CPU cores is expected to rise by a factor of at least four. High-bandwidth connectivity to external networks is essential to enable the use and sharing of data in these quantities and its provision, particularly to the EMBL Heidelberg site, is an urgent necessity.

X-ray based structural biology services

At the X-ray based structural biology sites in Hamburg and Grenoble, EMBL plans to make a new generation of unique infrastructures available to the life science community. The 'ESRF phase II upgrade' will convert the Grenoble synchrotron ring towards being the first diffraction-limited synchrotron facility worldwide by 2019. Helping exploit this development for structural biology will be a major task in Grenoble. In Hamburg, in addition to a continuation of the development of the EMBL@PETRA3 beamlines and structural biology support infrastructure towards robust service functionality, the European X-ray Free Electron Laser (European XFEL), the most powerful X-ray laser in the world, will start operation in 2017. XFEL has the potential to revolutionise structural biology in a similar way to the first use of synchrotron radiation 40 years ago on the site of the then-new EMBL Hamburg outstation. EMBL is eager to help make this opportunity available to a broad user community in Europe and is therefore playing an important role in the development of a dedicated sample preparation and characterisation facility for life science users at the European XFEL as part of an international consortium. We expect this facility to become operational from the beginning of the next Programme period and will ask the EMBL member states to fund staff positions required for its operation.

Access to high-resolution electron and light microscopy

Diverse recent technical developments in electron and light microscopy have greatly increased both the resolution that each of the methods can individually reach as well as the possibility to use both methods correlatively. These advances have led to an enormous demand in the European life science community for access to the new technologies and support in the accompanying sample preparation, data acquisition and image processing techniques. Imaging data is also a frontier where new resources and software must be developed to allow data storage and provision to the community. EMBL, with its leading expertise in both electron and light microscopy development and use and its broad experience in providing services and data resources for the European life science community, proposes to address this demand through an extension of its EMBL Heidelberg and EMBL-EBI-based services to external users. In order to create a multiplier effect that goes beyond simply providing access to a technology that is in high demand, the proposed facility will, like all EMBL facilities, have the mission to train users in the operation of the technology as well as in how to set up and operate similar facilities in their home countries.

Centre for Integrative Structural Modelling

EMBL Centres are horizontal activities that concentrate know-how in specific topics and techniques. They are an important resource to enable the integration of methods and insights from multiple disciplines and to combine experimental and computational approaches. To support EMBL researchers using integrated structural biology methods for research at the interface between the molecular and cellular scales, to advise the staff responsible for the molecular and cellular imaging facilities referred to above and elsewhere in this document, and to participate in both internal and external training in this area, EMBL plans to establish a new EMBL Centre for Integrative Structural Modelling.

Core Facilities

EMBL's Core Facilities serve internal technology needs and, where capacity permits, are available to users from the member states. Around one third of our 1000 annual Core Facility users are external users. In addition to providing scientists with access to cutting-edge technologies, EMBL's Core Facilities have an important training function. They not only train users in operating equipment, but also support them in setting up comparable equipment and facilities in their home institutions. To cater for the growing need to integrate studies of metabolism and metabolites, we will establish a Metabolomics Core Facility to provide services in quantitative analyses of small molecules based on mass spectrometry, imaging mass spectrometry and nuclear magnetic resonance methods. Apart from cutting-edge mass spectrometry equipment, the Core Facility will require expert staff to provide services.

Support for user access to EMBL facilities

User access to most of EMBL's service activities has been supported by external funding from the European Commission for many years. Unfortunately, overall European funding for user access has diminished in recent years and is now inadequate to meet the needs of the life science community, particularly for access to new technologies such as advanced electron microscopy. We believe

the EMBL member states should consider whether providing such funding for access to EMBL facilities should be a priority. It would allow all member state scientists to continue to make use of the cutting-edge infrastructures and technologies provided by EMBL in future.

2.3 Training and inspiring the next generation of stellar scientists

EMBL is a centre of excellence for training young scientists and has over its 40 years of history helped launch the careers of several thousand life scientists, many of whom now have leading positions in academia or industry in our member states. EMBL trains doctoral and postdoctoral fellows in world-leading programmes as well as young principal investigators (PIs) who, for the first time, independently lead research groups or service teams. Our internal training programmes frequently serve as best practice examples after which similar programmes in our member states are modelled.

To continue to attract the most talented young scientists, endow them with the best preparation for their future careers and provide our member states with highly skilled future scientific leaders, we constantly work to improve our training programmes. EMBL has paved the way towards truly interdisciplinary training with the establishment of its EMBL Interdisciplinary Postdoctoral (EIPOD) Programme, which we are planning to further refine in the period 2017-2021. At the time of writing, we have just received confirmation of a third consecutive round of Marie Sklodowska-Curie co-funding for the EIPOD programme from the EC. Such extended continuous EC funding is very exceptional and reflects the quality and uniqueness of the interdisciplinary training on offer. Together with national partners we have also begun to establish further specialised training schemes, which place fellows at the interface between basic and clinical or wet-lab and computational research.

Apart from training for our employees, external scientific training is another crucial service we provide to our member states. We provide ‘continuing professional development’ or user training in scientific methodology. Over the past three years more than 18,000 scientists from many sectors, from masters students to senior scientists, have been trained in 180 courses and conferences. Many scientists employed in the life science industries also benefit from our training and EMBL thereby contributes to strengthening the growth of this sector in Europe. In this Programme we will continue to develop and improve the quality and breadth of our external training, ensuring that it remains at the forefront of scientific and technological development.

2.4 Driving research, innovation and progress through technology development, interactions with industry and technology transfer

EMBL excels both in discovery research and in innovative development, particularly in the areas of imaging, structural biology instrumentation and computational technology, that drives progress in life science research at EMBL

and beyond. To make our discoveries and inventions broadly available, to benefit scientists, industry and society at large, EMBL and its knowledge transfer arm EMBLEM interact closely with industry in multiple ways, including actively engaging in technology transfer per se. During the 2017 – 2021 period we envisage two new activities in this area.

Imaging Technology Development Platform

The demand for EMBL's innovative imaging technologies among European researchers is high, but at present there is no effective mechanism for EMBL to provide access to these technologies before they are commercialised and can be made available through a Core Facility, which frequently takes 5-10 years. EMBL wishes, provided funding is available, to address this gap through the creation of a new Imaging Technologies Development Platform. The platform will employ staff that will provide both the hardware and software expertise to duplicate advances made in research groups at EMBL in a service setting and thus provide access to these imaging technologies to a wider community of users prior to commercialisation. The platform will focus on technologies invented and/or developed at EMBL, initially ranging from light-sheet and super-resolution microscopy to correlative light and electron microscopy.

Industry Relations

In the next Programme period EMBL will put further effort into forging new, mutually beneficial, strategic partnerships with industry to complement its existing ties with the private sector. Following the successful model of the recently established Centre for Therapeutic Target Validation - a strategic collaboration between EMBL-EBI, the Wellcome Trust Sanger Institute and GlaxoSmithKline - EMBL as a whole will continue to offer its expertise in areas such as data management and analysis or technology development to complement the needs of biotechnology, pharmaceutical and agri-food companies in pre-competitive research areas. In this way EMBL aims to develop additional larger and more sustained collaborations with commercial partners in order to ensure the translation of our skills and resources into innovative products that benefit society.

2.5 Taking a leading role in the integration of life science research in Europe

Because of its unique structure and role as Europe's only intergovernmental laboratory in the life sciences, promoting good relations and active interactions between EMBL and its member states is of fundamental importance to the Laboratory. EMBL actively promotes the integration of European science initiatives and helps shape science policy and strategy in Europe, as evidenced by its many interactions with member state communities (Box A1), its provision of pan-European service infrastructures and its leading role in establishing new European life science infrastructures such as ELIXIR and Euro-BioImaging. EMBL proactively engages with its member states and with the broader European and international scientific communities, European policy-making and decision-taking bodies including the EIROforum organisations (CERN, ESO, ESA, ESRF, Eurofusion, European XFEL, ILL and EMBL), the European Strategy

Forum for Research Infrastructures (ESFRI) and the European Commission as well as with the general public.

Member State Relations

EMBL's member states are very different from one another and in recognition of this our programmes and missions are designed so that we can respond to their diverse needs. The number of publications and research and service grants that are collaborative with our member states and the uptake of our infrastructures and training programmes (Box A1) reflect the fact that scientists in EMBL interact extensively with and provide important support to national communities.

To help develop and integrate Europe's scientific landscape in the life sciences and harness the scientific potential of all European states, the participation of as many European countries as possible in EMBL is desirable. EMBL is therefore actively engaged with interested non-member states. In the current Indicative Scheme, the Czech Republic joined EMBL as its 21st member state and EMBL Council also endorsed the membership of Malta (national ratification pending). Central and Eastern Europe have been and will remain a main strategic focus for EMBL's international relations and, together with EMBL and EMBC member states, a prospect membership policy was developed during the current Programme period that facilitates the accession of new countries to EMBL. Prospect membership has already led to Hungary, Poland, Slovakia and Lithuania joining EMBL, with a view to becoming full EMBL members in the course of three years. We expect more European states to join EMBL as prospect and full member states in the period of the next Programme. EMBL also engages in strategic international cooperation with countries outside Europe in the context of its Associate Membership Scheme. EMBL will seek to modestly expand its associate membership, which currently comprises Australia and Argentina, by integrating the scientific communities of non-European countries with a well-developed molecular life science programme.

Fostering gender balance in research at EMBL and beyond

Ensuring adequate gender balance among EMBL employees is a major goal of the laboratory and helps maximise the diversity of experiences, perspectives, and skills required to conduct world-class science. Increasing gender balance among EMBL personnel and fostering awareness and skills relevant to its promotion will contribute to research excellence at the Laboratory and ensure that EMBL serves as a model for basic research organisations across Europe. For this reason achieving better gender balance is one of EMBL's policy priorities for the Programme 2017-2021.

As a first step towards translating this commitment into action EMBL has established a Gender Balance Committee that promotes EMBL's culture of diversity by developing strategies, policies and other instruments aimed at:

- fostering an environment that values and supports gender equality
- achieving gender balance among EMBL's hires
- communicating EMBL's gender position internally and externally
- preparing EMBL personnel to act as role models in a gender diverse research environment, within and beyond EMBL
- measuring the effectiveness of its efforts to promote gender equality.

B. Research

EMBL Mission: Forefront research: uncovering the molecular basis of life

1. Backward look 2012–2014

The overall research framework of the EMBL 2012-2016 Programme was 'Information Biology' and centred on the challenge of unravelling the principles and logic of the complex functional organisation that defines living systems by extracting biological information from large and diverse datasets. This section illustrates the progress that has been made relative to these objectives by providing a selection of relevant research highlights. The selection has been made from close to 2000 publications produced by EMBL scientists in the period 2012-2014, since the beginning of the current Programme.

Our approach to Information Biology aimed to combine EMBL's traditional strength in mechanistic studies and hypothesis-driven research, using mostly real-time imaging, complementary structural biology technologies at different scales, and biochemical studies, with systems approaches employing unbiased genome-wide methods and computational biology. A fundamental aspect of this approach is the requirement for computational methods that allow the analysis of large biological datasets, modelling networks of gene and protein interaction and the simulation of biological systems and their defects, for example in disease states. A major strategic focus for the 2012-2016 Programme has therefore been to further strengthen EMBL's expertise in bioinformatics and computational methods. With the aid of new EMBL Centres, in the last three years the use of computational tools has increased across the Laboratory and has become standard across its research Units.

The following section provides a short overview of how far EMBL has come in its Information Biology endeavour and presents selected highlights from each of the four thematic research areas outlined in the Programme 2012-2016. Many of the research examples illustrated below present recent advances in the context of specific projects that were showcased as concrete, future-oriented research plans in the 2012-2016 Programme.

1.1 Bridging dimensions: from molecules to cells to organisms

The overarching goal of research conducted at EMBL is to bridge scales of biological organisation to gain a holistic understanding of organisms as biological systems. Thanks to the continuous and rapid development of new technologies, often spearheaded by our own scientists, and to the breadth of cross-disciplinary expertise available in-house, major milestones have been reached towards this goal in recent years. To obtain a deep and comprehensive understanding of the

properties of biological systems, scientists at EMBL have successfully combined detailed mechanistic analyses of individual functions at the molecular, cellular and organismal level with broad genomic-level approaches, thereby integrating a wide array of methods – from structural biology over biochemical to computational approaches.

1.1.1 Bridging molecular and cellular resolution: from protein-protein interactions to networks in cells

1.1.1.1 Determining molecular structures through integrated structural biology

EMBL has a traditional strength in structural biology, which is the main focus of the EMBL sites in Hamburg and Grenoble, and a core activity at the Structural and Computational Biology Unit in Heidelberg. Structure-based mechanistic understanding of cellular processes is pursued by combining multiple technologies, including macromolecular crystallography, small-angle scattering, nuclear magnetic resonance (NMR), electron microscopy and tomography, light microscopy, proteomics, chemical biology and innovative computational approaches. This integrated approach to structural biology has allowed EMBL scientists to solve a number of complex biomolecular structures over the recent years, and thus to gain structural insight into a variety of biological processes. Selected examples include:

- **Crystal structure of RNA Polymerase I**
  Scientists at EMBL Heidelberg have determined the high-resolution three-dimensional structure of RNA Polymerase I, the enzyme responsible for the production of ribosomal RNA. Their findings help explain the greater efficiency of this molecular machine relative to its better-studied counterpart RNA Polymerase II.

  Fernández-Tornero C. et al. (2013) Crystal structure of the 14-subunit RNA polymerase I. Nature 502:644-649. doi: 10.1038/nature12636

- **Architecture of the human general transcription factor TFIID core complex**
  Researchers at EMBL Grenoble uncovered the architecture of the core complex of human general transcription factor TFIID, which controls the expression of virtually all protein-coding genes in eukaryotic nuclei. For this analysis they utilised an integrated approach combining advanced protein expression techniques, proteomics, native and cross-linking mass spectroscopy, cryo-EM, and data from crystallography and homology models. This complex has long been a target of structural biologists and new protein expression technology developed in EMBL Grenoble was key to this success.

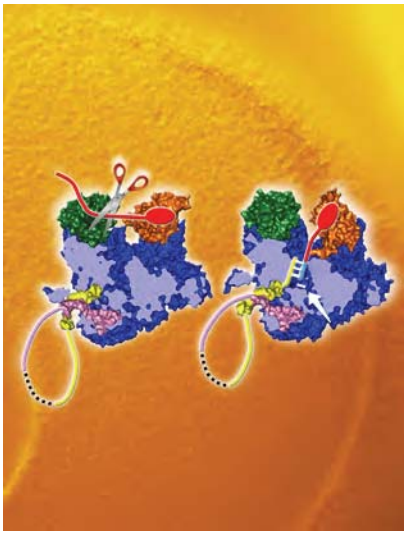
  Benossiek C. et al. (2013) The architecture of human general transcription factor TFIID core complex. Nature 493:699-702. doi: 10.1038/nature11791

- **Superhelical architecture of the myosin filament-linking protein myomesin**
  Members of the myomesin protein family are unusually long and elastic proteins that function as molecular bridges connecting filament systems in

muscle. EMBL scientists in Hamburg have unravelled the mechanism of molecular elasticity by superhelical linkers in these proteins by determining the overall architecture of myomesin by X-ray crystallography, electron microscopy, solution X-ray scattering, and atomic force microscopy.

Pinotsis N. et al. (2012) Superhelical architecture of the myosin filament-linking protein myomesin with unusual elastic properties. Plos Biol 10:e1001261. doi: 10.1371/journal.pbio.1001261

- **Complete structure of the influenza virus polymerase**
  Using X-ray crystallography, researchers in Grenoble were able, in culmination of work they had pursued for 25 years, to determine the atomic structure of the whole influenza virus polymerase. The results of this groundbreaking work allow researchers to understand how the polymerase uses host cell RNA to kick-start the production of viral messenger RNA, and could thus prove instrumental in the design of new anti-influenza drugs.



Pflug A. et al. (2014) Structure of influenza A polymerase bound to the viral RNA promoter. Nature 516:355-360. doi: 10.1038/nature14008

Reich S. et al. (2014) Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature 516:361-366. doi: 10.1038/nature14009

Figure B.1.1 The complete structure of the influenza virus polymerase allows researchers to understand how the polymerase uses host cell RNA (red) to kick-start the production of viral messenger RNA. Credit: EMBL/P.Riedinger

1.1.1.2 First protein crystal structure at 0.48 Å resolution

In 2011 scientists at EMBL Hamburg set the record in the structural resolution of a biological macromolecule obtained by X-ray crystallography. The structure of the small protein crambin was determined to 0.48 Å resolution using the PETRA II ring at DESY. A wealth of details of crambin's electronic structure were revealed by the results of this study, that used the PETRA II storage ring just before its conversion to the dedicated synchrotron-radiation source, PETRA III.

Schmidt A. et al. (2011) Crystal structure of small protein crambin at 0.48 Å resolution. Acta Crystallogr Sect F Struct Biol Cryst Commun 67:424-428. doi: 10.1107/S1744309110052607

1.1.1.3 Correlative microscopy of viral and cellular budding systems

The development and application of novel correlative microscopy techniques to study cellular events at high resolution over time was identified in the Programme 2012-2016 as a major research objective for the current Indicative Scheme period. A particular focus lay in the study of vesicle budding events, which are

involved in various cellular processes as well as being used by many viruses to invade host cells. Correlative microscopy methods have been successfully exploited at EMBL over recent years to produce high-resolution three-dimensional molecular-scale 'movies' of viral and cellular budding systems, as exemplified by the remarkable findings outlined below.

- **High-resolution retrovirus structure**
  Upon infection of a target cell, retroviruses such as HIV replicate to produce new immature viral particles, assembled from a medley of viral and cellular components. In 2012, scientists at EMBL Heidelberg used a combination of electron microscopy and computational methods to unravel how the key proteins of the viral shell assemble to form the immature virus, and how they rearrange themselves to give rise to its mature form. In subsequent studies, the same research team combined cryo-electron microscopy and tomography analysis to explore the structure of the HIV protein lattice and its rearrangements in even greater detail. The results of this work are of great significance as, for instance, they provide the basis for investigating the mechanisms of drugs known to inhibit HIV assembly and maturation.

  Bharat T.A.M. et al. (2012) Structure of the immature retroviral capsid at 8Å resolution by cryo-electron microscopy. Nature 487:385-389. doi: 10.1038/nature11169

  Bharat T.A.M. et al. (2014) Cryo-electron microscopy of tubular arrays of HIV-1 Gag resolves structures essential for immature virus assembly. Proc Natl Acad Sci 111:8233-8238. doi: 10.1073/pnas.1401455111

  Schur F.K.M. et al. (2014) The structure of the immature HIV-1 capsid in intact virus particles at 8.8 Å resolution. Nature 517:505-508. doi: 10.1038/nature13838

- **Dynamics of endocytosis by correlative light and electron microscopy**

  Endocytosis, like many dynamic cellular processes, requires precise temporal and spatial orchestration of complex protein machinery to mediate the membrane budding required for entry into cells of many essential cellular constituents. To understand how this machinery works, researchers at EMBL Heidelberg directly correlated fluorescence microscopy of key protein components with electron tomography. The resulting 3D ultrastructural movies accurately define the protein-mediated membrane shape changes that occur during endocytosis and underline the elaborate dynamics of recruitment and departure of the required proteins from the site of membrane budding.

  Kukulski W. et al. (2012) Plasma membrane reshaping during endocytosis is revealed by time-resolved electron tomography. Cell 150:508-520. doi: 10.1016/j.cell.2012.05.046

1.1.1.4 Structure and dynamics of nuclear pore complexes

The nuclear pore complex (NPC) is a highly dynamic and complex protein assembly, the largest stable protein assembly in the cell, and the organisation of its individual components has long represented a challenge for molecular biologists. The study of NPC architecture and function, identified as a key research topic in the last EMBL Programme, has come a long way in the first years of the current Indicative Scheme. By combining stochastic super-resolution

microscopy to directly resolve the ring-like structure of the NPC, with single particle averaging, to use information from thousands of pores, scientists in Heidelberg determined the average positions of fluorescent molecular labels in the NPC with unprecedented accuracy. This approach was applied systematically to a major motif of the NPC, the Nup107-160 sub-complex, to assess the overall structure of the NPC scaffold. This work shows that light microscopy can be used to investigate the molecular organisation of large protein complexes *in situ* in whole cells and is the first ever molecular resolution data obtained by light microscopy of cells. In a parallel study, an integrated approach based on electron tomography, single-particle electron microscopy, and crosslinking mass spectrometry was used to determine the structure of the Nup107 subcomplex, both in isolation and integrated into the NPC. The arrangement uncovered in this work may explain how changes in the diameter of the NPC that accommodate transport of huge cargoes are made possible.

Szymborska A. et al. (2013) Nuclear Pore Scaffold Structure Analyzed by Super-Resolution Microscopy and Particle Averaging. Science 341:655-658. doi: 10.1126/science.1240672

Bui K.H. et al. (2013) Integrated structural analysis of the human nuclear pore complex scaffold. Cell 155:1233-1243. doi: 10.1016/j.cell.2013.10.055

1.1.1.5 Systematic silencing of human genes and analysis by microscopy

In 2012 scientists from the Cell Biology and Biophysics Unit in Heidelberg, in collaboration with the Core Facilities, used an automated time-lapse microscopy-based RNA interference (RNAi) screen to systematically downregulate over 21,000 genes in cultured human cells. Their analysis revealed an unexpected variety of genes involved in cargo molecule transfer and secretion, and unveiled novel links between early secretory pathway function and key signalling pathways. The combination of quantitative RNAi and automated image analysis was subsequently used to characterise candidate genes associated with blood lipid levels, coronary artery disease, myocardial infarction and in regulating cholesterol levels in cells. In 2013 a similar approach was taken in airway epithelial cells to investigate the molecular bases of cystic fibrosis. The study uncovered a large set of genes that had not been previously linked to the disease and led to the identification of a promising new drug target. Taken together, these results show that large-scale high-content imaging-based RNAi screens provide an important resource for the integrative understanding of global cellular regulation and gene function, and can yield valuable insight into human disease.

Simpson J.C. et al. (2012) Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat Cell Biol 14(7):764-74. doi: 10.1038/ncb2510

Blattmann P. et al. (2013) RNAi-based functional profiling of loci from blood lipid genome-wide association studies identifies genes with cholesterol-regulatory function. PLoS Genet 9:e1003338. doi: 10.1371/journal.pgen.1003338

Almaça J. et al. (2013) High-Content siRNA Screen Reveals Global ENaC Regulators and Potential Cystic Fibrosis Therapy Targets. Cell 154:1390-1400. doi:10.1016/j.cell.2013.08.045

1.1.1.6 Cellular systems biology of a minimal bacterium

Mycoplasma pneumoniae, the causative agent of atypical pneumonia, carries a simple genome – consisting of only 689 genes – and yet shares many features with more complex cells in terms of molecular organisation and interactions. Hence, this human pathogen was chosen as a model organism to obtain the first-ever blueprint of a minimal cell. As reported in the Programme 2012-2016, EMBL scientists and collaborators from the EMBL-CRG Systems Biology Partnership Unit in Barcelona conducted a comprehensive and quantitative analysis of the bacterium's proteome, metabolic network, and transcriptome. This dataset has served, over the period of the current Indicative Scheme, as the basis for further interdisciplinary efforts at EMBL, aimed at deducing general principles underlying a variety of regulatory processes and networks. Selected examples of these studies are provided below.

- EMBL scientists from the Structural and Computational Biology Unit in Heidelberg used *M. pneumoniae* to systematically investigate the interplay of protein phosphorylation with other post-translational regulatory mechanisms, particularly lysine acetylation. The results of this work imply that previously unreported layers of post-transcriptional regulation, such as crosstalk between different kinds of post-translational modifications, is very likely to play an important role in defining the functional state of a cell.

  Van Noort V. et al. (2012) Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Mol Syst Biol 8:571. doi: 10.1038/msb.2012.4

- In collaboration with the EMBL-CRG Systems Biology Partnership Unit, scientists at EMBL Heidelberg reported the genome-wide identification in *M. pneumoniae* of small RNAs associated with transcription start sites (TSSs), termed tssRNAs. The evidence they present suggests that tssRNAs, found in several bacterial phyla, may represent a universal mechanism to halt bacterial transcription.

  Yus E. et al. (2012) Transcription start site associated RNAs in bacteria. Mol Syst Biol 8:585. doi: 10.1038/msb.2012.16

- EMBL scientists in Heidelberg and their collaborators in Barcelona took a systems biology approach to analyse the complex molecular events that regulate biological function and cellular responses to environmental perturbations in *M. pneumoniae*. Their study sheds light on these fundamental regulatory mechanisms by providing a detailed integrative analysis of cellular protein abundances and the dynamic interplay of mRNA and proteins under various external conditions.

  Mair T. et al. (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol 7:511. doi: 10.1038/msb.2011.38

1.1.2 From cells to organisms: dynamic organisation and imaging

1.1.2.1 Self-directed migration in zebrafish

Many research projects carried out at EMBL focus on understanding how cells aggregate and move within developing organisms to form differentiated tissues

and organs during embryogenesis. Two collaborative studies by scientists from the Cell Biology and Biophysics, Genome Biology and Structural and Computational Biology Units in Heidelberg have revealed important aspects of developmental signalling dynamics using the zebrafish lateral line primordium as a model. As the zebrafish embryo develops, cells migrate along chemoattractant gradients, leaving behind rosette-shaped clusters of cells which give rise to mechanosensory organs. In the first study, scientists applied a fluorescence timing approach to the chemokine Cxcl12a, a key guidance molecule, and combined it with intravital two-photon microscopy to show, remarkably, that the chemokine gradients driving the migration of cell collectives are self-generated by the clusters of migrating cells. The second study, which applied correlative light and electron microscopy, found that luminal signaling, particularly by fibroblast growth factor (FGF), is linked to the generation of novel forms of cell architecture that are involved in generating high local concentrations of signaling molecules and thus in control of the self-assembly of cells into the mechanosensory clusters. These findings, along with the innovative imaging methods that enabled them, pave the way for similar studies to investigate cell migration and self-organisation in other contexts, such as wound repair or cancer invasion.

Donà E. et al. (2013) Directional tissue migration through a self-generated chemokine gradient. Nature 503:285-289. doi: 10.1038/nature12635

Durdu S. et al (2014) Luminal signalling links cell communication to tissue architecture during organogenesis. Nature 515:120-124. doi: 10.1038/nature13852

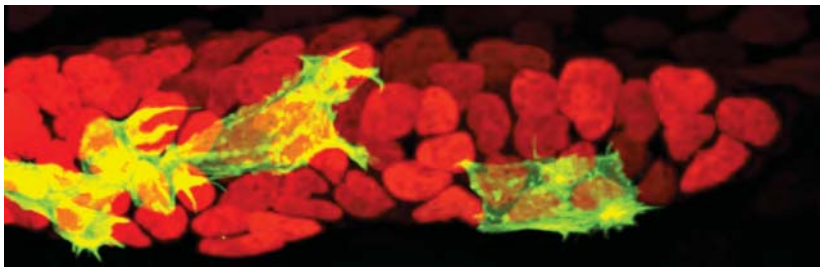


Figure B.1.2 Visualising actin dynamics (Life-Act-GFP) within the migrating primordium of the zebrafish lateral line
Credit: EMBL/C. Moehl

1.1.2.2 The control of developmental timing by biological clocks

A fundamental yet largely unanswered question in molecular biology is how inherent timing mechanisms ensure that biological processes occur in the right order and at the appropriate pace to obtain the correct functional output in order to generate correct multicellular organisation as organs and tissues differentiate. Following the research plans outlined in the last Programme, EMBL scientists have set out to unravel the basic principles of such biological clocks – particularly those controlling the temporally tightly regulated process of embryonic development. The somite segmentation clock, which controls the formation of the pre-vertebrae in early mouse development, has been the object of intense study by scientists in the Developmental Biology Unit at EMBL Heidelberg. By combining a novel *in vitro* segmentation assay with real-time imaging of gene activity, they identified a mechanism by which signalling oscillations allow mesoderm segments to remain proportional – that is, to scale – to overall embryonic size and thus ensure that embryos of different overall sizes are nevertheless perfectly organised internally The novel quantitative experimental

approach applied in this study opens new avenues for investigating how information is encoded at the level of signalling dynamics and how signalling oscillations are employed during embryonic patterning.

Lauschke V.M. et al. (2012) Scaling of embryonic patterning based on phase-gradient encoding. Nature 493:101-105. doi: 10.1038/nature11804

1.1.2.3 Regulators of body plan development

Formation of the correct body plan in the Drosophila embryo strictly relies on localised protein synthesis which in turn depends on the kinesin-dependent transport of maternal mRNAs. Researchers from the Developmental Biology Unit in Heidelberg have identified an mRNA structure which assembles when two RNA segments are spliced together in the nucleus. This structure is crucial for localisation of the *oskar* mRNA, and thus the *oskar* protein, to the posterior pole of the developing oocyte. This study implicates RNA splicing, whose disregulation can cause developmental defects and disease, in the regulation of mRNA localisation by promoting the formation of specific structures whose production can vary in different tissues.

Ghosh S. et al. (2012) Control of RNP motility and localization by a splicing-dependent structure in *oskar* mRNA. Nat Struct Mol Biol 19:441-449. doi: 10.1038/nsmb.2257

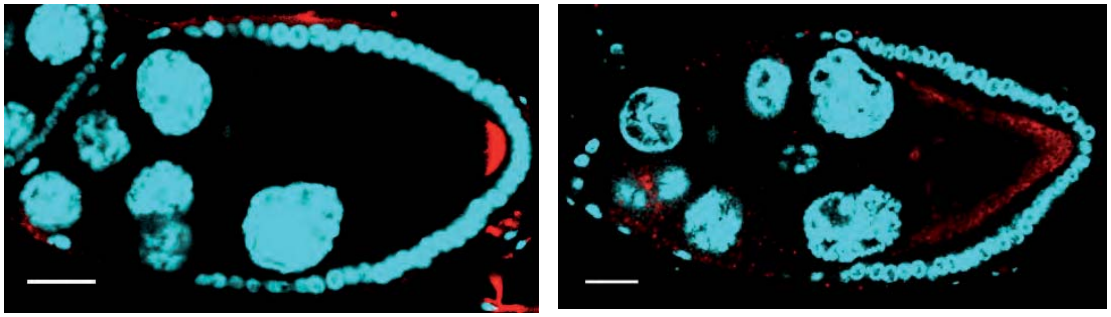
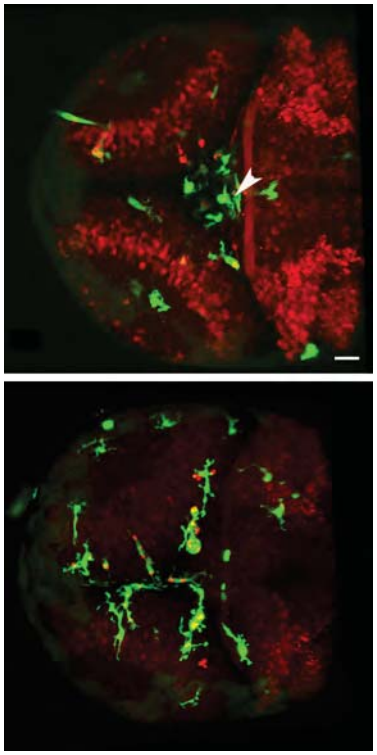


Figure B.1.3 *oskar* mRNA (red) is transported to the posterior pole in a normal Drosophila egg cell (left), but not in an oocyte with a mutated splicing tag (right).
Credit: EMBL/S. Ghosh

1.1.2.4 Understanding microglial migration towards injured neurons

Cells termed microglia are the resident phagocytes of the brain that are responsible for the clearance of injured neurons, an essential step in subsequent tissue regeneration. Scientists from the Developmental Biology Unit in Heidelberg have used the optically transparent zebrafish larval brain to uncover the relay of molecular signals that attract these cells toward sites of injury. By combining forward and reverse genetic approaches with quantitative imaging the researchers revealed a mechanism controlling microglial targeted migration to neuronal injuries that is initiated by glutamate and travels across the brain in the form of a Ca^{2+} wave. They also went on to show that microglia engulf dying neurons by extending cellular branches that form phagosomes at their tips, and identified two receptors that, through their combined action, allow microglia to remove dying neurons. This work provides important insight into processes that

lie at the heart of many neuronal degenerative disorders, such as Alzheimer's and Parkinson's diseases.



Sieger D. et al. (2012) Long-range Ca2+ waves transmit brain-damage signals to microglia. Dev Cell 22:1138-1148. doi: 10.1016/j.devcel.2012.04.012

Mazaheri F. et al. (2014) Distinct roles for BAI1 and TIM-4 in the engulfment of dying neurons by microglia. Nat Commun 5:4046. doi: 10.1038/ncomms5046

Figure B.1.4 Microglia (green) respond to brain injury by migrating towards injured neurons (top). This response is impaired when sensing of the molecular signals that attract microglia to the site of injury is impeded by knockdown of the relevant receptor (bottom).
Credit:EMBL/F. Peri

1.1.2.5 Regulation of cell and tissue shape during embryonic development

During morphogenesis, remodelling of cell shape requires the expansion or contraction of plasma membrane domains. By combining high-resolution imaging with genetics and biochemistry, EMBL scientists in Heidelberg have identified a mechanism underlying the restructuring of the apical surface during epithelial morphogenesis in Drosophila. This work resulted from collaboration between researchers from the Developmental Biology Unit, who developed a new strategy for imaging the Drosophila embryo, and their colleagues from the Structural and Computational Biology Unit, who combined light and electron microscopy to observe the membrane remodelling events in detail. These findings provide mechanistic insight into the control of cell shapes during embryonic development and could help understand how different cell types take on the specific shape required to perform different tasks, as well as how cell shape becomes abnormal during cancer progression.

Fabrowski P. et al. (2013) Tubular endocytosis drives remodelling of the apical surface during epithelial morphogenesis in *Drosophila*. Nat Commun 4:2244. doi: 10.1038/ncomms3244

1.1.2.6 Mechanisms of early mouse development and lineage segregation

A combination of quantitative live-imaging microscopy and single-cell transcriptomics has allowed scientists at EMBL Heidelberg to fully characterise the transition from meiosis to mitosis and the molecular profile of lineage segregation during early mouse development. The surprising finding was that the transition from meiosis to mitosis progresses gradually throughout the

preimplantation stage in the mouse embryo. The researchers from the Developmental Biology Unit, in collaboration with colleagues from the Genome Biology and Structural and Computational Biology Units, went on to investigate the genetic cues that determine cell fate – particularly those that distinguish the embryo from extra-embryonic tissue. By profiling the genetic activity of individual cells at different stages of development, they identified the exact time point at which random variations in gene activity give rise to two distinct cell populations, and characterised their genetic signatures. As a result of this work a new model was postulated that proposes a possible role of cell-to-cell expression heterogeneity in lineage segregation.

Courtois A. et al. (2012) The transition from meiotic to mitotic spindle assembly is gradual during early mammalian development. J Cell Biol. 198:357-370. doi: 10.1083/jcb.201202135

Ohnishi Y. et al. (2014) Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nat Cell Biol 16:27-37. doi: 10.1038/ncb2881

1.2 Unravelling biological complexity

A major challenge in the study of biological systems is the need to account for the combined action of multiple molecular or cellular components – ultimately, for the complexity of living beings – in the effort of linking genotype to phenotype. Over the current Programme EMBL scientists have made considerable progress in this direction by analysing and integrating systematic global measurements of diverse biological molecules (DNA, RNA, proteins, etc.) with bioinformatics and computational biology approaches, and uncovered fundamental aspects of genome organisation and transcriptional regulation.

1.2.1 Sequencing the HeLa genome

HeLa cells are the most widely used model cell line for studying human cellular and molecular biology, and have long served as a standard human model for understanding many fundamental biological processes. In a collaborative effort involving three different research groups, scientists from the Genome Biology Unit have successfully sequenced the DNA and RNA of a HeLa cell line to analyse its mutational portfolio and gene expression profile. Their work provides a high-resolution genomic reference that reveals striking differences between the HeLa genome and that of normal human cells. The scientists' analysis of the HeLa genome revealed a remarkably high level of aneuploidy, the loss of healthy copies of genes, and numerous large structural variants (genome rearrangements like deletions, duplications and translocations). The analysis of the HeLa transcriptomic profile revealed that several pathways, including cell cycle and DNA repair, exhibit significantly different expression patterns from those in normal human tissues. These results provide the first detailed account of genomic variants in the HeLa genome, yielding insight into their impact on gene expression and cellular function, as well as their origins. This resource could enhance the quality of basic research in human biology, as the knowledge of the genetic landscape of HeLa cells can inform the design and interpretation of future studies using these cells thereby strengthening the biological conclusions that can be drawn from them.

Landry J.J. et al. (2013) The Genomic and Transcriptomic Landscape of a HeLa Cell Line. G3 (Bethesda) 3:1213-1224. doi: 10.1534/g3.113.005777

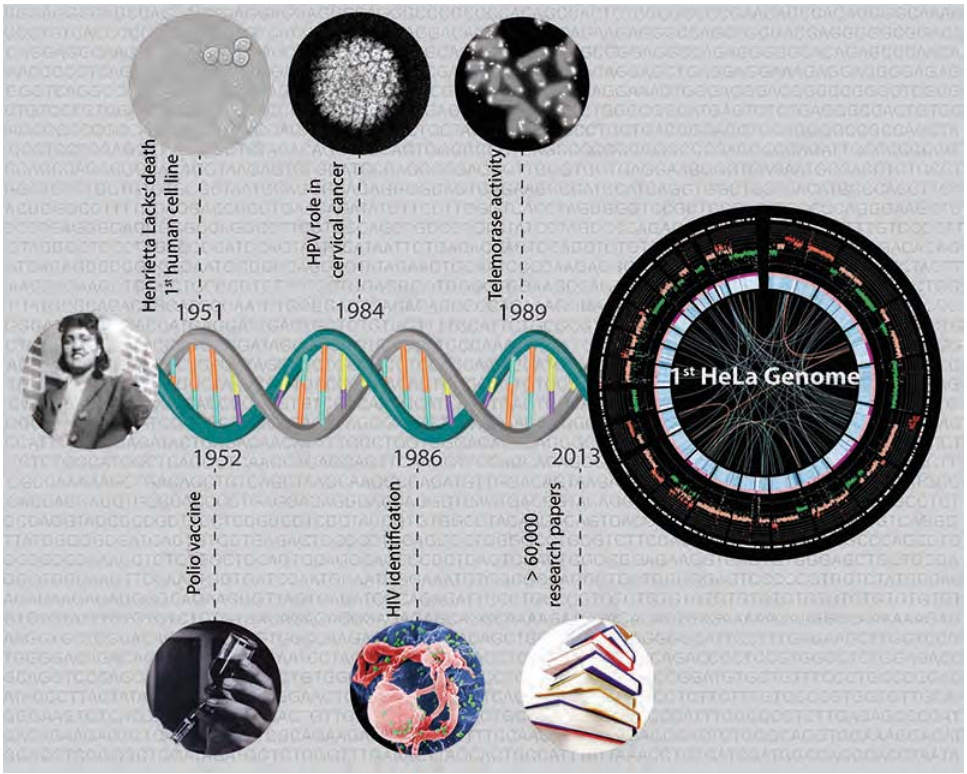


Figure B.1.5 Since 1951, HeLa cells have been a fundamental resource for scientists; their use has produced significant impact in many areas of biomedical research.
Credit: EMBL/J. Landry

1.2.2 Understanding transcriptome regulation

EMBL scientists in Heidelberg have revealed an astonishing and previously hidden level of variation in the structures of individual transcript molecules generated from a single genomic sequence. These findings were enabled by the development of a new high-throughput sequencing based technology that allows to precisely map mRNA boundaries, and thus to quantify individual transcript isoforms in yeast. This work represents the first-ever effort to measure transcriptional heterogeneity by whole genome isoform profiling, and sheds new light on transcriptome regulation and the importance of mRNA boundaries in determining the functional potential of genes.

Pelechano V. et al. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 497:127-131. doi: 10.1038/nature12121

1.2.3 Variability and expression at the single-cell level

Single-cell transcriptomic analyses are affected by unavoidable technical noise owing to the low amount of starting material. Through a joint effort across EMBL Units, researchers in Heidelberg and Hinxton developed a quantitative statistical method to distinguish true biological variability from these high levels of technical noise. Their approach, which determines the statistical significance of observed cell-to-cell variations in expression strength separately for each gene, was illustrated and validated using independent datasets from single-cell RNA sequencing experiments performed on plant and mouse cell populations.

Brennecke P. et al (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10:1093-1095. doi: 10.1038/nmeth.2645

1.2.4 Mapping the mRNA interactome

In 2012 EMBL scientists in Heidelberg used a systematic and unbiased approach to define the mRNA interactome – the whole set of protein-mRNA interactions – of proliferating HeLa cells. Their analysis revealed over 300 RNA-binding proteins (RBPs) that had previously not been thought to bind RNA, and shed light on diverse aspects of RNA biology including RBPs in disease, RNA-binding kinases and RNA-binding architectures. The researchers' list of newly discovered RBPs included a large number of metabolic enzymes that, if functionally relevant, could broadly connect intermediary metabolism with RNA biology and post-transcriptional gene regulation. This work offers an informative snapshot of RNA biology, and describes a novel approach that can be broadly applied to study mRNA interactome composition and dynamics in different biological settings.

Castello A. et al. (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149:1393-1406. doi: 10.1016/j.cell.2012.04.031

1.2.5 Advancing the understanding of enhancers

Over the past years EMBL scientists have engaged in elucidating enhancer function and its role in the regulation of gene expression, thereby making important contributions to the field. In 2012, researchers from the Genome Biology Unit in Heidelberg found that enhancer activity is precisely regulated by chromatin state during Drosophila embryogenesis. Using a novel approach, they obtained and used cell type specific information on chromatin modifications and RNA Polymerase II occupancy to make accurate predictions of dynamic enhancer activity during embryonic development. In collaboration with scientists at EMBL-EBI they also proposed a new model for enhancer function, showing that transcription factors can act in a highly cooperative manner at enhancers, with minimal DNA sequence requirements. Recently, new properties governing enhancer-promoter interactions and their dynamics during fruit fly embryogenesis have been uncovered. The finding that contact between the enhancer and its target promoter is established long before enhancer activation fundamentally changes the current view of transcription initiation. Furthermore, the observation that the prevalence of long-distance action and the degree of complexity of enhancer interactions are comparable to those seen in vertebrates carries important evolutionary implications.

Bonn S. et al. (2012) Tissue specific analysis of chromatin state reveals temporal signatures of enhancer activity during embryonic development. Nat Genet 44:148-156. doi: 10.1038/ng.1064

Junion G. et al. (2012) A transcription factor collective defines cardiac cell fate and reflects the developmental history of this cell lineage. Cell 148:473-486. doi: 10.1016/j.cell.2012.01.030

Ghavi-Helm Y. et al. (2014) Enhancer loops appear stable during development and are associated with paused polymerase. Nature 512:96-100. doi: 10.1038/nature13417

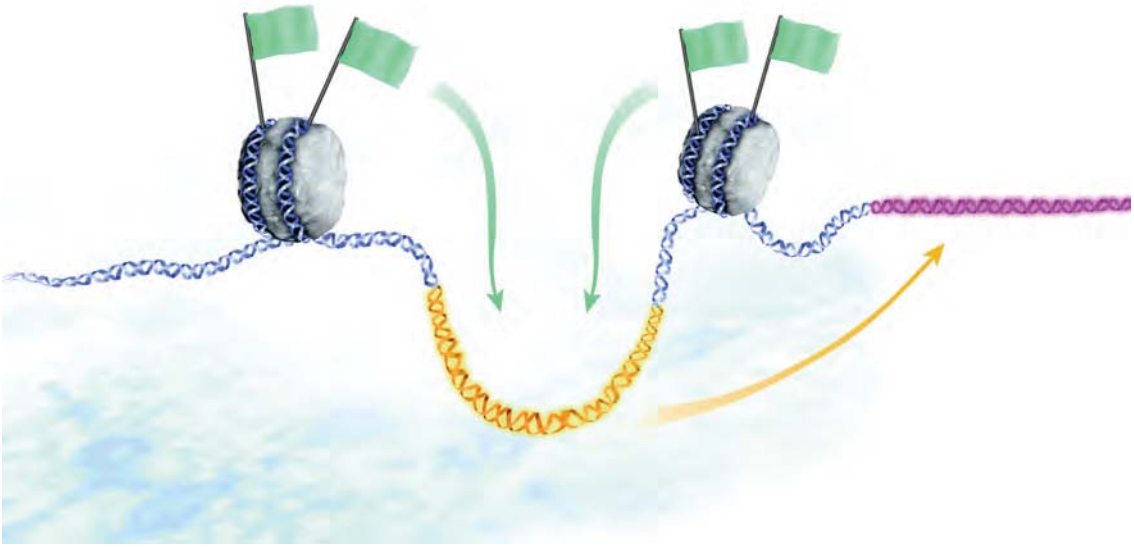


Figure B.1.6 Chromatin modifications (green flags) are chemical tags that activate enhancers (yellow), which in turn act as remote controls turning genes (red) on or off.
Credit: EMBL/P. Riedinger

1.3 Exploring biological variation

Biological variation was incorporated as a new research theme in the 2012-2016 Programme, following enormous advances in DNA sequencing technology which transformed the degree to which the genetic variation between species, and between individuals of the same species, and thus the genetic basis of healthy and diseased states, can be analysed. In the ongoing Indicative Scheme period a number of projects have successfully addressed how genetic variation influences basic biological function and organisation and its effects on complexity at various levels.

1.3.1 Evolution: inter-species variation

1.3.1.1 Evolution of the chordate notochord

Work by researchers in the Developmental Biology Unit at EMBL Heidelberg has provided new insights into the evolution of the notochord, the first vertebrate skeleton. The scientists identified the genetic signature of notochord cells, and found a group of cells showing similar gene activity in the developing larva of a marine worm. State-of-the-art light sheet microscopy, developed at EMBL, revealed that such cells give rise to a longitudinal muscle, whose morphology and location is similar to the notochord. The molecular relatedness strongly suggests that the vertebrate notochord has evolved from this ancestral muscular structure and thus has a much older evolutionary origin than previously assumed.

Lauri, A. et al. (2014) Development of the annelid axochord: Insights into notochord evolution. Science 345:1365-8. doi: 10.1126/science.1253396

1.3.1.2 Comparative analysis of model organisms

In the first three years of the current Programme, scientists at EMBL-EBI have participated in several international efforts aimed at investigating the evolutionary origin of given animal species, as well as how they genetically relate to higher vertebrates. One example is provided by a study that generated and analysed the draft genomes of the soft-shell turtle and the green sea-turtle to gain insight into the development and evolution of the unique turtle-specific body plan. In parallel, a similar study was conducted using zebrafish, a popular model organism for the study of vertebrate gene function and human genetic disease, to understand the extent to which zebrafish genes and gene structures are related to orthologous human genes. In the context of these studies, that heavily relied on the Ensembl data resource, which is jointly operated by EMBL-EBI and the Wellcome Trust Sanger Institute, genomic organisation and annotation, as well as comparative gene expression analysis and identification of orthology relationships, were obtained.

Wang Z. et al. (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet 45:701-706. doi: 10.1038/ng.2615

Howe K. et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498-503. doi: 10.1038/nature12111

1.3.1.3 Evolution of mouse gene expression

In recent years, researchers at EMBL-EBI have led studies aiming to investigate the contribution of different regulatory mechanisms to divergence in gene expression. For example, they took first steps to elucidate the evolution of transcription factor binding, and the underlying mechanisms, in mammals by characterising the binding profiles of three tissue-specific transcription factors in livers from six inbred rodents. The study revealed large, qualitative differences in the genomic regions bound between these closely related mammals, which contrasted with the smaller, quantitative binding differences among Drosophila species. This study points to the different population genetics of flies and mammals as a driver of profound differences in transcription factor binding stability between species, and could have implications for understanding differences in human gene regulation and in disease susceptibility between individuals.

Stefflova K. et al. (2013) Cooperativity and rapid evolution of co-bound transcription factors in closely related mammals. Cell 154:530–540. doi: 10.1016/j.cell.2013.07.007

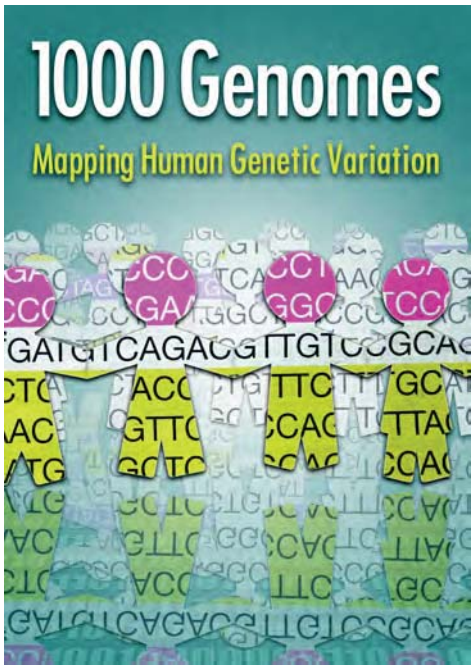
1.3.2 Genetic variation: intra-species variation

1.3.2.1 Map of human variation

In 2012 the 1000 Genomes Project achieved a major milestone towards the understanding of genetic contribution to disease by characterizing the geographic and functional spectrum of human variation at different scales – from single nucleotide polymorphisms to structural variants. Using a combination of whole-

genome and exome sequencing, the genomes of 1,092 healthy individuals from 14 populations were described. This vast dataset is an invaluable resource for studies ranging from biomedical research to human evolution, and has been made freely and publicly accessible to researchers worldwide. The great success of the 1000 Genomes Consortium is largely shared by scientists at EMBL-EBI and in Heidelberg that took lead roles in different areas of the project (i.e. the Data Coordination Centre and the Structural Variation Group). Recently, samples of 462 individuals from the 1000 Genomes Project were subject to deep sequencing and analysis of messenger RNA and microRNA, offering the largest-ever dataset linking transcriptome variation to the landscape of functional human genetic variation. The data produced in this study is available though the ArrayExpress functional genomics archive, led by scientists at EMBL-EBI.

The 1000 Genomes Project Consortium et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. doi: 10.1038/nature11632



Lappalainen T. et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506-511. doi: 10.1038/nature12531

't Hoen P.A. et al. (2013) Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol 31:1015-1022. doi: 10.1038/nbt.2702

Figure B.1.7 The 1000 Genomes Project is the first project to sequence the genomes of a large number of individuals, to provide a comprehensive resource on human genetic variation.
Credit: EMBL-EBI/S. Phillips

1.3.2.2 Cancer Genomics

Scientists in the Genome Biology Unit at EMBL Heidelberg use genetic variation as a starting point to unravel disease mechanisms, particularly those giving rise to cancer. In this context, our researchers aim at characterizing the extent, origin and consequences of these DNA variations, with a particular focus on genomic structural variants, and their links to different cancer types. This research, which combines experimental and computational techniques, has yielded several important findings over the past three years, as outlined below.

- EMBL scientists found that the development of medulloblastoma, the most common malignant brain tumor in children, frequently involves a process known as chromothripsis, where localised chromosomal shattering and reassemble in a one-off, massive DNA rearrangement event. Furthermore, an association between a hereditary mutation in the tumor suppressor TP53 molecule and chromothripsis in medulloblastoma was uncovered, that has

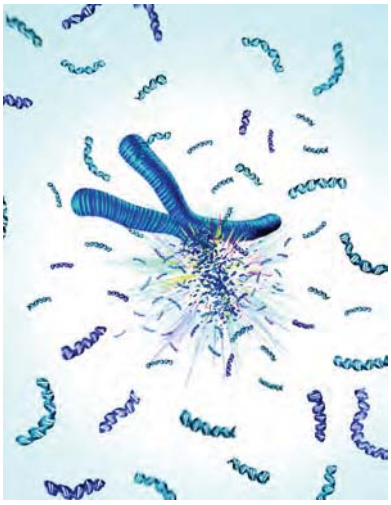
significant implications for diagnosis and treatment of particularly aggressive cancer subtypes.

Rausch T. et al. (2012) Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations in Cancer. Cell 148:59-71. doi: 10.1016/j.cell.2011.12.013

- Progress has been made in understanding the etiology of early-onset prostate cancer, the initiation of which was found to be largely driven by androgen-mediated somatic structural variants. These findings represent the first description of an age-related DNA alteration process in a common cancer, and could have clinical consequences for this common cancer.

Weischenfeldt J. et al. (2013) Integrative genomic analyses reveal androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell 23:158-170 doi: 10.1016/j.ccr.2013.01.002

- The proteins GFI1 and GFI1B were identified as two novel oncogenic drivers of Group 3 medulloblastoma, one of the most common medulloblastoma molecular subgroups. It was shown that oncogenic activity is triggered by a series of recurrent, highly disparate genomic structural variants, which place these genes in proximity to highly active enhancers, by a mechanism termed ‘enhancer hijacking’. Importantly, these results were supported by evidence from the first mouse models of human medulloblastoma.



Northcott P.A. et al. (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511:428-34. doi: 10.1038/nature13379

Figure B.1.8 A massive chromosomal shattering and gene rearrangement event, termed chromothripsis, is linked to different cancer types, including human medulloblastoma. Credit: EMBL/P. Riedinger

1.3.3 Disease models and mechanisms

1.3.3.1 Mouse models of human diseases

EMBL has a traditional strength in the generation of mouse models as a means to decipher mechanisms underlying biological and disease processes. Our long-standing expertise, along with an expanding repertoire of genetic technology, has allowed scientists at EMBL to accurately model a number of human diseases over the recent years. Selected examples of findings enabled by newly generated mouse models are listed below.

- **Anxiety**
  By combining the use of transgenic mice with behavioural tests, researchers in Monterotondo showed that functionally independent populations of neurons process different types of fear, such as fear of pain, of predators, and of aggressive conspecifics. These findings could have implications for addressing pathological fear including phobias and panic attacks in humans.

  Silva B.A. et al. (2013) Independent hypothalamic circuits for social and predator fear. Nat Neurosci 16:17311733. doi: 10.1038/nn.3573

- **Male infertility**
  Researchers at EMBL Grenoble found that disruption of the catalytic activity of a small RNA-guided endonuclease by a single point mutation caused male infertility and increased transposon activity, thus showing that the Piwi-piRNA pathway of small RNAs is critical for transposon control and fertility. Notably, mutations in this pathway that affect fertility in humans have also been described.

  Reuter M. et al. (2011) Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature 480:264-267. doi: 10.1038/nature10672

- **Autism**
  Microglia are phagocytic cells that invade the brain during early development, and whose functions include actively engulfing and eliminating synapses. EMBL scientists in Monterotondo used genetically engineered mice to show that microglia that fail to prune synapses efficiently during development cause reduced functional brain connectivity and behaviours commonly linked to autism. These findings open the possibility that disruptions in microglia-mediated synaptic pruning could contribute to neurodevelopmental and neuropsychiatric disorders.

  Zhan Y. et al. (2014) Deficient neuron-microglia signaling results in impaired functional brain connectivity and social behavior. Nat Neurosci 17:400-406. doi: 10.1038/nn.3641

- **Cleft lip/palate**
  EMBL researchers in Heidelberg genetically engineered a large region of the mouse genome, orthologous to a DNA interval associated with increased risk of cleft lip and/or cleft palate in humans. In collaboration with researchers in EMBL-EBI, they found that this DNA interval contains remote enhancers controlling expression of the regulatory *Myc* gene in the developing upper lip, and that deletions in this region cause mild changes in face morphology and greater susceptibility to facial clefts. This work provides insight into the genetic pathways accounting for the influence of this region in human facial development.

  Uslu V.V. et al. (2014) Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. Nat Genet 46:753-758. doi: 10.1038/ng.2971

1.4 The need for and use of Bioinformatics and Computational Biology

Computational approaches guide experimental research by generating new hypotheses and models, and are an indispensible requirement for meeting today's challenges in the molecular life sciences. The scope of bioinformatics and computational biology in EMBL's research landscape is broad and, accordingly, computational research is interwoven with many of the research projects illustrated in previous sections. Here we present a few examples of mainly computationally driven research to give a flavour of the variety of themes and applications encompassed by *in silico* analysis and modelling.

1.4.1 Map of human metabolism

An understanding of metabolism is fundamental to comprehending the phenotypic behaviour of all living organisms, including humans, where metabolism is integral to health and is involved in many aspects of human disease. High quality, genome-scale 'metabolic reconstructions', representing the entire network of metabolic reactions known for a given organism, are at the heart of systems biology analyses and instrumental to future biomedical studies. As part of a large international consortium, scientists at EMBL-EBI have presented Recon2, a community-driven, consensus metabolic reconstruction, which is the most comprehensive computational model of human metabolism to date. Recon2 contains verified information on thousands of metabolites and reactions. The metabolic model has been made available in the BioModels Database, hosted at EMBL-EBI, and contains thousands of cross-references to other databases such as ChEBI and UniProt. Enabling this kind of intra-network view of genes and their products – enzymes, hormones, nutrients, etc. – could give researchers a close-up view of a wide range of biological phenomena, including how diseases develop and which drugs may prevent or cure them.

Thiele I. et al. (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31:419-425. doi: 10.1038/nbt.2488

1.4.2 Using DNA to store digital information

Researchers at EMBL-EBI have tested a reliable and scalable method for using synthetic DNA to store data. By translating binary digital files into non-repeating strings of A,T,G and C in a four-base code that they devised and applying an error-correction algorithm the scientists encoded several large computer files, and subsequently manufactured them as DNA using synthetic biology techniques. Remarkably, the original files were reconstructed with 100% accuracy upon sequencing of the DNA sample. Such a DNA-based storage scheme could be scaled far beyond current global information volumes and offer a realistic technology for large-scale, long-term and infrequently accessed digital archiving in the future.

Goldman N. et al. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 94:77-80. doi: 10.1038/nature11875

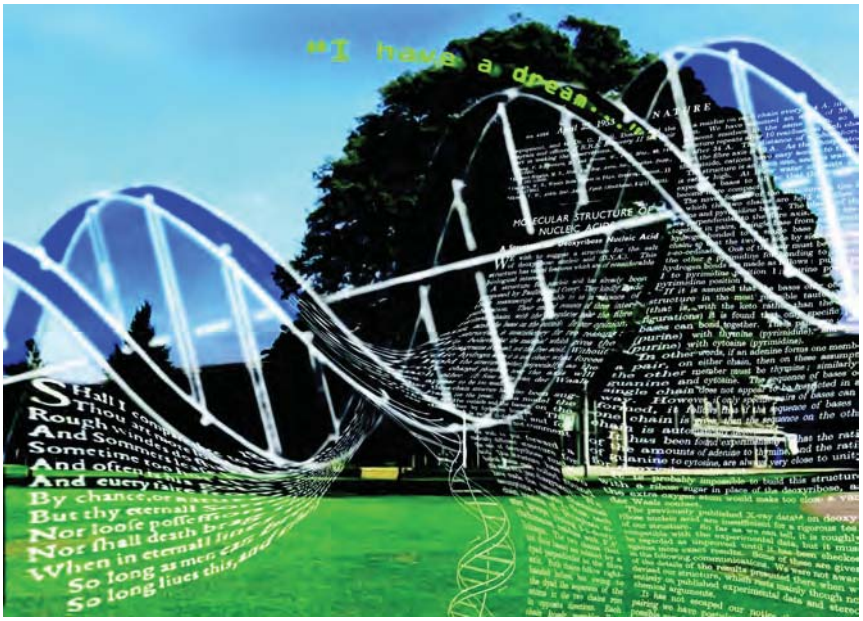


Figure B.1.9 Practical, high-capacity storage of digital information in synthesised DNA is becoming a reality. Credit: EMBL

1.4.3 Metagenomics

The growing field of metagenomics responds to the need of incorporating information on environmental context in order to gain a complete understanding of the human body as a biological system. The power and the huge potential of this approach were already highlighted in the context of the research themes for the Programme 2012-2016. In the current Programme period EMBL scientists have contributed significantly to this area of research, particularly with studies focusing on the microbial communities in the gut. Innovative computational tools were developed to analyse the gut metagenome, that is the collection of genomes of all the microbes in the human intestinal tract, at high resolution. It was found that each human carries a unique set of bacterial strains and mutations, which is stable over time. By analysing faecal samples, the researchers identified three principle community compositions, termed enterotypes, and uncovered microbial genetic markers that correlate with traits like age and weight. Furthermore, the first disease association was discovered: variations in the gut microbiome are associated with increased risk of obesity-related conditions. These findings could lead to the development of new approaches in the disease diagnosis and, in the longer term, may open new therapeutic avenues.

Schloissnig S. et al. (2013) Genomic variation landscape of the human gut microbiome. Nature 493:45-50. doi: 10.1038/nature11711

Arumugam M. et al. (2011) Enterotypes of the human gut microbiome. Nature 473:174-180. doi: 10.1038/nature09944

Le Chatelier E. et al. (2013) Richness of human gut microbiome correlates with metabolic markers. Nature 500:541-546. doi: 10.1038/nature1250

1.4.4 Protection against mutations

A central tenet in evolutionary theory is that mutations occur randomly with respect to their value to an organism, and that selection then governs whether they are fixed in a population. This principle has recently been challenged by the work of researchers at EMBL-EBI, who used a novel approach combining phylogenetic and population genetic techniques to study how mutation rates vary between different sites within the genome of the bacterium *Escherichia coli*. Our researchers observed that mutations occurred non-randomly – at a lower rate in highly expressed genes and in those undergoing stronger purifying selection (i.e. those showing signs of having been subject to selective pressure) – thus suggesting that the mutation rate has been optimised to reduce the risk of deleterious mutations. Understanding the extent to which this takes place and the mechanisms that modulate genetic mutation rate in different organisms is important for our understanding of the evolution of genomes as well as the bases of many human diseases.

Martincorena I. et al. (2012) Evidence of non-random mutation rates suggests an evolutionary risk-management strategy. Nature 485:95-98. doi: 10.1038/nature10995

Figure B.1.10 A unique set of bacterial strains is harboured within an individual's gut.
Credit: EMBL

1.4.5 ENCODE

Researchers at EMBL-EBI led the analysis of the data produced in the context of the ENCyclopedia Of DNA Elements (ENCODE) project, carried out by an international consortium as a contribution towards generating a comprehensive catalogue of all components of the human genome that are crucial for biological function. In 2012 ENCODE presented a systematic and detailed genome map including regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled researchers to annotate 80% of the genome as being involved in one or more of these function-associated processes, many sites lying outside of the well-studied protein-coding regions. The analysis uncovered many candidate regulatory elements and identified elements corresponding to sequence variants linked to human disease. In a joint effort between the Genome Biology Unit in Heidelberg and EMBL-EBI in Hinxton, EMBL researchers combined the transcription factor binding maps generated by ENCODE with other sources of genomic variation data for Drosophila and humans to specifically investigate transcription factor binding site variability. Overall, the ENCODE project provides new insights into the

organization and regulation of our genes and genome, and represents an extensive resource of functional annotations for biomedical research.

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

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

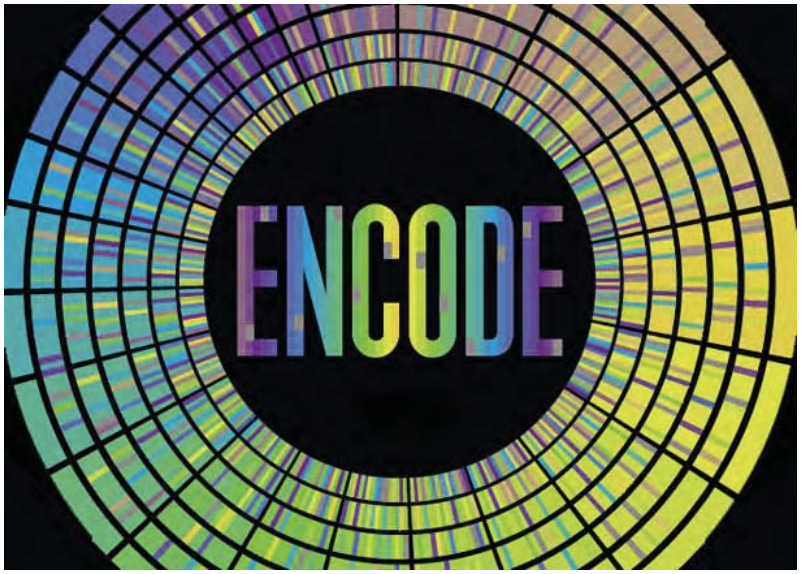


Figure B.1.11 The ENCODE project aims at generating a comprehensive catalogue of all components of the human genome and providing insight into their biological functions.
Credit: Nature journal

2. Research themes 2017–2021

2.1 Analysing biological data: from molecules to organisms and beyond

Biological research is being increasingly dominated by technologies that generate Big data (as, for example, illustrated by Figure C.1.3 in Section C.1). Next-generation sequencing is already providing a flood of data, and with whole-genome programmes emerging around the world, this trend will only increase (Sections B.2.2 & B.2.3). In addition, high-throughput imaging technologies are developing rapidly (Sections B.2.4 & E.1.1.2) and soon image data will compete with genome data in quantity. The output from other areas like proteomics and metabolomics is also rapidly increasing in both quantity and quality. For the first time in the history of biology the challenge of data gathering is matched by the challenge of data interpretation.

The data challenge is multifaceted. Biological data are complex, heterogeneous, often interdependent and biased and contain significant technical noise due to new technologies. The processes they describe result from molecular interactions and networks in space and time, with multiple levels of control. The data cover multiple different organisms, from small viruses and bacteria to humans and even entire ecosystems approaching a planetary scale (Box B.2.1). Data quantities are overwhelming: for example, the increasingly affordability of sequencing an individual's genome coupled to the medical use of this information will see us move from having a single reference human genome ultimately towards having 7 billion genomes, one for every human being on the planet. This introduces new research areas and applications, including studying the effects of individual mutations or monitoring the impact of the environment. As with all complex systems, it is the emergent properties that define how the organism as a whole operates, given certain environmental conditions. In addition to the more complex biological problems that come into reach with new data types and quantities, the sheer amount of data – rapidly approaching and soon perhaps even exceeding the data magnitudes seen in high-energy physics and astronomy – imposes a multitude of challenges (Section C.1) that have to be addressed if we are to use the data collected to efficiently further our knowledge.

Another major foreseeable shift is the number of potential 'near-market' applications in the health, agricultural and biotechnological sectors with expected shorter translational phases. The increase in quantitative molecular data and therefore rational understanding of living systems is bringing basic research closer to the commercial interface and this again brings opportunities and challenges. The opportunities include the introduction of genomic medicine, the potential to develop improved crops with better nutritional value and stress resistance, and the potential for green chemistry in industry. The challenges are to harness the data, to abstract the knowledge and to translate it for everyone's benefit. For this purpose, EMBL strongly encourages the open sharing of data to optimise translation, with commercialisation downstream. Thanks to new internet technologies, the public will have much more direct involvement in science, from selective 'crowd-funding' to information provision via social networks or 'citizen science', participation in distributed digital and educational experiments.

2.1.1 Computational biology and bioinformatics at EMBL

Making sense of Big data in biology requires interdisciplinary scientists with expertise not only in data handling, analytics and interpretation, but also in specific biological fields. EMBL has a strong tradition in computational biology and bioinformatics and in recent years, this area has become increasingly important across the board in biological research. The importance of computational approaches and expertise in everything done at EMBL is the reason we have chosen this chapter to begin the description of our research activity in detail. Today, almost every research group in EMBL is involved in some aspect of computational biology, whether as users or developers. Clearly, EMBL-EBI has the greatest concentration of bioinformaticians but already more than 40% of the scientists in EMBL Heidelberg spend more than 50% of their time using computers for scientific work. Regardless of whether all these people are classified as bioinformaticians, modern life science research relies heavily on computational biologists involved in data integration and data interpretation. In addition, it is becoming increasingly important that computational and analytical considerations are integrated into experimental research early on to inform experimental design and choices. This is particularly critical for those areas that are not yet benefitting from high-throughput technologies and where data is still comparatively sparse. For this reason, while some EMBL research groups focus exclusively on computational biology, most are increasingly involved with active data gathering and interpretation. Similarly, the majority of experimental groups arm themselves with bioinformatics expertise to be able to cope with the data they generate. This brings infrastructure and training challenges and we have therefore established several computational biology Centres (Section B.3.1) to provide expertise in identified subfields and the Bio-IT Project (Section B.3.2) to coordinate the various biocomputing activities in the experimental labs. In this way, EMBL, which was at the forefront of the development of bioinformatics as a discipline, has retained a leading position in computational biology, which we believe is instrumental in EMBL's success in all its research areas.

The increasing quantification of biology enlarges the complexity of biological systems that can be studied and makes them accessible to computational analysis. Bioinformatics already integrates a large range of data types at all spatial scales in biology, ranging from modelling individual molecules or complexes, such as the nuclear pore complex or the endocytosis machinery (Boxes B2.6 and B2.7), to modelling the properties of whole organisms (Section B.2.4.3) and communities (Section 2.4.5). Thus, computational biology functions as a bridge between historically disparate biological disciplines such as molecular, structural, cellular, developmental and organismal biology and also reaches out to ecology and research areas beyond biology.

During the current Indicative Scheme, EMBL has made major contributions in many of these areas by developing computational methodologies that enable the integration and interpretation of many different types of data to enhance our biological understanding. Broadly, our work follows the overall scheme of starting with data capture, following the genetic information flow (genomes to transcriptomes to proteomes, etc.) and from there across spatial scales from molecules to organisms, with bioinformatics being required throughout. We have developed methods for genome assembly and annotation and have proposed a

practical scheme to use DNA as a data-storage medium (Section B.1.4.2). We continue to study and model individual molecules and their properties, including the effects of post-translational modifications and the classification of enzyme mechanisms. Our interpretation of transcriptome data has revealed great evolutionary divergence in the regulation of gene expression, even between relatively closely related organisms and, by careful statistical analysis, helped to differentiate normal and cancer cells and even identify possible carcinogens. We have discovered how to establish pristine stem cells by re-programming human induced pluripotent stem cells (iPSCs), which can then be induced to differentiate *in vitro* into any cell type. With the advent of single-cell technology, new methods to decipher noisy data have enabled novel discoveries, for example in the homeostatic control of T-cell differentiation by the secretion of a steroid molecule (Box B.2.2). New computational methods to deconvolute confounding factors and to enable correlated image and single-cell data hold great promise for bridging the scales between molecular, cellular and whole organism data, as exemplified in whole brain images. Increasingly, we are involved in the analysis of medical genotypes, such as pan-cancer genomes (Box B.2.4), and phenotypes, such as changes in heart structure or in studying the effects of drugs on normal and cancer cell growth. Lastly, the analysis of metagenomic data has offered glimpses of how the human microbiome is organised and associated with diseases (Box B.2.12) and has provided examples of how the interactions between organisms and the environment affects life on earth (Box B.2.1).

2.1.2 Computational biology and bioinformatics at EMBL, 2017–2021

We envisage that during the next Indicative Scheme, computational biology will continue to develop rapidly at EMBL, integrating and abstracting knowledge from public and in-house data (Box B.2.1). We envisage a time when we will understand life and its underlying complex processes at the molecular level and be able to model them *in silico,* predicting the outcomes of genetic variation, infectious diseases and therapeutic agents. Although there is still a long way to go towards this goal, major advances are expected in the next five years in crossing from molecular to cellular scales, by combining molecular and imaging data (Section B.2.4.1) and understanding how genetic and environmental variations can lead to disease. Below we highlight the key areas in which we expect computational biology and bioinformatics at EMBL to play a major role, often coupled with the expected emergence of Big biological data.

Genes, genomes and their variation

As technologies for improved genome sequencing develop, we will work on algorithms to interpret long-read sequence data, which is predicted to become the new standard technology over the coming years. Using our current expertise in phylogenetics, we will be involved in the analysis and interpretation of pathogen genomes as a basis for clinical decision making. Furthermore, medical data increasingly presents challenges in linking genomic information and high-dimensional clinical phenotypes, (e.g. the structure of the heart) and we will develop computational methods to explore causality and predict outcomes.

Regulation of gene and protein expression and their variation

Computational biology at EMBL will benefit from single-cell genomics efforts in Hinxton and Heidelberg, and will explore how to incorporate spatial and temporal information to understand phenotype by integrating population-level variation and single-cell variability (Section B.2.2.1). These projects will generate huge amounts of sequence data that will require careful analysis and interpretation. We see many applications in human iPSCs, evolution, development (e.g. the roles and regulation of mRNAs; lncRNAs, proteins and chromatin) and in understanding cancer. We also anticipate further exploration, using systems approaches, of transcription factor-binding sites using CrispR technology to determine evolutionarily conserved and species-specific regulatory wiring and disease-causing variants. Beyond method development for the analysis of genomic, epigenomic and transcriptional regulation as well as their variations, we will also incorporate data on translational control (e.g. via ribosome profiling) for furthering our understanding of all the regulatory contributions to protein expression. Variation will be analysed between cells, cell types, organs and individuals, to differentiate between stochastic processes and regulatory constraints.

Proteins, proteomes and their variation

New proteomic technologies, including novel computational methods to increase both the sensitivity and interpretability of the data, will be used to chart the architecture of different cell types and states. Developing methods to combine these data with new structural data will help us to understand the organisation of proteins into complexes, as described in Section B.2.4.1, in a cell type- and cell state-specific manner. Based on this ‘proteotyping’ of cells and its variation between individuals, we will further improve our functional understanding of a cell’s molecular machinery as well as its evolution using bioinformatics approaches. We will perform large-scale analyses of the various post-translational modifications (PTMs) that considerably influence the functional potential of a protein. Only the entirety of expressed proteins with their various PTMs, resolved at spatiotemporal resolution, and their interactions with other small molecules such as metabolites and macromolecules including other proteins, lipids or RNAs, will reveal the organisational principles of a cell state. As a step towards this goal, we will interpret the effects of proteome variation among cell types, cell states and individuals, considering stability, interactions, function, PTMs, localisation and abundance, and develop models to explore intermediate phenotypes of interest and their links to diseases.

Beyond proteomes: chemical biology

A particular focus of the coming years will be the large-scale spatiotemporal study of metabolites and their integration with proteomes, mostly via enzymes or other proteins that interact with small molecules (Section B.2.2.4). This will also involve the perturbation of cells by exogenous factors such as drugs. For example, we will study the effect of target variation on drug response or the pharmacogenetics of pro-drug activation. We will develop new, more sensitive, methods for mass spectrometry and nuclear magnetic resonance (NMR)-based metabolomics as well as imaging mass spectrometry for providing spatial context. Metabolic modelling will be instrumental for deducing the fluxes and to understand variations between organisms, organs and cell types, as well as the differences observed during the lifetime of cells and the aging of the organisms to which they belong.

Interactions, networks and pathways

Here, we will continue to develop tools to handle complex novel datasets including single-cell proteomics and transcriptomics. We envisage specific applications in epigenetics in stem cells, single-cell RNA-seq in T cells, and transcription pathways and drugs in cancer. In the future we will perform integrative single-cell studies, correlating imaging, proteomics and transcriptomics. We will use and integrate such multi-omics data to uncover underlying biochemical pathways and mechanisms and modelling techniques to both validate and predict features of the systems that will be experimentally studied. The models will become increasingly complex as both cellular interactions and larger complexes continue to be elucidated but will be essential to explain key biological processes such as development, the cell cycle and infection and disease, including in humans. We aim to computationally integrate more measurements in model organisms, in which we also have access to organ functionality and organismal biology. We envisage an explosion not only in data around such topics but also in multi-scale modelling to combine inter-molecular, intra-cellular and increasingly also inter-cellular interactions to bridge molecular and higher-order processes.

Beyond the cell: cell–cell and species–species interactions in the context of populations and environment

Driven by data on cell–cell communication or shotgun sequencing of entire microbial communities, computational biology at EMBL will increasingly tackle the interactions between cells or (single-cell) organisms at molecular resolution (Section B.2.4.5). Spatiotemporal information, provided by imaging, mass spectrometry, microfluidics and longitudinal studies, will be used computationally to deduce and model networks of cells and species within environments, analogous to the modelling of molecules within cells discussed above. Populations of cells reveal stochastic and self-organisational elements, which are important for understanding processes such as development or the evolution of multicellularity. Modelling such systems requires a controlled environment to allow the study of adaptation to particular exogenous factors or perturbations introduced experimentally. Such studies will reveal basic interaction principles as well as identifying specific interactions within populations of cells or species. In many ecosystems, the environment can be considered as a variety of physicochemical parameters and other abiotic factors that can often be defined and measured. However, outcomes are also influenced by lifestyle or interactions with other species. Capturing and modelling these effects, by learning how to abstract knowledge from heterogeneous datasets that all attempt to describe the environment, is a major challenge ahead that we are working towards.

Taken together, EMBL's strategy is to exploit computational biology to advance our scientific understanding of basic biology – from molecules to organisms to ecosystems – using computational approaches to capture, analyse and interpret large amounts of quantitative molecular data. In this way, borders between disciplines will diminish as information, such as medical and environmental data, become part of the integration and modelling process. The overarching role of bioinformatics, bridging between scales and fields, will also ensure that the common molecular foundation underpinning all life, captured as data, resources and derived knowledge, will empower future developments in a wide range of applied disciplines such as agriculture, biotechnology and medicine. The key to all these advances in the years to come is the handling, integration and interpretation of Big data.

Box B.2.1: Data integration at various spatiotemporal scales - Tara Oceans project

Biological research in the 21st century has become increasingly quantitative and public repositories are being populated with data at an exponential rate. Biological systems range from individual proteins – whose structures are determined at atomic resolution and whose dynamics are modelled in femtoseconds – via cells, tissues, whole organisms and even to entire ecosystems, which are studied in an evolutionary context that covers the 4 billion or so years of evolution on earth. All these systems are interrogated by an array of '-omics' technologies, tracking molecular processes that capture the information flow that is inherent in life, from the genome via the transcriptome, to the proteome and metabolome and capturing changes over many time-scales. While DNA, RNA, proteins and metabolites all represent cellular components, the readouts are not restricted to cells *per se*, but they can spatially comprise any biological system ranging from single cells, whole tissues and organisms to species communities or even entire ecosystems.

The vast amount of these different types of data being collected provides a unique opportunity for integrative bioinformatics to unveil biological patterns across multiple scales – both spatial and temporal – in previously unimaginable ways. Some challenges and opportunities of multi-omics integration across space and time are illustrated by the Tara Oceans project, which was initiated by researchers at EMBL and advocates a holistic approach to advance our understanding of ecosystems. During a four-year journey, the research schooner *Tara* systematically collected more than 35,000 biological samples, which were subjected to molecular, cellular, morphological, environmental and geochemical analyses. The latter provide context for the organisms captured in those samples, for example in the form of microscopic pictures or DNA profiles. The wealth of heterogeneous data and metadata collected is unprecedented and will remain in use for integrative analyses for years to come.

So far, Tara Oceans has provided environmental 'shotgun' sequencing of plankton from all over the world, encompassing organisms ranging in size from small DNA viruses to fish larvae, and complemented these efforts by single-cell sequencing of some of the most abundant but poorly described organisms on earth. Researchers at EMBL have already compiled an extensive 'parts list' of the species and genomic content of hundreds of sites in the oceans, which has been made available to the scientific community. Correlating genomic data with imaging-derived morphological features provides a new method to better characterise the unknown fraction of organismal life. Integration of contextual and metagenomic data has already revealed temperature as a main driving factor for species compositions and the evolution of populations; complementary metatranscriptomic sequencing provides information on the physiological responses of plankton, e.g. in response to the natural fertilisation of ocean waters in the wake of islands.

One huge future challenge will be to better understand evolutionary adaptation to changing environmental conditions on a global scale. For example, with the rising seawater temperatures caused by global warming, microbial communities in the oceans will be considerably reshaped and it will be important to develop methods to predict the consequences. Taken together, multi-omic integration through space and time provides the framework to comprehensively address the fundamental question: "Who is doing what, where, when, and why?", which becomes particularly important given our goal of understanding and predicting the impact of contemporary climate change on global ecosystems.

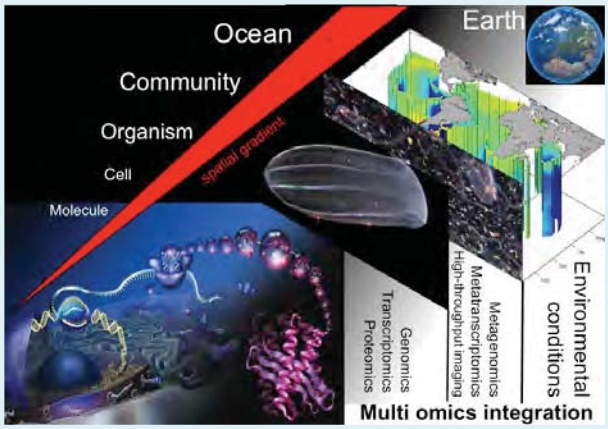


Figure B.2.1 Illustration of different data types generated and derived for planktonic organisms and their environment in the context of the Tara Oceans project. They cover spatial scales from single molecules to entire ecosystems, but also temporal scales, ranging from femtoseconds to years.

2.2 Decoding molecular processes: understanding how the genome gives rise to cellular, organismal and disease states

Genomics – the study of the genome and its derivative products (such as RNA, proteins and metabolites) – has played a major methodological role in advancing life science research over the past decade. The ongoing development of next-generation sequencing (NGS) has had a dramatic impact, permeating almost all research domains including cell, developmental, and organismal biology. Genomic research at EMBL therefore spans multiple Units, with the aim of understanding how the genetic blueprint of the genome ultimately gives rise to form and function in living systems and, through its variation, to individuality and evolution.

During the current and past Programmes, EMBL scientists have, for example, charted genomes of model organisms and humans, discovered pervasive non-coding transcription, defined enhancer architecture, identified genes that control cell-cycle events, uncovered catastrophic genomic rearrangements in cancer patients and identified the populations of microorganisms that are present in human gut samples (Section B.1). Although the technical limitations to sequencing have been dramatically reduced – enabling the enormous recent expansion in our knowledge of which parts of the genome are transcribed into RNA and which bind to regulatory factors – we are still trying to catch up in our understanding of genome function, with the role of only a very small percentage of this sequence in humans currently being defined. One major challenge is to identify the function of the 98.5% of the genome that is not protein coding. There is increasing evidence that these regions contain many segments with diverse regulatory roles, but it will take extensive research to untangle these functions. This is true not only in humans but also in model systems, and at EMBL we have, for example, recently identified general and specific roles for non-coding RNAs, including yeast antisense RNAs and vertebrate long non-coding (lnc)RNAs. EMBL also leads the data analysis efforts in the international ENCODE (Encyclopedia of DNA Elements) effort that aims to identify all functional elements in the human genome sequence. To achieve breakthroughs, the development and application of new genomic approaches remain at the core of EMBL's aims in this still young, yet rapidly advancing, field.

EMBL is well positioned to continue its contribution to biological discovery through the application of genomic technologies due to the interactive structure between its creative research groups and the Core Facilities (Section C.3), which disseminate these technologies not only within EMBL but, capacity allowing, also to scientists in our member states. EMBL scientists have played a major role in innovative use of genomic technologies, such as tiling arrays and NGS applications, and have made important contributions to the development of new computational methods for the analysis of different emerging data types. The synergy and close proximity between research groups developing new experimental methods, the workshops that help to implement them, the Core Facilities, and computational groups that provide the statistical expertise to analyse the data, provides a unique environment. In the examples below, specific areas are highlighted in which we expect major developments over the course of the next EMBL Programme and in which EMBL is in a strong position to take a leading role.

2.2.1 Single-cell diversity

Recent studies have uncovered a surprising degree of heterogeneity in gene expression between individual cells within 'homogenous' populations. Together with the known genetic evolution of somatic tumour tissues, this observation makes it increasingly clear that irrespective of whether they are healthy or diseased, single cells differ in their genetic composition or in how they utilise their genomes for gene expression. This heterogeneity can have important phenotypic consequences. For example, changes in gene expression can provide additional means for single cells to acquire fitness advantages over their neighbours, even when they are genetically identical. Understanding differences between individual cells will be important for the treatment of human diseases (Section B.2.3). Cancer, for instance, results from single cells escaping regular growth control, and reoccurs by single cells escaping therapy. These escapers proliferate to cause tumours. Studying the genetic and molecular composition of individual cells will therefore be informative for understanding and treating cancer. Today, single-cell sequencing of DNA or RNA using next-generation sequencers enables the detection of cellular differences at high resolution and provides new insights into how single cells of the same type function, sometimes differentially, in the context of their microenvironment.

EMBL has made significant contributions to this new era of single-cell genomics. It has contributed to analysing heterogeneous cell samples, characterising transcript isoform diversity, correlating transcriptomic and *in situ* data, and dissecting cell-lineage relationships. Our investment into microfluidic methods that can be used, among other things, to separate single cells from a variety of sources have contributed to our ability to study diverse systems in this area. Currently, for example, single-cell RNA sequencing is being applied to understand heterogeneity in gene expression in cells of the thymus, where heterogeneity seems to be crucial for the development of self-tolerance in the immune system. Advanced computational tools and statistical models for the analysis of these challenging datasets have been developed at EMBL. These enable biological and technical variations to be reliably distinguished, and identify and correct for interesting and unavoidable biological confounding factors such as cell-cycle effects during data analysis. In addition, EMBL scientists have applied these methods to study cellular bet-hedging strategies by which individual cell-to-cell variation can be used as a means to promote survival of the population of cells (and in many cases the organism) when environmental conditions change. While it is well known that bacteria exploit gene expression 'noise' to implement such evolutionary bet-hedging strategies, the characterisation of the extent and consequences of transcriptomic variability in higher organisms, even in yeast, has only recently begun. Recent studies by EMBL scientists, which are currently being extended, indicate that such stochastic cell-to-cell expression heterogeneity followed by signal reinforcement underlies lineage separation events in the early mouse embryo.

Given the current technical challenges, and the fast pace of methods development, groups interested in single cells at several EMBL sites will continue to exchange ideas and technologies. In 2013, the Sanger Institute-EBI Single-Cell Genomics Centre was established on the Wellcome Trust Genome Campus in Hinxton, UK, to pioneer technological and computational approaches for single-cell genomics.

In the next Indicative Scheme, these tools will be applied to different biological questions (Box B.2.2), including a detailed analysis of bet-hedging strategies in clonal populations of single-celled organisms, cell-fate decisions during the development of multicellular organisms and cancer-cell survival strategies that lead to cancer recurrence. Future

projects will also allow the characterisation of rare cell types, such as microbes from mixed populations that cannot be cultured, and thereby provide molecular insights into organisms that have so far escaped laboratory analysis. A major future goal is also to combine different types of single-cell information (e.g. DNA and RNA sequencing, bisulphite sequencing to identify epigenetically regulated DNA methylation sites, etc.) to understand the relationship between DNA sequence, RNA expression and epigenetics at single-cell resolution. Combination of these methods with single-cell imaging technologies, including imaging mass spectrometry, also has the potential to provide the link between the genome expression status of an individual cell, aspects of its environment and its phenotype.

Cellular decisions are often thought of as intrinsic autonomous events, especially in the light of recent studies using embryonic stem cell models. In reality, a cell within a multicellular organism rarely acts in isolation but rather continuously interacts, via bi-directional signalling, with its three-dimensional environment (Section B.2.4.3). Examples of this are cues provided by cellular niches to give stems cells their pluripotent properties or inductive signals that trigger cell identity or reciprocal contact-based inhibition to reinforce and maintain cell states. Also, tumour cells are influenced by, and in turn influence, the surrounding 'normal' tissues. Tackling this inherent multicellularity poses a major future challenge. In the coming years, several EMBL groups across multiple Units will develop new strategies to combine single-cell RNA sequencing of animal model embryos with three-dimensional spatial information from imaging cell positions (brainbow, RNA FISH) to map a cell's molecular fingerprint back into its original *in vivo* position in the tissue or embryo from which it was isolated. New methods in large-scale quantitative *in situ* hybridisation will also be explored. To complement these efforts, expertise in modelling how cells are organised into three-dimensional tissues will be strengthened and developed through the creation of a new EMBL outstation for Tissue Biology and Disease Modelling (Section B.2.5). Together, these approaches will link gene expression and genome regulation to spatial position and thereby aid our understanding of how cells respond differentially within a tissue or a field of cells depending on their environment (e.g. distance to a gradient source of a diffusible signalling molecule, developmental history) and facilitate the prediction of developmental trajectories.

Perturbation-based studies will be used to measure the functional relevance of expression changes at the single-cell level. Efforts in this direction will include using light-activated switches in gene expression that can be triggered in single cells within an organ or organism at precise developmental stages. The ability to generate such tools builds on EMBL's complementary expertise in cell, genome and organismal biology, our access to Core Facilities in the area of genomics and advanced light microscopy and to the Sanger Institute-EBI Single-Cell Genomics Centre, which provides state-of-art analysis platforms.

2.2.2 Analysis of causality and mechanism

In parallel to dramatic increases in data quantity, the past decade has seen a revolution in the qualitative and quantitative nature of large-scale data acquisition. In the study of gene expression, for example, new advances in NGS have enabled quantitation of single transcript molecules simultaneously for all transcripts that are expressed in a cell at a given state. NGS provides crucial information on where the regulators of gene expression, such as transcription factors, bind throughout the genome with base-pair resolution, a degree of precision that was unobtainable only five years ago. This shift to more quantitative and high-resolution data is opening up new possibilities in genomics,

facilitating mechanistic insights into how the genome is utilised and regulated. During the next Programme, we will continue to develop and make use of advanced technologies to move from correlation to causation, combining exploration with new mechanistic insights. EMBL aims to play a leading role in three key areas:

Deciphering fundamental principles in the regulation of gene expression

The information stored within the genome defines the totality of programmes available to regulate biological processes. Understanding the precise regulation of gene expression is therefore essential and requires information on multiple steps, including the transcriptional, post-transcriptional and post-translational levels. EMBL has a long tradition in deciphering mechanisms that act on these processes. Recent examples include new models of how enhancers (*cis*-regulatory elements) function, the discovery that antisense transcription is both pervasive and regulatory, and a new understanding of how a specific type of non-coding RNA – Piwi-interacting RNAs (piRNAs) – can silence regions of the genome.

In the next five-year Programme, our goal is to link each regulatory step of gene expression, from transcription-factor binding, chromatin remodelling, RNA polymerase II recruitment and elongation, mRNA capping and processing, all the way to the recruitment of the mRNA by the translational machinery. Many of these steps could previously only be studied using single-gene approaches, but they can all now be measured genome-wide. At the chromatin level, EMBL scientists will, for example, integrate genomics, biochemistry, structural biology and cell biology to understand how mutations in histone-modifying enzymes, commonly found in patients with neurodevelopmental disorders and cancer, affect histone function and cellular phenotypes (Box B.2.11, see also Section B.2.4.4). They will also study mechanisms involved in the regulation of untranslated regions of mRNAs in various contexts ranging from single-celled organisms to embryonic development and embryonic stem cell differentiation.

Measuring dynamics and real-time kinetics of genome regulation

Currently, very little information is available on the kinetics and dynamics of how the genome is regulated in real time. This includes both the tight deployment of protein function in time and space to regulate genome activity throughout development, as well as stochastic and sporadic properties of transcription at the single-cell level. However, this temporal dimension is essential to understand and model the inherent robustness of cellular decisions and phenotypes.

One exciting recent development is the ability to investigate how chromatin structure is organised in three dimensions within the nucleus. New developments in single-cell genomic technologies and super-resolution imaging, such as 3D single molecule localisation microscopy (STORM and PALM), open up the possibility to study some of these processes in real time (Section E.1.1.2.3). During the next five-year Programme, EMBL will invest in pushing the resolution at which genome regulation can be studied to its limits, making use of our strengths in genomics, imaging and bioinformatics. Several EMBL groups across multiple Research Units will image different aspects of genome topology and gene expression in cell culture-based systems and in developing *Drosophila* and early mouse embryos. This will involve new technology development to facilitate, for example, the study of enhancer looping to determine its properties and dependencies during transcriptional regulation (Box B.2.3). For this purpose, new super-resolution microscope development, as well as strategies to tag specific genes, loci and transcripts, both for visualisation and targeted biochemistry, will be developed.

Causality modelling

Identifying the molecular consequences of genetic variation is central to decoding genome function, predicting phenotypic effects and developing therapeutic interventions. It is hampered by the difficulty of identifying the causal pathways that mediate a genetic variant's effect on phenotype and distinguishing these from pathways modulated as a consequence of either the phenotype itself or of other sources of variability such as environmental conditions. Molecular profiling is an excellent approach to address this issue and EMBL has played a leading role in this area, making important contributions to the study of natural variation in model organisms, genomic abnormalities in cancer, and genetic variants that predispose individuals to inherited diseases. During the past five years, many studies (often involving EMBL scientists) have successfully linked molecular data from transcriptomics, proteomics or metabolomics with specific DNA-sequence variants (Section B.1). The challenge now is to identify the causal relationships within these extensive correlative datasets. Recently, we have been successful in deriving causal inference from linked molecular datasets for gene expression traits measured in yeast segregants by determining which among multiple genetic variants that lead to quantitative differences in the levels of expression of genes (eQTLs) were causally important for fitness. The conclusion of this work was that it was necessary to study a large population of yeast genetic variants in multiple environmental conditions to identify which were causal. These studies have implications for the interpretation of cell-to-cell variability in multicellular organisms, for example in the choice of which of the potential targets identified in genomics studies should be used for intervention in precision medicine. In the next Programme period, we therefore wish to extend this approach to more complex eukaryotic systems.

Similar causality modelling will be used in the context of the Pan-Cancer Analysis of Whole Genomes project, an international consortium co-directed by EMBL scientists that aims to integrate all human cancer data into the largest world-wide resource to date that links genome information to phenotype (Section 2.3.5, Box B.2.4). EMBL is working on collecting extensive genomic and molecular data for several tumour samples and using causal inference to identify the molecular pathways that lie immediately downstream of genetic variants. Given EMBL's role in integrating these data, its expertise in generating quantitative large-scale data and its strength in computational analysis and statistical modelling, we are well positioned to make a strong contribution to the derivation of predictive models of biological processes and their perturbation by sequence variation.

2.2.3 Synergistic application of proteomics with other genomics technologies

Despite the enormous success of NGS technologies, the predictive power of how transcriptional regulation ultimately affects the proteome remains limited. mRNA and protein abundances are frequently poorly correlated because translation rates and degradation of proteins are often regulated independently from transcription. Post-translational modifications, small molecule interactions and the formation of complex protein-RNA interaction networks are examples of processes that further modulate protein function. Therefore, the synergistic combination of proteomics and genomics technologies is required to ultimately understand how genomic variation and gene expression lead to phenotypes.

Mass spectrometry-based proteomics has become increasingly comprehensive and can provide quantitative information on protein abundance and post-translational modifications. EMBL scientists have developed new mass spectrometric techniques to

study protein-RNA interactions, protein-lipid interactions and the architecture of protein complexes. In the past three years at EMBL, these technological advances facilitated e.g. the discovery of hundreds of novel RNA-binding proteins (see below), a new mechanism of lipid transport from the ER to the cytoplasmic membrane and the elucidation of nuclear pore complex architecture (Box B.2.7). Proteomics technologies hold great potential to bridge across different disciplines and to integrate genome and proteome research with cell, structural and developmental biology. In the next Indicative Scheme, the Proteomics Core Facility will continue to work closely together with scientist from all Units to develop novel experimental workflows tailored to their specific requirements.

Although mass spectrometry-based proteomics is not yet applicable to the single cell level, EMBL scientists have made critical contributions to making protein and proteome analysis more sensitive. Thanks to these advances, samples of limited availability, such as small pools of hundreds of cells can nowadays be analysed. In the 2017 - 2021 period, we anticipate that these advances will contribute to a more rapid and comprehensive proteotyping, potentially even in a clinical setting. The integration of proteomic state information on protein isoform expression, post-translational modifications, conditional protein localisation, molecular interactions and protein conformation with data from structural models, genome expression, imaging and metabolome data will open new avenues to understanding complex regulatory networks.

2.2.4 Integrating metabolism and genome regulation

Although metabolism has been studied in isolation as a ‘housekeeping’ function for decades, an emerging view emphasises that metabolic activities and cellular functions are tightly intertwined and interdependent. How the metabolic state is affected by environmental cues and how it is connected to the cellular state, and especially the elucidation of the underlying molecular mechanisms and functional role of these interconnections, is a common interest among several EMBL groups across multiple Units. Metabolites represent about 50% of the content of a cell, yet we understand very little about their localisation, activity and function. The interpretation of the relationship between genomes and phenomes is hampered by the absence of a comprehensive analysis of all cellular building blocks and in this respect, charting metabolites is crucial.

Metabolites have widespread functions beyond being building blocks or metabolic intermediates. Many metabolites are essential co-factors for post-translational and epigenetic modifications. Metabolic intermediates exert unexpected signalling roles, for instance as ligands for G-coupled receptors, and several families of lipid-derived small molecules participate in a broad range of cellular processes. In addition, many metabolic enzymes have moonlighting (that is, additional) functions in the regulation of gene expression, as highlighted by a recent EMBL study that integrated our strengths in biochemistry, genomics and bioinformatics to identify hundreds of new RNA-binding proteins, many of which are ‘classic’ metabolic enzymes. This led to the hypothesis that RNA, enzymes and metabolites are tightly functionally linked and that RNA may have a regulatory role in controlling enzyme activity as, for example, observed in T-cell activation. In the next Indicative Scheme, we will use a new proteomic technology developed at EMBL to determine the sites at which these so-called ‘enigmatic RNA-Binding Proteins’ (enigmRBPs) interact with RNA. In addition, a collaborative project, involving several Research Units and the Proteomics Core Facility, whose aim is to comprehensively identify the RNAs that are bound by the hundreds of enigmRBPs as well as the exact position of binding sites within each RNA, will be undertaken. The project will record metabolic profiles of strategically chosen mutants and investigate the

interaction of enigmRBPs with (effector) RNAs through structural studies. In select cases, the investigation of clinical specimens could be informative regarding human (metabolic) disorders. The resulting information will determine whether genomes can directly affect the functions of mature proteins via effector RNAs, and may provide a fundamental link between diet and genome usage.

In the next Programme, EMBL will focus on visualising, measuring, and annotating metabolic activities, identifying molecular mechanisms linking metabolism and cellular functions, and studying the physiological effects of metabolic state/environment in their *in vivo* context, using high-sensitivity metabolomic and dynamic flux measurements. To enhance these efforts, a new group in imaging mass spectrometry has been recruited who will be able to visualise spatiotemporal dynamics of metabolites and metabolic activities. Linked to this research group, a new Metabolomics Core Facility will be established to provide access to these technologies. Novel fluorescent sensors for key metabolites and *in vivo* testing using model organisms will also be developed (Section E.1.1.2.1). Community standards and databases for metabolomics data will be generated at EMBL-EBI. Ultimately the data for transcriptomes, proteomes and metabolomes will need to be analysed collectively and, as further elucidated in the section Analysing Molecular Data (Section B.2.1), many of the most difficult challenges in this area lie in computational biology.

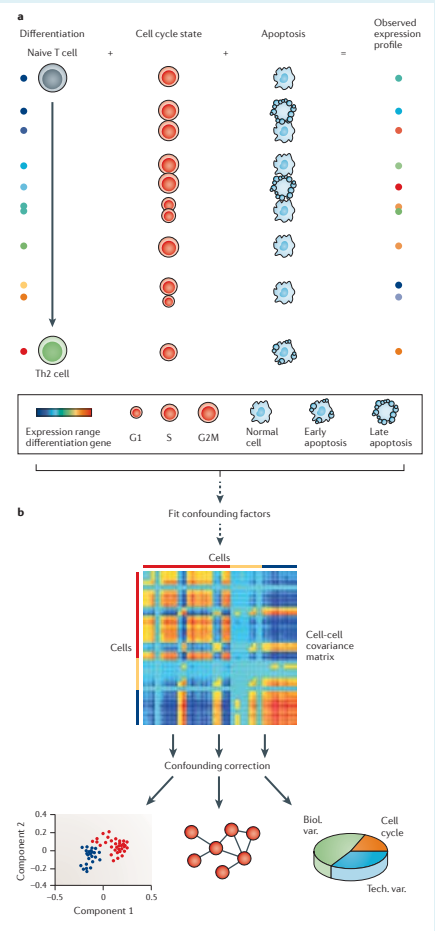
2.2.5 Conclusion

In summary, genomic science has permeated most life science research domains through the precision and comprehensiveness afforded by systematic, comprehensive molecular measurement. Recent developments have been particularly facilitated by ongoing advances in NGS and mass spectrometry. As these methods enable accurate measurements across thousands of genes, genomics can reveal patterns in biological processes that are not detectable by single-gene level analyses. Thus, it is only through genomics that many biological mechanisms have become apparent. Examples include the universality of bidirectional promoters and the extensive functional importance of what was previously considered ‘junk’ DNA. Over the next EMBL Programme, we expect further advances in the resolution of genomics that will enable population measurements to move towards the single-cell level. This step will provide insights into biological and evolutionary processes and their robustness. The Programme will also focus on the further integration of dynamic genomic data across layers of biological information by enabling the integration of imaging data into our current repertoire and improving approaches to causality inference among correlated data to understand information flow and mechanisms. This will enable the systematic interrogation of increasingly complex processes such as the cross-talk between metabolism and genome expression.

Box B.2.2: Single-cell genomics to analyse cellular diversity

Cell identity and function is often characterised at the molecular level by unique transcriptomic signatures. Characterisation of the extent and consequences of transcriptomic variability in higher organisms has only begun recently, driven by advances in single-cell transcriptomics. Gene expression variability can potentially provide explanations for many long-standing questions in disease and developmental biology. For example, why do some cancer cells survive chemotherapy, even in tumors that are genetically relatively homogeneous? How do adult stem cells find the right balance between self-renewal and differentiation? What explains the low penetrance of genetic and environmental factors in neurodegenerative disease?

In the current Indicative Scheme EMBL has built up important expertise and infrastructure in the area of single cell genomics, e.g. the Sanger – EBI Single-Cell Genomics Centre and complementary resources for very sensitive ChIP (Chromatin Immunoprecipitation) protocols developed in EMBL Heidelberg. On the computational side EMBL scientists are developing normalisation strategies, methods that allow controlling for confounding variables such as the cell cycle (Figure B.2.2) and approaches for modelling single-cell gene expression within a spatial context. These resources put EMBL in a strong position to push the future investigation of transcriptomic variability in higher organisms.



Developing these approaches and combining them with strategies for placing cells on temporally defined lineages will facilitate insights into stochastic cell fate decisions, e.g. during haematopoesis, embryonic development and adaptive immune responses. In the latter area EMBL scientists are for example investigating immune homeostasis and have already discovered a steroidogenic T-cell subtype that drives immunosuppression through the secretion of pregnenolone. Understanding cell type identity in a multicellular tissue requires the integration of each cell's expression profile with its spatial location within the tissue under study. EMBL scientists have developed a high-throughput method to identify the precise spatial origin of cells assayed using single-cell RNA sequencing.

In the next Programme, we are planning to apply this method to several model organisms, including sponges, cnidarians and amphioxus, to obtain insights into brain evolution and development. EMBL has also started collaborations with clinical researchers, who can provide access to human tumour and adult stem cell samples, to apply these technologies to characterise the transcriptomes of 1000s of single cells, for example before and after chemotherapy. These strategies will ultimately unmask mechanisms driving functional diversification of cellular populations and their relevance in human disease.

Figure B.2.2 Modelling single-cell RNA sequencing data. a) Each cell's expression profile is a combination of factors. b) Identifying confounding factors and correcting for them in downstream analysis.

Box B.2.3: Towards dynamic structure and function of the genome

Genome function relies on its highly compacted three-dimensional (3D) topology in the nucleus, the disruption of which impacts on transcription, replication and stable inheritance of the genome. To date, the 3D organisation of the genome and its dynamic properties have been studied by two largely separate communities. On the one hand, genomics methods including chromatin immunoprecipitation (ChIP) and multi-C DNA methods have revealed dramatic changes in global chromatin accessibility during development and provided new insights into the proximity of all genome sequences to each other. As these methods average over large populations of nuclei, they reveal invaluable insights on the most frequent structures, but miss rare or dynamic interactions. Imaging approaches, on the other hand, have provided single-cell views of the relative proximity of individual loci and of the position of whole chromosome territories. However, as the structural elements of chromatin lie below the diffraction limit of light, imaging has not been able to map the linear genomic sequence into physical space. In addition, most, of our current knowledge comes from fixed samples, with very little information on the real-time kinetics of changes in chromatin topology.

To bridge this gap, EMBL scientists will integrate new advances in genomics, super-resolution microscopy, and genome editing to answer fundamental questions about genome structure and transcriptional regulation. We will push the spatial resolution of imaging approaches to systematically map the location of sequence elements within a single nucleus, combining newly developed isotropic super-resolution microscopy and computer simulation of DNA sequence to generate the first 3D physical map of the human genome in single cells. Live imaging of diagnostic structural loci will be used to understand the topological transformation required to form compact mitotic chromosomes. To assess the dynamics of chromatin topology, changes in enhancer-promoter interactions identified by Hi-C during *Drosophila* development will be imaged in live embryos in the early blastoderm and manipulated through targeted deletions. These research endeavours require major efforts in technology development, including advanced algorithm, labelling tools and microscopy to explore genome structure in single cells and live embryos. EMBL's strength in technology development, genomics, microscopy and computational biology puts us in a strong position to tackle these challenging issues, and provide fundamental new insights into the structure and dynamic function of the genome.

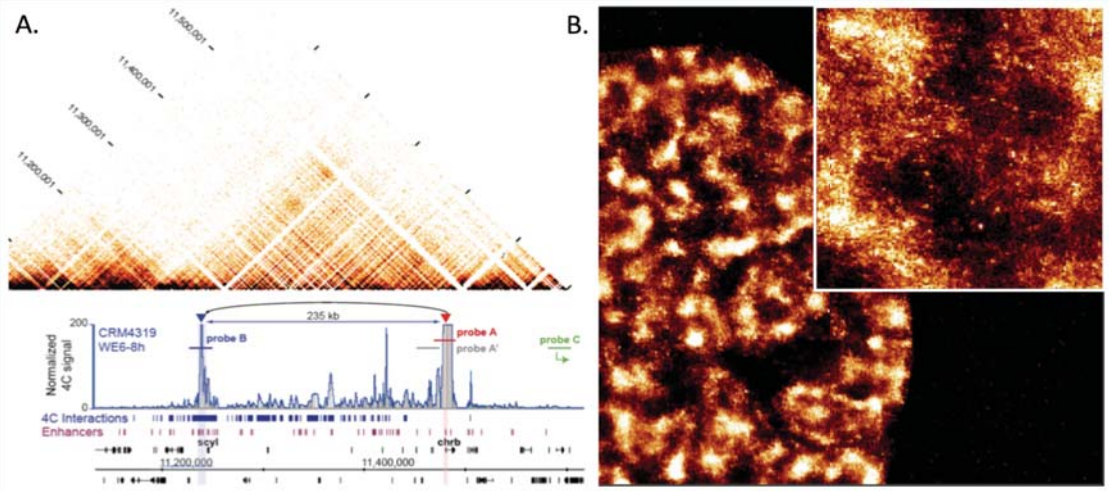


Figure B.2.3 Intersecting molecular and imaging views of chromatin structure. A. Hi-C (above) and 4C (below) molecular data showing chromatin proximity within a given locus, from a *Drosophila* embryo. B. Super-resolution microscopy image of chromatin fibres emanating from the surface of a human chromosome in late G2.

2.3 Unravelling molecular processes in humans

The past decade has seen a renaissance in human genetics as a result of technological advances that have transformed opportunities for basic and translational research in human systems. The advent of high-throughput technologies has paved the way for the global interrogation of the human genome, transcriptome, proteome, cellular phenome and, most recently, metabolome (Section B.2.2). There has been a precipitous drop in data-acquisition costs, both by nucleic-acid sequencing and imaging at the cellular, tissue and whole organism level. The size of the human genome (diploid size, 6 Gbases) is no longer a significant impediment to genetics experiments, with the cost of a whole-genome polymorphism array below €50 and whole-genome sequencing below €1,000. Proteomics and metabolomics technologies are also poised to experience important improvements in sensitivity and cost. Studying the human microbiome (the complex composition of bacteria that inhabit the human body) is an example of a new, important area of research on humans that has been made possible only by low-cost, high-throughput sequencing.

New cell-based systems and approaches have also changed the way human genetics is studied. For example, human induced pluripotent cells (iPSCs) enable a defined cocktail of factors to convert a skin fibroblast or blood cell precursor into an embryonic-like state from which a variety of cell types can be derived. Such protocols are now applied in numerous laboratories worldwide. Similarly, the development of organoids has started to revolutionise the way in which we can investigate molecular processes in humans. Organoids are cell collections that either self-organise or can be induced to differentiate when grown in defined three-dimensional (3D) matrix supports (Section B.2.5). They allow the creation of *in vitro* tissue ensembles (e.g. gut or brain) that mimic many aspects of *in vivo* tissue organisation and function. Finally, the remarkable developments in CrispR/Cas9-based technology allow the introduction of specific targeted mutations to human cells and a variety of model systems. Thus, genetically defined human cells or organoids can now be phenotyped by high-content imaging and genomic readouts. Robotic and microfluidic solutions allow these technologies to be applied at high throughput, and offer an exciting possibility to directly correlate high-content imaging with single-cell genomics.

In addition to imaging approaches, EMBL's strengths in high-throughput molecular profiling technologies, such as RNA-sequencing, proteomics, and large-scale drug screens, and their advanced computational analysis, will be applied to human systems. In the next period EMBL will pursue multi-omics approaches in cells derived from patients to investigate the genetic basis of complex hereditary traits and diseases to shed light on potential disease mechanisms. This will build on recent work in which EMBL scientists have used molecular profiling in HapMap cell lines and found extensive variation of the activity of regulatory elements (such as enhancers and promoters) across healthy human individuals. As these regulatory variants are highly overlapping with genetic variants associated with complex genetic diseases, they point towards potential disease mechanisms.

These technical developments and the unique ability of EMBL to integrate all of them, both experimentally and analytically, have led to a remarkable increase of

molecular and cellular research occurring in human systems at EMBL, in the context of health and disease alike. Building on these ongoing activities, we are planning to make important contributions to unravelling molecular processes in humans in the three areas outlined below.

2.3.1 New imaging technologies for human model systems: from cells to organoids and tissues

Recent breakthroughs in imaging will make it possible to directly assess molecular processes and measure their dynamics in human systems (Sections B.2.4.2 & E.1.1.2). Studies on human cells, cell-based in vitro 3D cultures and organoids directly profit from the imaging technology developments pioneered in research on model systems. High-throughput microscopy allows to rapidly phenotype human single-cell systems such as disease model cell lines, iPSCs, or their genome edited versions. At EMBL this approach has already revealed genetic variance and molecular mechanisms behind cystic fibrosis, hepatitis C infection or regulation of cholesterol metabolism.

Due to the increasing size and complexity of disease-relevant multicellular human experimental systems, large volume or ‘mesoscopic’ imaging based on light sheet technology, first invented at EMBL, is also going to be a major enabling technology for human molecular biology (Section E.1.1.2.4). Applying mesoscopic imaging approaches at the tissue level, EMBL’s new outstation in Barcelona will investigate phenomena directly relevant to human health, such as diseases of the immune system, cancer, or congenital abnormalities. Model systems and organisms ranging from human 3D cell and organ culture to the mouse, will play an important role in the new outstation for Tissue Biology and Disease Modelling (Section B.2.5), because they allow the predictions of computer models about disease mechanism to be tested experimentally.

Finally, to be able to report on defined molecular mechanisms directly in primary cells from patients, where genetic manipulation is not an option, new fluorescent reporters and tools that can be rapidly applied ex vivo to patient samples are needed. EMBL’s strength in chemical biology has enabled the development of many such probes, that allow to non-invasively image dynamic processes in living cells with high spatial and temporal resolution (Section E.1.1.2.1). Researchers at EMBL Heidelberg and Monterotondo have successfully applied these reporters to mouse models of human disease, e.g. to investigate chronic obstructive pulmonary disease (COPD), chronic pain states, tumour formation or to monitor lesions in an osteoarthritic animal model. In the context of the Molecular Medicine Partnership Unit (Section F.1.1.4), EMBL scientists and collaborators from the University Clinic Heidelberg have applied fluorescent probes to cells from patients suffering from cystic fibrosis and COPD. Such molecularly defined functional assays have the potential to serve as more precise biomarkers than global quantitation of RNA or protein molecules. They are non-invasive and require little cellular material, which makes them well suited for future use in clinical diagnosis. Together with their clinical collaborators, EMBL scientists will work towards introducing such imaging based diagnostic tools into the clinic.

2.3.2 Defining disease-relevant molecular processes in human systems

Harnessing our strengths in developing new tools, methods and technologies and applying them to complex biological systems, we will work towards understanding aspects of human biology at the molecular level. This effort will be based on EMBL's extensive experience with more tractable experimental systems, including a variety of cellular models (including standard human tissue culture cell lines) and model organisms (such as yeast, fruit fly, zebrafish and mouse). These experimental systems are where new techniques can be developed, refined and proven in cost effective ways. Since the model systems, unlike humans, allow more extensive experimental perturbation, it is often in them that new principles of biological organisation and mechanism are revealed. Once a new technique has been pioneered and proven in model systems and once a mechanism is understood, it can be transferred to the human systems to test if similar mechanisms are at work. At the same time hypotheses generated from increasingly available human data such as genome sequences, which can often only be based on correlation, can be transferred back into an experimental model system for functional validation and mechanistic dissection.

A good example of how research in model organisms fosters our understanding of human biology is EMBL's work on cell fate transitions. Cell fate determination as it occurs in stem cells, either during early embryogenesis or later in development for organogenesis, can be studied and perturbed in model systems. The same fundamental processes underlie major questions in human molecular biology and disease mechanisms, for example the early embryonic cell state changes that can lead to infertility or severe congenital diseases, or those that occur during initial tumorigenesis or when tumors relapse from therapy. Mouse embryonic stem cells provide one model to rapidly access the molecular mechanisms behind these differentiation processes. New breakthroughs in light-sheet microscopy developed at EMBL, have made it possible for the first time to observe the earliest physiological fate decision in stem cells in the context of the live mouse embryo in real time and will allow us to analyse the responsible molecular regulatory and structural protein networks in the future (Box B.2.8). This is vital to understand how cell states can become susceptible to change initially, and subsequently become locked in. At the same time it provides a unique opportunity to understand fundamental questions about the molecular mechanisms of human embryology and fertility, for example why early human embryos display a very high rate of mosaic aneuploidy (mis-segregation of chromosomes during early embryonic cell divisions) and how they can survive this severe perturbation.

Similar processes of cell state transition, often in the reverse direction of de-differentiation, are key during tumorigenesis and need to be studied in multicellular systems that recapitulate the physiological tissue environment that cancer arises in. Transgenic mouse disease models based on oncogenes that can be switched on and off at will in the appropriate cancer-prone tissue – for example the mammary gland as a model for breast cancer – provide unprecedented opportunities for molecular understanding. For subcellular and molecular investigation, the animal models are being complemented by human organoid cultures such as mammary acini, which can be accessed with the same powerful light sheet imaging methods as early mouse embryos but will in the

future allow us to observe the process of metastasis or tumour relapse in real time and to isolate and analyse the single cells involved in the earliest stages.

2.3.3 Systems medicine: Novel therapeutic approaches based on systems biology of disease mechanism in model systems

Another barrier in translational research that is more efficiently first addressed in model systems is the identification of molecular intervention points that can be therapeutically targeted. Such endeavours are necessary because knowledge of the causative genetic defect underlying a disease is frequently insufficient for the development of therapeutics. Systematic approaches have often helped identify effective molecular targets that are related to, but not directly implicated in, the disease and thus would not have been selected as potential targets. The relative complexity of molecular networks in humans makes the identification of such intervention points much more challenging. Yeast has proven effective as both a genetic and pharmacological model of numerous inherited diseases, due to its experimental tractability and extremely well-characterized biology. For example, the severe and largely untreatable mitochondrial ATP synthase disorders have been studied in yeast models by an international team led by researchers in the Genome Biology Unit at EMBL Heidelberg. Their integrated approach that includes screens of drug repurposing libraries and chemical genomics in addition to biochemical and genetic validation assays elucidated the causative mechanism behind this group of diseases and demonstrated that mitochondrial protein import is a target that can rescue the disorders in the model. Future work will focus on extending these approaches to move from the yeast to the human system, as well as tackling other Mendelian diseases, including further mitochondrial deficiencies, with the same multi-pronged approach. Such systems medicine approaches, that capitalise on the interdisciplinary expertise and collaborative atmosphere at EMBL, hold significant promise for uncovering novel strategies to intervene in the progression of disease.

2.3.4 Computational integration and mining of heterogeneous data from human systems

The experimental techniques described above generate large amounts of data, which have to be processed and analysed jointly, taking into account the more limited availability of patient samples and therefore much greater requirement for statistical sampling compared to animal model studies. Extracting knowledge requires integrative analysis of high-dimensional data – with many genes/molecular species being measured, potentially in many individual cells, coupled for example with multiple imaging modalities, often done on many cells simultaneously, i.e. in 3D and over time.

The analysis of such datasets is challenging for a variety of reasons. Firstly, it requires professional data management at large scale, while maintaining the metadata of experimental details and biological sample descriptions to ensure comparability with other studies. EMBL has a strong track record of delivering such data management solutions for both research and service, for example in the context of the Genome to Phenome Archive and the BioSamples database (Section C.1). Secondly, raw data files (sequence reads; pixel readouts) have to

be converted into standardised and useful parameters for downstream analysis, and EMBL has substantial experience in extracting validated metrics and representations from large datasets to allow comparisons with reference data. Thirdly, different experiments have to be integrated to analyse them together, which requires considerable statistical and algorithmic innovation. There are two, broadly complementary, approaches: data-driven modelling (where machine learning techniques are used to discover patterns in the data) and model-driven data analysis (where parameters that can be applied to well understood mechanistic models can be estimated from the data). EMBL scientists have a strong record in creating, applying and serving some of the most cutting-edge approaches in both these areas; for example, the development of feature models for human cell division, the development of accurate noise estimation in single cell genomics, and the use of latent variable analysis to model unwanted confounders in genomics data. We are therefore confident that we can continue to make important contributions to methods of data analysis in this area.

2.3.5 Molecular medicine: bridging between basic and clinical research

As a consequence of the important advances in understanding human biology at the molecular level, molecular biology is becoming increasingly relevant to clinical research and even clinical medicine. The application of basic biological knowledge to create diagnostic methods, clinical treatments and drugs has been greatly enhanced by the common language of molecular data, to the point where basic research and translational research, which have traditionally been quite distinct, now overlap.

EMBL is not a translational research institute, but its focus on molecular and cellular technology, in both human and experimental systems, coupled with its experience in data management and innovative integrative analysis, make it an obvious player in bridging the world of basic with translational and clinical research. Section C.1.3.5 describes how EMBL is planning to provide support for clinical researchers and practicing healthcare professionals in the future by providing bioinformatics services to deal with patient genome data. In our research activities we have recently experienced a great increase in requests for interaction with medical and clinical research communities and expect to see even more of these in the coming years. These interactions can be grouped into two broad categories: targeted partnerships and alliances with leading biomedical and translational institutions in the EMBL member states and participation in international consortia, often centred around cohort studies that generate large scale data.

Partnerships and alliances in molecular medicine

EMBL engages in two institutional Partnerships (Section F.1.1.4) for molecular medicine, the Nordic EMBL Partnership for Molecular Medicine and the Molecular Medicine Partnership Unit in Heidelberg. One strategic goal of these is to allow the duplication of the successful EMBL research model into member state organisations engaged in translational and clinical research. A second aim is to enable EMBL researchers to identify interested collaborators in order that they can apply their expertise in basic research directly to the understanding of disease states to the ultimate benefit of human patients.

The Nordic Partnership involves a network of partners in the Scandinavian countries – the Centre for Molecular Medicine Norway (NCMM), the Laboratory for Molecular Infection Medicine Sweden (MIMS), Institute for Molecular Medicine Finland (FIMM) and the Danish Research Institute of Translational Neuroscience (DANDRITE). Each of the partner institutions contributes unique and complementary expertise in biomedical disciplines, including epidemiology, cancer, clinical microbiology and translational neuroscience. In addition to their partnership with EMBL, the individual Nordic research centres engage in collaborations with other national partners, including research and public health institutes, hospitals and research councils, and have thereby established an extensive Nordic network for molecular medicine.

Since 2002, EMBL has cooperated with one of the leading Medical Centres in Germany, at the University of Heidelberg, within the Molecular Medicine Partnership Unit (MMPU). The MMPU presently consists of about a hundred clinical and basic researchers within eight teams that are jointly led by a member of the Faculty of Medicine and an EMBL group leader. These teams enjoy full access to the University hospital clinics as well as the facilities of EMBL. They address disease mechanisms using the full spectrum of molecular, genomics and imaging technologies as well as data analysis. The medical focus is on common, usually complex multifactorial, diseases including haematological diseases, metabolic disorders, chronic airway disease, chronic pain and HIV.

In addition, scientists at EMBL Heidelberg closely collaborate with the German Cancer Research Centre (DKFZ) and the National Centre for Tumour Diseases (NCT). For example, recent work on cancer genomes involving researchers in EMBL's Genome Biology Unit and collaborators at the DKFZ has provided new insights into the formation of complex rearrangements in childhood medulloblastoma, thus identifying the juxtaposition of enhancers through DNA rearrangements into the vicinity of oncogenes (i.e. 'enhancer hijacking') as a common mechanism in childhood brain tumours. They further discovered a genetic mutation that defined a subset of medulloblastoma cases and that is relevant to diagnosis and treatment.

Additional recent progress in understanding cancer aetiology came from EMBL's alliance with the University Hospital Eppendorf (UKE) in Hamburg, from a joint endeavour to understand the genetic changes connected to prostate cancer in young men, that is frequently more aggressive than later-developing prostate cancer. By analysing data collected from more than 12,000 patients at UKE, the team has found significant changes in the expression of hormone-regulated genes that are highly prevalent in younger patients and which correlate with disease severity. These are now under further investigation. .

Not surprisingly, given the rapid and ongoing increase in the collection of genomic data from patients, and the requirement in the medical profession for assistance in its analysis and interpretation, EMBL‑EBI also has an increasing number of close collaborations with local translational and clinical research institutions. One example is the newly established joint postdoctoral programme with the NIHR Cambridge Biomedical Research Centre that supports projects that apply computational approaches to translational clinical research involving human subjects.

The new EMBL outstation in Barcelona (Section B.2.5) is ideally situated to forge links with local clinical institutions and can build on several already existing connections with the Hospital del Mar Research Institute, located in the same building as the future outstation, the Vall d'Hebron University Hospital, and many specialized biomedical research institutions. Future collaborations are for example expected to provide access to primary cell lines from patients, tissue-banks and biopsies.

International consortia and cohort studies

Computational biology not only bridges between different biological research disciplines (Section B.2.1) but increasingly reaches into medicine. Bioinformatics is becoming the conduit to allow translation of new discoveries in human biology in research laboratories to enter clinical practice. This process is only just beginning and at this stage our aim is to ease the bottleneck to translation by increasing biological knowledge, especially for humans and pathogens; by handling and integrating data and by championing open data, as well as ethics and security of data standards.

We will identify and quantify environmental, demographic, social and lifestyle determinants of diseases in addition to genetic factors and make reference datasets available to allow access to clinical researchers. In pharmacogenomics we will work on applications to specific diseases and use data integration and systems readout for therapeutic target validation. We will utilise novel biomarkers such as human microbiota or metabolomics data for improved diagnostics and search for novel ways to overcome imbalanced gut microbiota in the disease state, and also to mitigate antibacterial resistance.

Another important areas for biology and medicine are dynamic profiles for regulatory networks and longitudinal profiles in medical samples. Over the next years we expect to see a steady increase in research and service activities with a stronger medical research emphasis, in particular in the area of bioinformatics, mostly at EMBL-EBI but also involving bioinformatics and genomics research in Heidelberg. This will build on our successful existing research and service activities in regulatory networks ranging from identification and characterisation of regulatory elements in the human genome to modelling complex biological processes, as well as data management activities for large-scale omics projects. Through these activities we have gained experience with the challenges associated with such datasets, the access to which is controlled by Data Access Committees; as well as the complexity introduced by the additional dimension of time. The latter aspect in particular will require the development of new storage methods and integrative analysis tools.

Much of this work takes place in the context of international consortia, such as the Pan-Cancer Analysis of Whole Genomes project (PCAWG, Box B.2.4), rare disease consortia, the International Human Microbiome Consortium (Box B.2.12) and many others, where EMBL often takes a leading role. Many of these consortia aim to systematically profile cohorts of patients or healthy individuals genetically, molecularly and, particularly with the help of imaging methods, also phenotypically. These cohorts are proving informative on clinical research questions as well as informing basic biology and they are likely to drastically increase in size and diversity over the coming indicative scheme. These studies demand innovative statistical analysis coupled with mechanistic insight to transform them from phenomological observations into mechanistic models of

biology and disease. EMBL, by being able to apply the latest technology and models, coupled with appropriate analysis, is in a very favourable position to provide this link (Box B.2.4).

In addition to these large cohorts, specific subsets of individuals are likely to be regrouped by specific genotypes for further analysis. In these focused cohorts more extensive molecular and cellular research will be undertaken, and these more detailed cohorts will be a specific opportunity for EMBL's to obtain new molecular, cellular and ultimately mechanistic insights.

Rare diseases are an area where the application of high-throughput genomics to many individuals will enable a large leap forward over the next few years. Already one can find the disease-causing mutation(s) by sequencing both the affected patient and his/her parents. So far the causative gene for about half of all Mendelian diseases is known. The challenge however remains to use this knowledge to help in developing a therapeutic intervention. One Mendelian disease EMBL has become active in, in the context of its alliance with Stanford University whose purpose is to link EMBL's expertise in basic research to the strengths of Stanford in clinical technology development and application, is NGLY1 deficiency. N-glycanase 1 (NGLY1) is a part of the misfolded protein degradation machinery in the cell. Non-functional NGLY1 manifests in patients as a global mental and physical developmental delay along with a host of other complicating clinical phenotypes. As part of an international consortium of researchers, EMBL scientists are involved in characterizing the molecular phenotypes of NGLY1 deficient cells and tissues in order to screen for genetic and chemical suppressors of the disease state. The clinical molecular analysis and screening carried out for this project is unique in its scope, patient and parent participation, and potential for immediate effect with the goal of laying down a blueprint for future rare disease research efforts.

Another example is the Human Induced Pluripotent Stem Cell Initiative (HipSci, Box B.2.5), which aims to generate iPS cells from over 500 healthy individuals and 500 individuals with genetic disease to discover how genomic variation impacts cellular phenotype. Information obtained from the study of these iPSC lines will be computationally integrated with genomic data and patient records in various settings to elucidate the connection between genotype and phenotype for specific disease states. EMBL scientists will play a key role in the statistical integration of genomic, epigenetic and cellular imaging data.

As will be evident from the above, EMBL is involved in several initiatives whose goal is to provide support to medical and clinical researchers and improve the understanding of human disease. This change in EMBL's focus is accelerating, driven by advances in technology, which enable EMBL researchers to make real contributions to the understanding of multiple disease states by applying their expertise in genomics, in imaging and in advanced data analysis, to human systems. This trend will expand during the next Programme period

Box B.2.4: Pan-cancer analysis of whole genomes

Cancer research is one area that will continue to benefit tremendously from the new possibilities provided by large-scale sequencing of patients and from the integration of genome data with transcriptomic and epigenomic datasets as well as environmental and clinical data. The participation of cancer patients in cohort studies is high and increasing numbers of cancer genomes and tissues are being analyzed for research, diagnostics and clinical care.

In the next Programme, EMBL scientists will play leading roles in pursuing integrative analyses of cancer genomes and associated data types to shed new light on tumor biology. Within the recently initiated Pan-Cancer Analysis of Whole Genomes (PCAWG) project, tumor and blood samples from 2,500 cancer patients determined by whole genome sequencing, and often transcriptome and DNA methylome profiling data, have begun to be mined to uncover tumor biological principles both specific to individual cancers and across over 20 different cancer types. Within the coming indicative scheme, cancer genomic analyses will be scaled to tens and hundreds of thousands of cancer genomes. The resulting genomic data will be integrated with environmental and comprehensive clinical data, including histological and radiological images. A major focus will be to enable standardized integrative analyses of data generated in diverse contexts, including basic cancer research studies and clinical studies, where sharing of raw patient genomic data is complicated or impossible. Within PCAWG, EMBL scientists will be aiming to achieve data harmonization by devising standardized computational approaches to facilitate joint integrative analyses of cancer genomic data available via either open or restricted access mechanisms. PCAWG will generate strong incentives for participation by offering data submitting institutions prioritized access to large harmonized pre-processed datasets disseminated through EMBL-EBI.

By pursuing integrative analyses of these harmonized datasets, EMBL scientists will obtain novel insights into basic and disease human biology, for example, by uncovering the full landscape of cancer 'driver' genetic alterations (*i.e.*, DNA alterations promoting tumor growth) including point mutations in intergenic regions and DNA rearrangements mediating 'enhancer hijacking'. EMBL researchers will aim to identify important biological markers for disease states, and uncover commonalities and differences in tumor evolutionary patterns among cancer types. Data integration will further enable EMBL scientists to develop hypotheses on causal molecular relationships relevant to tumor biology, for example influences of the germline genome and environmental factors on cancer development, which can subsequently be transferred back into an experimental model system for more detailed mechanistic dissection. Furthermore, new findings made through these efforts may help pinpoint promising new intervention points for treatment, fostering interactions between basic research and clinical translation.

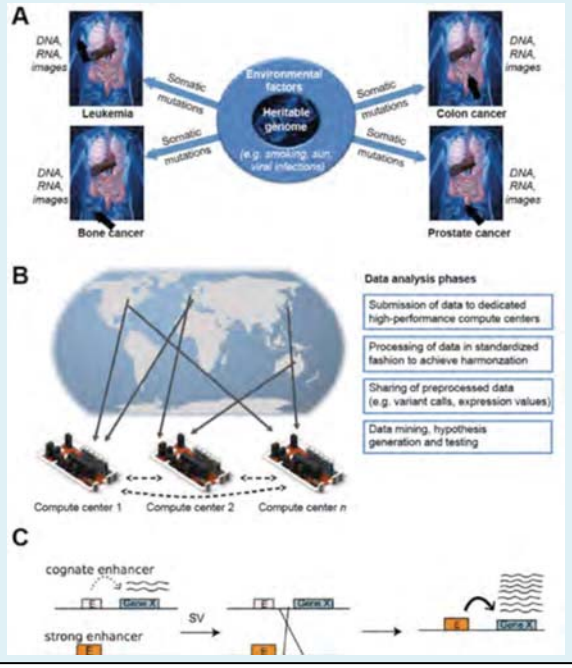


Figure B.2.4 Pan-cancer analyses. A. Integration of genomic, environmental and clinical data for numerous tumour entities. B. Approach taken to harmonize cancer genomic data generated in different institutions and countries. **C.** Pan-cancer analyses will provide novel insights into basic and disease human biology including DNA rearrangements affecting the *cis*-regulatory landscape by mediating 'enhancer hijacking'.

Box B.2.5: Genotype to phenotype using human induced pluripotent stem cells

The UK Human Induced Pluripotent Stem cell Initiative (HipSci) is a flagship project that seeks to create a human induced pluripotent stem cell (iPS) reference panel of healthy normal individuals and a selection of rare disease samples. The initiative is unique in that it brings together leading expertise in iPS technology, genomics, proteomics and cell biology. HipSci has already demonstrated the feasibility of large-scale iPS line generation, delivering more than 100 lines derived from genetically diverse donors.

The production of HipSci lines will soon reach the final goal of generating 500 normal iPS lines and a similar number of rare disease samples. In doing so, HipSci will create one of the first poly-omic variation datasets, combining genotype data, epigenetic state, gene expression levels, as well as quantitative proteomics and cellular traits in a large sample. This unique resource of banked iPS cells together with deep molecular and cellular phenotype data will stimulate and drive future developments in several strategic areas at EMBL.

First, to fully exploit the data being generated, new statistical and computational tools to integrate large-scale molecular variation data across molecular layers will be needed. Through their involvement in HipSci, EMBL research groups will be in an excellent position to pioneer the development of this technology. By combining multivariate association genetics with principles from causal reasoning, we will be able to trace the effect of individual genetic and epigenetic variants across the levels of transcription and translation to manifest in cellular phenotypes.

Second, by using the lines being generated, the HipSci project will stimulate important applications of emerging technologies, including single cell genomics. By leveraging the recently established EMBL-EBI/Sanger single-cell genomics center, we are planning to generate single-cell transcriptome data from thousands of differentiating iPS cells from diverse genetic backgrounds. By tying together the single-cell readout with genotype information and other HipSci assays, we will be able to derive new insights into how genetic variation shapes the transcriptional state across endoderm differentiation, one of the most fundamental developmental processes. In the future, the correlation with imaging based assays, is likely to add an additional very valuable phenotypic information layer.

Finally, the generation of iPS lines from rare disease samples in particular will provide unprecedented opportunities to study human disease. The HipSci resource will form the basis of several application-driven projects, differentiating iPS cells into disease-relevant tissue types or organoids. Moreover, by comparing iPS-derived tissues to primary cells from the same samples, we will be able to objectively assess the accuracy of iPS models for regenerative medicine.

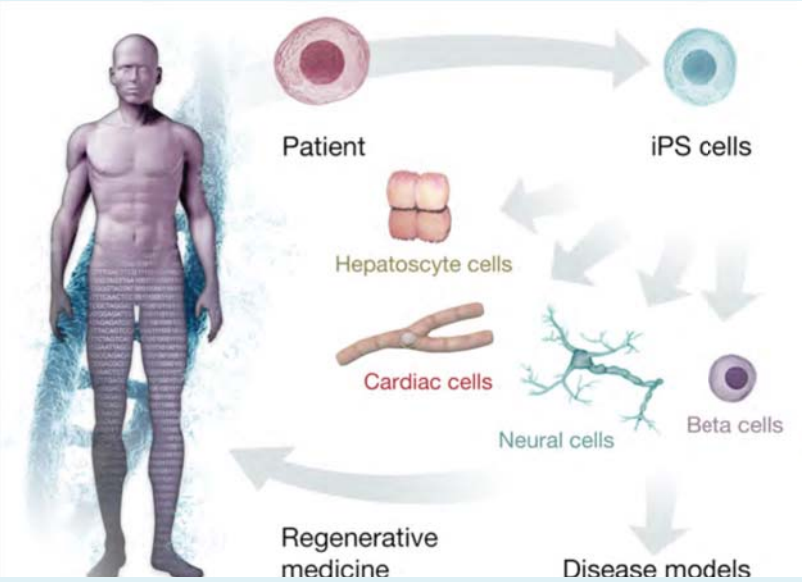


Figure B.2.5 Induced pluripotent stem cells (iPSCs) are an important resource to derive disease models and to investigate cell differentiation. The UK human induced pluripotent stem cell initiative (HipSci) will establish one of the largest collections of iPS cells from genetically diverse donors. The involvement of EMBL in HipSci will stimulate and drive a number of key future research aims in the area of integrative genomics, stem cell biology and regenerative medicine.

2.4. Molecular Processes in Space and Time

In the three preceding chapters we have discussed how to identify the molecular components required for specific biological functions. Through genomic, genetic, bioinformatic and biochemical analysis, we have begun to understand how many of these components interact to form regulatory and functional networks. The resulting networks are very useful maps of biological function, but they are essentially two-dimensional, mostly still lacking information on where in space and when in time the activities they participate in take place. The remaining critical dimensions of biological systems, spatial, multicellular and temporal organisation, can be derived from methods that visualise biological function, an area of particularly rapid current technology development, and from organismal-level studies. This chapter is devoted to EMBL's plans in biological imaging across multiple scales and in organismal biology.

2.4.1 From Molecules to Complexes - Structural Biology

One goal of molecular biology is to understand biological processes in terms of the underlying chemical reactions and molecular interactions. Biological molecules such as proteins, nucleic acids, lipids, carbohydrates and other metabolites assemble into complex molecular structures that carry out functions that go far beyond the properties of the individual components. Structural biology elucidates how these components assemble into biologically relevant complex structures and visualises their dynamics, three-dimensional architectures and interactions in space and time, thereby providing molecular and chemical insight into their biological functions and how these are carried out.

Studying multidimensional complexity at the molecular level

Understanding the functions of macromolecular assemblies must take into account multidimensional complexity, including variations in composition, conformation, post-translational modification and interaction with other cellular components, including ligands and metabolites. Dynamic transitions between these different states are functionally relevant. The interactions and dynamics displayed by a given assembly depend on cellular context. Studying multidimensional complexes both *in vitro* and *in situ* therefore requires the broad integration of various structural biology techniques (nuclear magnetic resonance (NMR), X-ray crystallography, single-particle electron microscopy (EM), electron tomography, small angle X-ray scattering (SAXS), etc.) across different scales of resolution. Complementary approaches such as proteomics, single-molecule studies, functional light microscopy, biophysical approaches and chemical biology provide additional insights and must be integrated to provide a comprehensive description of complex biological systems. With its cutting-edge instrumentation and infrastructure in structural biology and imaging, its broad expertise in chemical biology and proteomics, and its culture of interdisciplinary collaboration, EMBL is perfectly placed to meet this challenge. The excellence of the Laboratory in this area is documented by progress in structural biology in recent years (Section B.1.1.1).

Three examples can be chosen to illustrate EMBL's excellence and impact. The first, is the structure–function analyses of the nuclear pore complex, which is the largest non-polymeric protein complex in cells. Combining single-particle EM, electron tomography, super resolution light microscopy, quantitative and chemical crosslinking proteomics as

well as advanced protein labelling strategies, EMBL scientists from several Units have obtained detailed and complementary information about the nuclear pore scaffold. This project lays the groundwork on which future efforts will build (Box B.2.7). The second is the use of cutting-edge protein expression technology in insect cells combined with X-ray crystallography for the structural analysis of the complete influenza RNA polymerase, and the third the combination of single-particle EM, X-ray crystallography and chemical crosslinking/mass spectrometry for the analysis of eukaryotic RNA polymerases I and III and their pre-initiation complexes.

The coming years will see further integration of different experimental approaches, but also an increased ambition to unravel multidimensional complexity by better monitoring compositional and conformational variations in different cell stages and cell types. Projects that will be tackled by groups at EMBL in the coming years include: the functional roles of the epigenetic chromatin states of regulatory protein complexes and of long non-coding RNAs in gene regulation using combinations of structural, proteomic, genomic and genetic approaches in order to make further links between structure and function in gene expression; further work on different conformational and functional states of viral RNA polymerases with a view to understanding these unusual polymerases better and guiding the development of antivirals; the role of biological membranes in the organisation of many cellular processes such as endocytosis (Box B.2.6), viral entry and viral budding as a contribution to the still difficult areas of the structure of membrane proteins and membrane shaping; and the structural, biophysical and compositional characterisation of, for example, the RNA-containing granules found in poorly understood cellular structures involved in cell differentiation such as P-bodies, germplasm and the nuage of *Drosophila melanogaster* germline cells as examples of organised structures formed by components that, on their own, are largely unstructured.

The ‘resolution revolution’ in electron microscopy and its implications for structural cell biology

Electron microscopy has a key role in forming the connection between molecular information obtained *in vitro* by X-ray crystallography or NMR with information about cellular structures and dynamics obtained *in situ* by light microscopy techniques. Recent hardware and software developments, including stable high-end microscopes and direct detectors, now allow single-particle EM and electron tomography to generate reconstructions at much higher resolution and to resolve smaller and more flexible complexes than was previously possible. Single-particle EM reconstructions can now reach below 3 Ångstrom (Å) resolution, allowing the technology to be used for the *ab initio* structure determination of both large and small protein complexes, similar to X-ray crystallography but requiring less sample. It is also possible in some cases to solve the structures of multiple conformations from a single sample thereby providing information on dynamics related to function. The resolution limit for structures obtained by averaging electron tomography data (subtomogram averaging) is now at about 8 Å, and this technique can be applied not only to purified samples but also to heterogeneous, complex systems including intact cells, enabling the structures to be seen within their functional context. These advances – sometimes referred to as the ‘resolution revolution’ – will have a profound impact on structural and cell biology. The uses of X-ray crystallography and NMR will also be affected by these developments as they provide complementary high-resolution and/or dynamic information about individual subunits and sub-complexes that can be more precisely fitted into the improved EM reconstructions. As a result, obtaining the detailed structures of large complexes in combination with different ligands and interactors or in different conformational states will no longer be limited to a few model systems (e.g. ribosomes, RNA polymerases, exosomes, proteasomes etc.), but instead will become possible for many multiprotein complexes. At

the same time, higher-resolution electron tomography will allow faster and more reliable positioning of protein complexes and their higher-order assemblies within a cellular context. EMBL is very well positioned to contribute to (as outlined in Section C.2.3) and take full advantage of these exciting developments. At EMBL Heidelberg, electron microscopes are equipped with direct detectors for single-particle EM and electron tomography, whereas EMBL Hamburg and EMBL Grenoble have access to high-end electron microscopes through their local partnerships and strategic alliances, the Centre for Structural Systems Biology (CSSB, Section F.1.1.1) in Hamburg and the Partnership for Structural Biology (PSB, Section F.1.1.4) in Grenoble. Recent electron tomography reconstructions at EMBL have already provided important structural insight into human immunodeficiency virus (HIV) assembly, the irregular structure of coat protein 1 (COPI) vesicles and the nuclear pore scaffold (Sections B.1.1.1.3 and B.1.1.1.4).

These developments, together with advances in correlative imaging will allow us to move closer to the long-term goal of understanding and visualising multiple cellular processes at the molecular level. Fundamental to this goal is the ability to model across different resolution scales as well as to integrate very different types of complementary data. The new EMBL Centre for Integrative Structural Modelling (Section B.3.1.5) will be instrumental in achieving these aims.

New opportunities in structural biology with the X-ray Free Electron Laser

Four decades ago, synchrotron radiation began to revolutionise structural biology. Since then, the use of high-brilliance synchrotron radiation has become an essential workhorse in obtaining the high-resolution structures of thousands of proteins, including large integral membrane proteins and a number of large protein complexes, that have collectively had enormous impact in the life sciences. Other important applications in structural biology are X-ray scattering analyses, a range of spectroscopic methods and X-ray imaging, which is a potentially powerful emerging field of structural research. The two EMBL Units in Hamburg und Grenoble are intimately linked to the on-site synchrotron radiation infrastructures provided by the German Electron Synchrotron (DESY) and the European Synchrotron Radiation Facility (ESRF), respectively and have a strong complementary record in developing innovative instruments and approaches (Section E.1.1.1), thereby implementing leading research services for the member state research communities (Section C.2).

During the past decade a new type of X-ray radiation generated by the Free Electron Laser (FEL) has become available. This new technology has a number of key advantages: FEL-based X-rays have peak intensities orders of magnitude higher than the brightest current synchrotron radiation sources; the laser emits X-rays in femtosecond (fsec) pulses with a time structure that is completely different from all other X-ray sources; and it displays full radiation coherence. At present, two X-ray laser infrastructures are in operation – the Linac Coherent Light Source (LCLS) facility in Stanford, USA and the SACLA facility in Japan. Because of the huge demand and limited availability, however, it is very difficult for users to access these facilities. The new European XFEL in Hamburg, Germany, is expected to open in 2017 and will be by far the most powerful source of this type. Unlike the other existing facilities, the European XFEL is unique in using superconducting technology and will produce 27,000 light flashes per second, compared to a maximum of 120 at the current facilities. This will markedly reduce the quantity of sample required for analysis and will also provide completely new possibilities for studying time-dependent structural dynamics. Pioneering research at the X-ray lasers in recent years has shown that it is possible to perform completely novel experiments that make use of the scattering properties of many types of biological material, ranging from crystalline diffraction to single particle experiments.

The possibilities and limitations of the use of X-ray lasers in biology are however still largely unexplored. Amongst many other applications, this approach allows previously unreached time resolution, in the fsec regime, and the structure of sub-mm samples to be elucidated. In parallel, during this early phase, there has been a growing wave of research in and development of novel instrumentation, methods of sample transfer and data processing, all targeted at overcoming various technical hurdles. In many respects these developments parallel those that took place several decades ago when synchrotron radiation first became available. The next Indicative Scheme period will correspond to the first wave of developments and exploration of scientific possibilities at the XFEL in Hamburg. Based on its past excellence and major role as a service provider for life-scientists, EMBL is now preparing to be part of the cutting-edge XFEL based activities in both research and service provision to the life science community.

The EMBL Unit in Hamburg, as a provider of three state-of-the-art synchrotron beamlines at PETRA III, is in a unique position to become engaged in future structural biology applications and provision of research infrastructures using both synchrotron radiation and FEL radiation. Proof-of-principle experiments by scientists at EMBL and CFEL/DESY have demonstrated that these approaches hold considerable future potential. Exploiting these opportunities will allow EMBL to expand its structural biology portfolio through the structural analysis of biological processes in ways that were not previously possible (e.g. following viruses in real time as they enter their host cells or assemble at the cell membrane during the budding process).

Structural systems biology and networks

Structural biology at EMBL comprises detailed mechanistic as well as broad systemic approaches. EMBL has a long tradition in characterising protein–protein interactions in molecular networks using a combination of structural and biophysical techniques, which provides a quantitative understanding of molecular networks *in vivo* and *in vitro*. In the next EMBL Programme, extending these activities towards the charting of interaction between proteins and metabolites or small molecules, and analysing both the structural aspects and the biological consequences of these interactions, will be a high priority. Pioneering studies conducted at EMBL in the area of protein–lipid binding will be extended to other classes of metabolites. Furthermore, we will explore how post-translational protein modifications alter protein networks and their interactions with small molecule metabolites. Finally, broad chemical screens will be used to understand in detail which molecules in the network are specifically targeted by a given metabolite or chemical and how these interactions affect the properties of the network *in vivo*. To quantitatively understand and model molecular networks, we will combine different mass spectrometry (MS) techniques such as quantitative MS, native MS, and targeted and untargeted metabolomics, which will become available at EMBL through the establishment of a Metabolomics Core Facility (Section C.3), with structural biology and complementary biophysical techniques (e.g. isothermal titration calorimetry (ITC), thermophoresis, surface plasmon resonance). In these activities, we will exploit the possibilities of miniaturisation using, for example, microfluidics devices. The feasibility of this overall approach has been demonstrated in previous work carried out at EMBL with model systems such as the minimal bacterium *Mycoplasma pneumonia,* the yeast *Saccharomyces cerevisiae* and the thermophilic fungus *Chaetomium thermophilum.* In the new EMBL Programme, we anticipate that these approaches will be rolled out to study more complex systems such as healthy and malignant human cell lines, the pathogen *Mycobacterium tuberculosis,* interactions underlying the formation and stability of microbial communities, host–pathogen interactions involving bacteria and viruses, all the way to human patient samples such as pathological cancer specimens or gut microbiome interactions with the host.

Box B.2.6 An integrative approach to endocytosis

To ingest molecules from their environment and from their own surface, eukaryotic cells use clathrin-mediated endocytosis, in which cargo molecules to be imported are packaged into small membrane vesicles formed from the cell's membrane. Cargoes include not only essential nutrients and signaling molecules, but also viruses that exploit endocytosis to infect eukaryotic cells. The basic concept of endocytosis was described five decades ago, but we still have a very limited understanding of the molecular mechanisms by which cell selects cargo molecules and form the endocytic vesicle.

The dynamic nature and the complexity of the endocytic process make it very challenging to assess the underlying molecular mechanisms. A mature endosome has a size of about 200 nm - just below the diffraction limit of visible light and is composed of over 50 different proteins - that assemble and disassemble in a highly dynamic fashion on the cell membrane, which deforms within a few tens of seconds as a consequence of the endocytic process.

To overcome these challenges several EMBL groups have teamed up to create interdisciplinary approaches. Scientists from the Cell Biology and Biophysics and Structural and Computational Biology Units in Heidelberg have developed a correlative light and electron microscopy method (Section E.1.1.2.2) to combine the advantages of the high temporal resolution of light microscopy, which provides detailed information about the locations and the assembly dynamics of endocytic vesicles, with the high spatial resolution of electron microscopy, which reveals the shape changes of the endocytic membrane (Fig B.2.6 A,B). In future work, superresolution light microscopy will be used to visualise the detailed organisation of different endocytic proteins (Fig B.2.6 C). Together, these methods will allow of the dynamic architecture of the endocytic protein machinery to be visualised with high spatio-temporal resolution in the context of ultrastructural membrane shape changes.

For a more detailed analysis still, the endocytic machinery must be taken out of the cellular context. Structural biologists from Hamburg and Heidelberg aim to use structural electron microscopy and crystallography to reveal the structure of isolated parts of the endocytic machinery *in vitro* (Fig B.2.6 D). This information can then be combined with the overall view obtained by *in situ* analysis. To understand how structure and function of the endocytic components are related, these methods will be applied to normal cells and cells in which specific activities are perturbed either genetically or chemically. Modelling groups will integrate the structural and functional data obtained to build mathematical models with the aim of describing the vesicle budding process quantitatively (Fig B.2.6 E). The integration of all these approaches will produce the first complete dynamic and structural molecular model of the endocytic process.

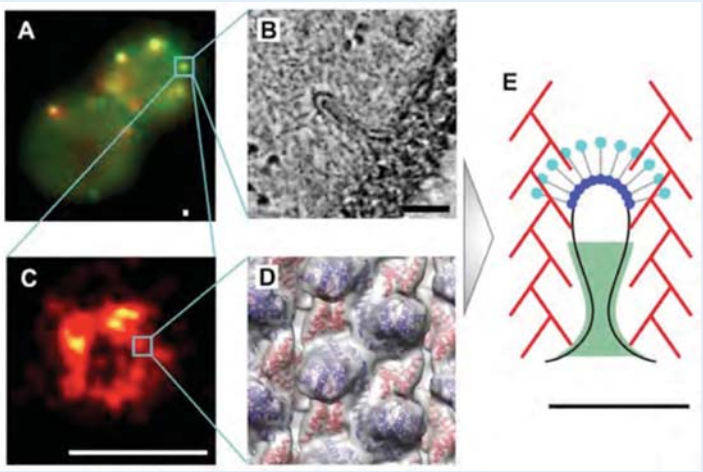


Figure B.2.6 An illustration of the proposed integrative approach to endocytosis.
A) Live cell imaging reveals the assembly order of the endocytic machinery. B) Correlative light and electron microscopy determines the membrane shape changes at endocytic sites. C) Superresolution light microscopy maps the distribution of different proteins at endocytic sites. D) X-ray crystallography determines endocytic protein assemblies at atomic resolution. E) Computer modeling is used to integrate the multiscale spatio-temporal data obtained by different approaches. A-E) Scale bar, 100 nm.

Box B.2.7 Towards a complete structural model of the largest protein complex: the nuclear pore

Nuclear pore complexes (NPCs) are the largest known non-polymeric protein complexes. About 1000 proteins (nucleoporins) per complex span the inner and outer nuclear membranes to form a central transport channel. The composition of NPCs varies between organisms, cell types and most likely also within single cells. Formerly seen as static structures, NPCs are nowadays appreciated as multifunctional ensembles that exist in numerous structural states and are involved in various cellular processes besides nuclear transport. These 'moonlighting' functions of the NPC include DNA repair, transcriptional regulation and translational control during cellular differentiation, reprogramming, and malignant transformation. To understand how compositional and conformational variability as well post-translational modification patterns change nuclear pore function depending on cellular needs is a challenging goal for cellular and structural biology in the coming decade.

Tackling this challenge requires the integration of various experimental approaches, which have however already been implemented across EMBL for other purposes. Super-resolution fluorescence microscopy and fluorophore counting allow the exact position and quantity of the individual components that constitute native nuclear pores to be measured (Fig B.2.7 A). Electron tomography can generate moderately resolved global maps of NPCs that serve as an overall framework for structural modeling (Fig B.2.7 B). Mass spectrometric measurements together with chemical cross-linking and internal peptide standards reveal protein interfaces and stoichiometries (Fig B.2.7 C). The power of this integrative approach has been demonstrated by determining the position of one major scaffolding component already identified by structural biology (Fig B.2.7 D,E). Such integrated efforts undertaken by multiple EMBL groups will be increased to achieve a complete picture of nuclear pore architecture, to ultimately generate pseudo-atomic models of the entire NPC in different functional states.

By combining quantitative imaging, proteomics and gene expression analysis EMBL groups are aiming to chart a dynamic map of nuclear pore architecture. This includes the variability of the NPC and the incorporation of these data into otherwise static structural models will be instrumental in functionally understanding spatiotemporal dynamics. To achieve this goal, advanced protein labeling strategies, correlative and superresolution imaging as well as proteomics approaches will become increasingly important. Ultimately, a complete structural model of the NPC will lead to a molecular understanding of the many essential functions of the complex.

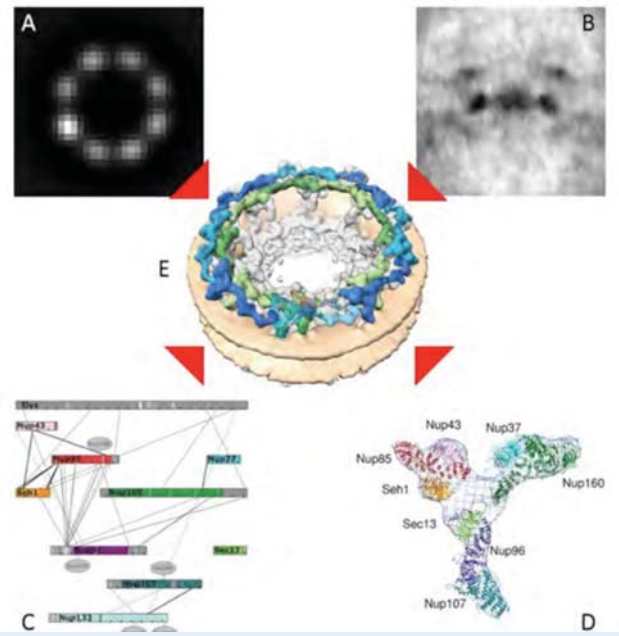


Figure B.2.7 Towards a complete atomic model of the nuclear pore complex (NPC) by integrative structural biology. A) Super-resolution light microscopy of individual NPC protein components. B) Averaged side view of a mature NPC obtained by electron tomography. C) Cross-links between different NPC protein components determined by mass spectometry. D) Major scaffolding components of an NPC complex analysed by single-particle electron microscopy and X-ray crystallography. E) Integrating data from different techniques converge in a structural model of the NPC.

2.4.2 From complexes to cells

Mapping the cell: Visualising the molecular processes of life inside the cell in time and space

Digital biology has moved on from the genome to single proteins and protein complexes and is beginning to be applied to the cellular level of biological organisation. At the scale of the cell – the smallest autonomous unit of life – we are now in a position to seriously tackle the molecular complexity and dynamic properties of the central processes of life. Key to this is unlocking the series of molecular processes that work together over time to lead to higher-order structural and functional properties. What appears 'magic' at a higher level, such as the beautiful and symmetrical shape of a dividing cell, can be mechanistically understood in terms of its underlying molecular self-organising principles. We cannot understand biology without understanding the dynamics of molecular processes and their interplay. Driven largely by the rapidly increasing power of fluorescence labelling and imaging technologies, we can directly visualise any protein-based component within the cell and determine its biophysical and biochemical parameters in time and space with ever-increasing resolution. By integrating such biophysical and biochemical data on all members of a protein complex, and the networks linking the action of many complexes, we can start to 'visualise' the molecular machinery of life in action and directly watch molecular mechanisms unfolding.

In the foreseeable future, including the period from 2017–2021, the main challenges will be to continue to push our analytical methods towards higher precision and robustness, and to systematically map out the molecular machinery of life rapidly and comprehensively at the cellular level (Section E.1.1.2). At the same time, we will need to develop both the computational tools to integrate and archive the resulting data and new theoretical approaches to enable systematic mechanistic understanding. We will also need to be able to represent this rapidly accumulating mechanistic understanding in large computational models that are amenable to refinement based on newly acquired data (Sections B.2.5 & E.1.2.2.3). Even though working at the cellular level is already a huge opportunity and challenge, we are at the same time pioneering this comprehensive molecular mechanistic approach to multicellular functions in the context of tissues (Section B.2.5) and organisms (Section B.2.4.3). This section provides examples of where we stand today in our ability to visualise molecular processes in space and time and our aims for the next period.

Visualising molecular machines in action in their functional context

New imaging technologies at the interface of structural and cell biology are rapidly revolutionising our ability to identify and map the proteins required for specific cellular functions in space and time at the nanometer scale (Section E.1.1.2). EMBL excels in the interdisciplinary integration of different imaging approaches. Examples are high-resolution and large-volume electron microscopy (EM), in which EMBL has recently implemented both serial EM tomography as well as focused ion beam scanning EM (FIB-SEM) technologies. These are seamlessly linked, using correlative methods developed at EMBL, to quantitative fluorescence imaging modalities, including live cell and single-molecule imaging as well as super-resolution microscopy implemented, for example, as gated stimulated emission depleted (STED) microscopy or ground state depleted (GSD)/dSTORM localisation microscopy. The computational integration of the resulting data using mathematical modelling (Section E.1.2.2.3) has allowed us to reach unprecedented structural and/or functional understanding of key supramolecular machines in cells.

One of the first breakthrough successes of this integrative approach is the nuclear pore complex (NPC), which is the largest non-polymeric protein complex in cells and for which we can now measure the stoichiometry of its subunits, their relative position at nanometer scale and even start to assign their atomic structures inside cells within a giant assembly of several hundred proteins. In the next period, a complete atomic level structural model will be within reach (Box B.2.7). Obtaining this model in different functional states by correlation with dynamic imaging will reveal the structural basis and molecular mechanism of selective nucleocytoplasmic transport. A second example of success is endocytosis, which is the essential process of cellular nutrient uptake that involves the transient assembly of a complex of over 50 proteins on the plasma membrane. The formation of this complex initiates the membrane invagination that gives rise to a coated vesicle inside the cell. Here, dynamic supramolecular assembly and function are one and the same, and structural analysis is only mechanistically informative when performed in the context of a precisely defined functional state. The integration of electron and light microscopy has allowed us to understand how the assembly order of the individual proteins drives the sequence of membrane-shaping events. In the future, a molecular scale mechanistic model that resolves the interplay between membrane-shaping and force-producing protein modules, as well as the increasing integration of atomic level structural detail, will reveal the molecular mechanics of endocytic vesicle formation (Box B.2.6). Applying the same strategy of time-resolved structural and functional analysis to the NPC, for which complex formation precedes function, will allow us to reveal the mechanism of self-assembly of even the largest cellular machines.

Activities in molecular and cellular imaging will also increasingly be integrated with proteomic and genomic studies. For example, as discovered at EMBL during the current Indicative Scheme, the composition of the nuclear pore varies between different cell states. These changes can be monitored by genomics and proteomics but their structural and functional consequences can only be observed via high-resolution imaging (Box B.2.7). Similarly, eukaryotic transcription will be followed *in situ* by combining detailed *in vitro* information about the structure and dynamics of the transcription complexes involved (RNA polymerases, general transcription factors, chromatin remodellers) with high-resolution imaging of chromatin, proteomics, chromatin immunoprecipitation sequencing (CHIP-seq) and RNA sequencing (RNA-seq) data.

Progress made in the endocytosis and NPC projects have been pioneering achievements. A key feature of projects at EMBL is that any technological and computational solutions are turned into general strategies via EMBL's scientific services and facilities. This allows scientists, in EMBL and elsewhere, to apply the same principles to other large complexes and cellular machines. We will establish robust protocols and ensure data compatability in order to build up an inventory of mechanistic understanding for the major supramolecular machines of the cell. Future research at EMBL will thus be able to take on new major challenges such as unravelling the structural and functional basis of the largest and most important polymer: the organisation in time and space of DNA in chromatin in the nucleus of single cells (Box B.2.3).

Imaging the non-protein molecular machinery

Biological systems not only comprise protein-based macromolecules and complexes, but also nucleic acids, sugars and lipids as well as a host of smaller metabolites. Molecular machines are often hybrid in composition – for example, RNA and protein or lipid and protein – and to fully understand them also requires an analysis of the non-protein partners. Although nucleic-acid imaging with fluorescent stem-loop-binding

proteins and fluorophore-binding RNA-aptamer sequences has been successful, we are still limited in our ability to investigate lipids and are virtually blind to the sugar world inside cells when it comes to visualising and mapping defined molecular species. Significant progress has been made at EMBL in the development of new chemical biology approaches to couple lipids to small fluorescent dyes or reactive groups *in vivo*. The dynamic visualisation of lipids and protein–lipid interactions has already provided tantalising new insights into membrane biology and lipid signalling. In the future, we will need strategies to ensure the functionality of modified lipids and to incorporate their analysis much more systematically into our quest to map the molecular landscape of the cell.

Reconstructing dynamic molecular networks

Supramolecular machines do not act in isolation but are regulated by and are part of molecular networks that orchestrate the essential functions of life via the coordinated action of several machines in space and time. The most basic cellular processes such as division, growth, shape changes, movement and differentiation often require hundreds of proteins and other macromolecular components. For each such network, understanding how the right sequence of events (or information flow) is ensured and how the correct supramolecular structure is formed at the right time and in the right place to generate higher-order emerging properties requires us to reconstruct the network and its dynamic changes over time. Such dynamic spatial proteomics seemed impossible for many years. However, recent technological breakthroughs at EMBL in automating single molecule-based concentration and interaction measurements by fluorescence correlation spectroscopy (FCS) combined with the establishment of fully functional fluorescence tagging of proteins by homozygous genome editing now enable us to reconstruct the first dynamic networks in single human cells. As genetically encoded fluorophores can currently distinguish only a few molecular species at the same time, we first measure all network components sequentially in individual living cells during the process of interest in reference to the same spatial and temporal landmarks and then integrate all the information back into a computer model based on these landmarks (Figure A). In parallel, we are working on generating additional chemical tags so that more components can be assayed simultaneously (Section E.1.1.2.1)

A challenge in creating models of dynamic biological processes is that the spatial and temporal reference system is constantly changed by the coordinated action of the dynamic molecular networks inside it. A dividing cell, for example, completely reorganises its structure before genome segregation owing to the action of its mitotic protein networks. The kinetics of these molecular actions in time are non-linear. An appropriate dynamic computer model thus has to be based on the biological space and time (i.e. four-dimensional (4D) kinetics) of the process of interest. The 4D protein network can be inferred by comparing the biophysical and biochemical properties of each protein in the network, including for example their absolute local copy number, spatial distribution changes and mobility. Connections between nodes of the predicted network can then be directly probed by single molecule measurements of binary interactions, again in a space and time-resolved manner, which allows us to watch complexes form and fall apart again like a precisely choreographed molecular ballet. As a pilot cellular process, we will focus on the assembly of the complete network involved in cell division in a human cell. The experimental and computational tools that need to be developed will not only be technological breakthroughs but will also provide generalisable strategies for carrying out such dynamic network studies for any cellular function of interest.

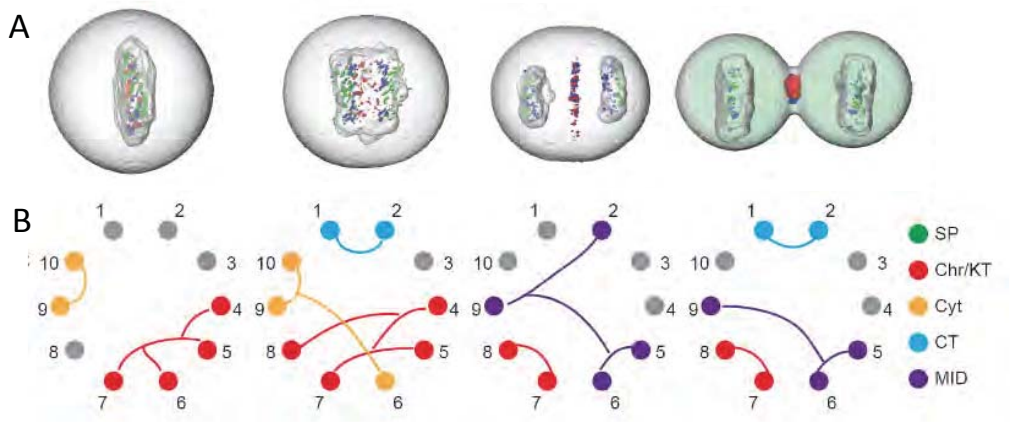


Figure A Computer model of an average dividing human cell and illustration of an inferred set of dynamically changing protein–protein interactions. (A) Proteins of interest (POI) are labelled in colour, and spatiotemporal reference landmarks (i.e. the surface of the genome and the cell) in grey. (B) Localisations of 10 POIs (1-10) on landmark structures (spindle (SP, green), chromosome/kinetochores (Chr/KT, red), Cytoplasm (Cyt, yellow), centrosome (CT, light blue), midbody (MID, purple)) were registered over the course of the cell cycle. Colocalisations of POIs on the same structures are depicted and highlighted in the same colour and through coloured lines.

The canonical human cell - a generative computer model to integrate, mine and navigate molecular information in space and time

As outlined above, reconstructing dynamic protein networks requires computer models of the network components in space and time. Initially, such models will rely on fluorescence-based spatiotemporal landmarks, but will increasingly incorporate coordinates from higher resolution cell-mapping approaches such as FIB-SEM and super-resolution microscopy of libraries of functional fluorescently 'barcoded' proteins. These methods are currently being developed at EMBL to reach the throughput and robustness required to sample cell-to-cell heterogeneity and different functional states (Section E.1.1.2) to allow us to build atlases of cells at molecular resolution. Such models, once built for a prototypical cell type and cellular process of interest, can be extended by any new molecular measurement that contains the same reference information so that it may be mapped into the model. We will spearhead this approach with a model of the human cell based on quantitative experimental data throughout the cell cycle, which represents our growing state of knowledge about the 'canonical human cell'. This canonical cell model integrates all available quantitative molecular and dynamic information in space and time in a standardised and parameterised form. This makes it invaluable for mining the data for network construction and inference of molecular mechanism. All molecular information in it is linked to the relevant biomolecular databases (e.g. the ENSEMBL genome) and provides the spatial and temporal link to connect the genome sequence with cellular phenotype. Furthermore, the standardised geometry and time of the canonical human cell will allow the generation of a visual output of any observed or predicted state (Figure B). To be able to navigate the available data for any set of proteins of interest is incredibly powerful and we will make the data available to allow creative and intuitive hypothesis generation. At the same time, it will allow the development of machine learning algorithms to suggest hypotheses automatically and classify different types of molecular and regulatory mechanisms in order to distill the common principles underlying the dynamic action of molecular networks.

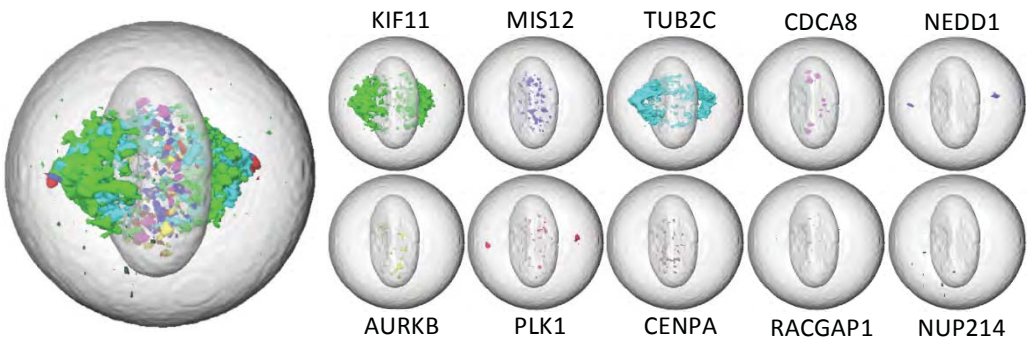


Figure B Computer-generated model of the 'canonical human cell' in late metaphase of mitosis. The average metaphase cell was created from 100 cells, providing the spatial reference landmarks, i.e. the surface of the genome and the cell in grey. Localisation of 10 different proteins of interest, registered to a virtual coordinate system, is indicated in different colours.

2.4.3 Organismal biology: understanding the principles of development and homeostasis

In their natural context, in animals and plants, cells of multicellular organisms function not in isolation, but in interaction with other cells. The first two sections of this chapter outline how we intend to incorporate high-resolution structural analysis into its cellular context, thus linking the molecular and cellular dimensions. Cells however also have a larger context, either in single-cell communities, such as those formed by microbes or in organs, tissues and organisms in multicellular systems. The arrangement, environment and the interactions in which cells engage in these communities ultimately govern the development and homeostasis of organisms from the time of their birth, at fertilisation, until their death. This is why organismal biology – the study of the structure and function of biological processes at the level of the organism – is an important part of EMBL's scientific structure and provides the natural context for many of the research plans concerning molecules, networks, cells and tissues outlined in previous sections of this document. Key to understanding how organisms develop is the ability to visualise and perturb a process and analyse the effect at high spatial and temporal resolution through technologies that are often first developed in biochemical or cell biological contexts. In the next Programme, a major goal for organismal biology will be to exploit the new technology developments – including single-cell genomics, imaging, chemical biology and optogenetics (Sections B.2.2.1 & E.1.1.2) – in the context of whole organisms, to provide a deeper and increasingly quantitative understanding of both normal development and disease states.

Studying the molecular mechanism of cell state changes: from single cells via tissues towards organisms

The spatiotemporal origins of first fate choices in organismal development

As the fertilised egg cell – the single-cell embryo – cleaves and multiplies, the cell collective self-organises into a multicellular organism. Recent advances in live high-resolution microscopy at EMBL have revealed an unexpected fluctuation and stochasticity during cell-fate specification and fate transition in developing organisms. This feature is particularly prominent in cells with high plasticity and potency for multi-lineage differentiation, such as embryonic stem cells. Gaining a mechanistic understanding of such complex systems will require the ability to monitor dynamic processes at single-cell resolution and to integrate molecular, cellular and tissue-scale information.

To achieve an integrated view of the early steps in organismal development, an interdisciplinary effort by EMBL researchers from the Cell Biology and Biophysics and Developmental Biology Units aims to develop advanced microscopy for imaging organisms at high spatiotemporal resolution, automated image analysis tools, and a pipeline to integrate all information into a lineage map (Box B.2.8). Classical cell lineage information is highly variable across samples. In the new map, quantitative information of lineage-marker gene expression and other cellular parameters will be integrated, including cell position, geometry and volume. Furthermore, cell and tissue mechanics can now be measured in living embryos to monitor the change in cortical tension, cell contractility, viscosity, elasticity and cytoplasmic flow. The future systems-level lineage map will be used to reveal emerging patterns and extract principles hidden in the apparent (and real) complexity of organismal development. Once a reproducible 'node' is identified within the map, a specific cell can be isolated and, for example, single-cell

transcriptome, epigenome or proteome analyses performed to determine the molecular properties that underlie its function. Such comprehensive information will also provide the basis for modelling organogenesis, and will allow the prediction, testing and, eventually, the control of cell-fate transitions.

This integrative EMBL project will elucidate emerging properties of multicellular organisms at the cellular and molecular levels. It will be particularly interesting to examine the potential roles of noise, cell-state fluctuations and cell–cell heterogeneity in cell-fate transitions and the acquisition of robustness by the system. An interdisciplinary team is currently using a relatively simple organism – the mouse pre-implantation embryo comprising up to 64 cells developing in a closed environment – to design tools that will then be applied in other contexts, including later stages of embryonic development (Box B.2.8).

Transitions in cell state drive the assembly of tissues and organs during development, allow tissue repair after injury and, when cancerous, can lead to organ destruction. Although numerous genetics and genomics studies have identified a multitude of factors associated with changes in cell state, the underlying cellular mechanisms that control these transitions within tissues are less well described. Because of their robust development and optical accessibility under experimental conditions, cell-state changes during organogenesis are primarily studied in model systems such as fly and fish embryos, for example the development of the lateral line sensory organ in zebrafish. This approach uncovers mechanisms that cannot be studied in single cells. A recent breakthrough at EMBL revealed that the collective morphogenesis of a group of organ precursor cells encloses a common extracellular lumen in the process of polarisation from a mesenchymal to an endothelial state. This multicellular structure is crucial because the lumen allows the participating cells to concentrate a secreted molecular signal to sufficiently high levels to drive their differentiation as a group. This ‘locks in’ the decision to become an organ precursor from which fully differentiated sensory cells emerge. Here, single-cell molecular signalling and multicellular morphogenesis are linked to each other to generate a positive feed-forward loop, ensuring the progressive and coordinated development of a new organ.

Cellular mechanisms of organ formation and tissue mechanics: development of experimental and data analysis pipelines

The luminal signalling mechanism described above reveals that changes in multicellular architecture can play a central role in driving cell-state transitions during differentiation and organ formation. Such mechanisms allow dynamic changes in multicellular phenotype, such as shape and mechanics, to feed back on gene expression and drive stable cell-fate determination. We expect to discover more such novel regulatory linkages in the next Programme period through the combined set of tools available across EMBL.

A joint aim of EMBL groups across several Units for the next Indicative Scheme is to quantitatively define tissue architecture using standardised, dynamic readouts such as cell-polarity markers, and correlate these with the expression of genes known to define cell state, using quantitative reporters of gene expression *in vivo*. Potential roles for tissue architecture in the control of cell type-specific gene expression will be directly interrogated using acute, targeted perturbations including optogenetics and laser ablation strategies combined with modelling. Similar perturbation approaches will be used to investigate the role of candidate molecular regulators of these pathways. Together, we expect these experiments to provide further novel insights into how tissue architecture feeds back on gene expression during cell-state transitions. These *in vivo* studies will be complemented with *ex vivo* approaches that allow cell explants to be

manipulated using biophysical strategies (e.g. micro-fabricated substrates) and will provide an excellent interface with the new EMBL outstation in Spain (Section B.2.5).

Cell niches and organoids

A niche provides a specialised tissue environment that is advantageous to the survival and growth of particular cell types. The regulating cues provided by the niche may include both soluble molecules and direct physical interactions involving both cells and complex molecular structures. There are several niches that allow haematopoietic stem cells (HSC) to differentiate into various blood cell types and these provide well-known examples of niche function. Studies of the characteristics, structure and role of cellular niches, and how they change in space and time, are crucial for our understanding of cell heterogeneity, many differentiation processes, and homeostasis. Such efforts are therefore integral to several ongoing and future organismal studies at EMBL.

Organotypic organoid cultures are *ex vivo* experimental systems that aim to reflect, at least in part, complex *in vivo* morphologies and 3D structures. By contrast, 'neurosphere' or 'mammosphere' cultures have the more restricted ambition of facilitating cell growth and differentiation through suspension culture and 3D expansion, without necessarily fully recapitulating orientation and proper structure. These 3D *ex vivo* systems can have markedly improved differentiation capabilities, provide improved insights into cell–cell interactions and cell signalling and facilitate live-cell imaging, cell-fate tracing and the characterisation of cellular heterogeneity. As such, they are already being exploited at EMBL in several areas to study complex, multicellular phenotypes including cancer remission and drug resistance and are expected to form a major pillar of the organ and tissue studies to be pursued in the outstation in Spain (Section B.2.5).

Molecular dynamics: from signal oscillations to self-organisation

A central challenge in biology is to understand the role of timing and dynamic transitions in controlling cellular states, multicellular assemblies and, ultimately, organisms. What information is encoded at the level of molecular dynamics, such as oscillations? How are signalling oscillations synchronised in a multicellular context, leading to spatiotemporal order at the mesoscopic and organismal scale? Technical advances in real-time imaging enable researchers at EMBL to simultaneously observe molecular and morphological changes directly within the context of a developing organism. Major advances have already been made at EMBL to link the production and phase of such oscillations to their roles in organism segmentation, through gene-expression and signalling networks (Box B.2.9). We expect to progress further in establishing linkages between such processes and, for example, the production of metabolic oscillations in the next Programme.

Another hallmark of dynamic biological systems, which often consist of weakly interacting components, is the ability to self-organise at multiple scales. EMBL researchers are addressing this phenomenon at the molecular and cellular levels by analysing how qualitatively novel properties emerge during self-organisation. The developing embryo represents an ideal physiological context in which the role of timing and such dynamic transitions can be addressed in a complex physiologically relevant context (Box B.2.9).

Visualising, perturbing and quantifying molecular processes in living organisms

Perturbing dynamic processes with high precision in living organisms

Achieving a complete understanding of organismal development and homeostasis will require a quantitative and dynamic description of the endogenous molecules and

biochemical reactions, which in turn requires the ability to perturb endogenous processes with high spatial and temporal precision. The success of the fruit fly, zebrafish and the mouse as animal models has largely been due to the sophisticated and continuously evolving genetic tools and genome engineering that enable precise perturbations to be created and hypotheses to be tested. A particularly interesting new tool is optogenetics, which allows light-based activation or inhibition of biochemical reactions, or the specific relocalisation of molecules in individual cells of living organisms.

The ability to control protein activity in living organisms with high spatiotemporal precision (at cellular and sometimes subcellular resolution) is opening new frontiers for the investigation of developmental pathways and the cellular biology underlying complex processes such as morphogenetic movements. Recent advances in genome engineering allow very precise control of the activity of endogenous proteins (e.g. signalling receptors, key regulators of intracellular trafficking or cytoskeletal dynamics) during animal development. The combination of optogenetics and two-photon laser illumination will allow researchers to achieve a localised pattern of light activation or repression. Together with light-sheet microscopy, the development of which continues to be pioneered at EMBL (Section E.1.1.2.4), researchers will be able to observe the global organismal response to a localised perturbation of, for example, tissue mechanics or signalling cues. The powerful combination of optogenetics and genetics will open the door to synthetic morphogenesis: for instance, by activating gene-expression within a defined four-dimensional context in selected mutant backgrounds, EMBL researchers will be able to test the impact of pre-defined tissue geometries on morphogenetic output (Box B.2.10).

Imaging of intracellular machineries and biochemical activities in dynamic tissue contexts

How cellular organisation is coordinated to allow the emergence of tissue-level properties is a topic of fundamental importance. For example, recent work has shown that the mechanical properties of tissues are dependent on the formation of tensile ‘supracellular’ actin networks that span entire cells. Likewise, the functions of most epithelial organ systems depend on the coordinated reconfiguration of intracellular trafficking machinery to allow polarised cell transport. An understanding of such multicellular functions has been limited by a general inability to perform high-resolution analysis over broad spatiotemporal scales. Thanks to a number of recent technical developments, discussed in detail in Section E.1.1.2, researchers at EMBL are making progress towards this goal.

The entire array of visualisation methods being developed and used at EMBL will increasingly be applicable to developing systems. Correlative light and electron microscopy (CLEM) approaches established at EMBL (Section E.1.1.2.2) will provide an essential bridge between tissue and subcellular scales by allowing the ultrastructural analysis of subcellular processes in organs and embryos. This has already been established by recent successes include studies of the endocytic machinery in the *Drosophila* embryo and luminal signalling in the zebrafish embryo and will be further enabled by super-resolution microscopy methods.

A major obstacle to understanding how tissues emerge through self-organising molecular machineries has been the absence of methodologies for imaging and quantifying biochemical activities *in vivo*. EMBL teams have recently pioneered generic approaches that are robust and reliable enough to apply to whole tissues and organisms. Examples include spectroscopy-based methods (fluorescence correlation spectroscopy (FCS) and fluorescence cross-correlation spectroscopy (FCCS) and

tandem fluorescent protein timers (tFT), an imaging tool to quantify protein turnover in cells. The latter enabled a proof-of-principle study in which tagging of a chemokine receptor with a tFT allowed ligand-triggered receptor turnover to be measured for the first time in living embryos and led to the demonstration of a novel ‘self-directed’ mode of tissue migration.

To obtain 3D reconstructions of tissues at ultrastructural resolution, automatic block-face imaging will be performed using scanning electron microscopes (SEM) equipped with a focused ion beam (FIB) or an ultra-microtome. The automatic alignment capability of this technology is ideal for tissue-scale problems that require large numbers of serial sections. Applying this methodology to tissues undergoing cell-state transitions, such as the lateral line primordium in zebrafish, the ventral furrow in *Drosophila* and the presomitic mesoderm in mouse, will provide an unprecedented understanding of the fundamental process of tissue generation via cell differentiation and mutual interaction.

Box B.2.8 Self-organisation of the first fate choices in the mammalian embryo

Mammalian development begins with the multiplication of identical cells in the early embryo. This initial symmetry between cells has to be broken during development to establish the different tissues and body plan of the foetus and adult organism. Recent studies unexpectedly revealed that morphogenesis and gene expression are highly dynamic and stochastic during this process. The identity of the decisive symmetry-breaking cue for mammalian development and how it leads to the reproducibly patterned blastocyst are fundamental open questions about the beginning of mammalian life.

A multidisciplinary team of EMBL scientists across the Cell Biology and Biophysics and Developmental Biology Units recently developed new imaging and experimental systems to monitor early mouse development *in toto* at unprecedented spatiotemporal resolution. This allowed them to establish complete cell-lineage maps from zygote to blastocyst (Figure B.2.8 A) and identify the precise moment of symmetry breaking in the mouse embryo. This advance now provides the basis to investigate the cellular and molecular mechanisms of symmetry breaking.

So far, the team have identified that the epithelial polarity that emerges in the 8-cell stage blastomeres (Figure B.2.8 B) plays a key role in symmetry breaking in the mouse embryo and that cell positions become predictive for their future fate at the 8-16 cell transition. Concomitantly, the initially stochastic cell-to-cell gene expression differences progressively stabilise into a reproducible pattern segregating the first lineages of the blastocyst. This self-organisation process likely relies on a feedback loop between gene expression and cell and tissue mechanics to achieve a coordinated developmental programme.

Future research will focus on identification of the all-important symmetry-breaking cue, molecular characterisation of the *de novo* formation of epithelial polarity, and the mechanism by which cell mechanics contribute to the self-organisation of patterning in the embryo. An integrative model based on the complete lineage maps will allow prediction and testing of the emerging properties within the developing system.

A **B**

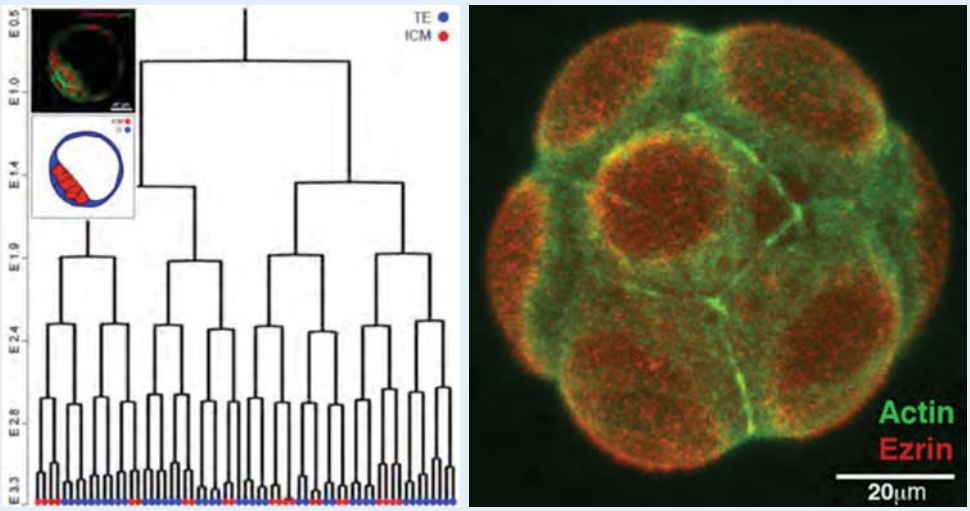


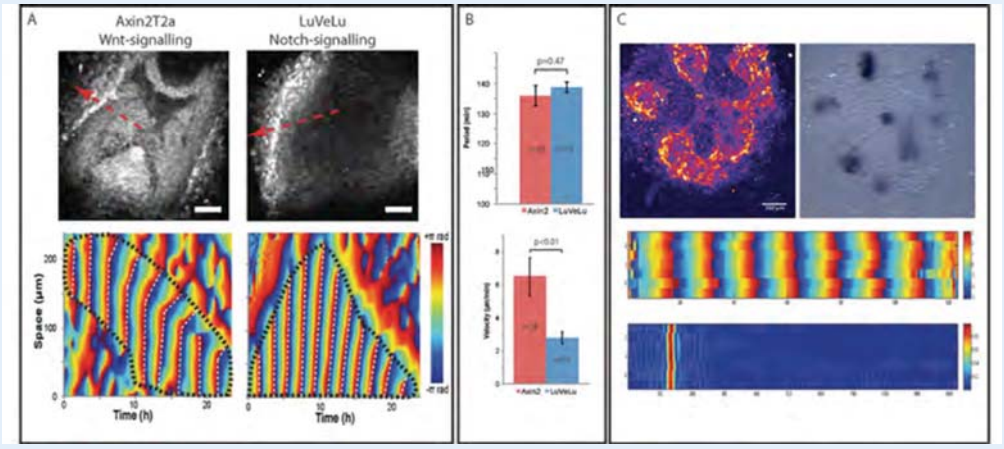
Figure B.2.8 A) Lineage map of the early mouse embryo from zygote to blastocyst. B) The moment of symmetry breaking in the mouse embryo undergoing compaction.

Box B.2.9 From signal oscillations to spatiotemporal self-organization

The process of mesoderm segmentation in vertebrates offers a particularly suitable context to study the principles underlying synchronization of genetic embryonic oscillators and to address their role in developmental patterning. Our quantifications of Wnt- and Notch-signalling oscillations revealed significant differences at the level of wave dynamics and oscillation phase distribution (Figure B.2.9). Conceptually, we therefore propose that in this context of mesoderm segmentation, periodic activity waves of different pathways, i.e. Wnt and Notch signalling, interfere dynamically in order to define both temporal and spatial information. In future, we will investigate how oscillations in several signaling pathways are integrated and decoded in pre-somitic mesoderm cells to control segment formation. We will employ microfluidic technology to perform periodic perturbations of key signaling pathways in a subset of cells and score the effect using dynamic quantifications at both molecular and phenotypic level. Combining an interdisciplinary approach with theoretical modeling, we will aim to reveal fundamental principles of dynamic signal encoding.

In the context of mouse mesoderm development and using an ex vivo randomization assay, we have found clear evidence for self-organization of genetic, cellular oscillators showing, in addition, emergent properties at the level of oscillation dynamics. Our future goal is to address how cellular properties emerge during self-organization as a result of downward causation. To this end, we will investigate the effect of changes in local cellular environment, cellular coupling and spatial constraints, on the process of self-organization. We will combine this ex vivo approach used for exploration and discovery of the working principles with *in vivo* experiments, addressing the role of self-organization at the origin of coherent and synchronized oscillations during the earliest stages of mesoderm development. Our goal is to reveal the basis for self-organized and emergent behaviors and functions, which are characteristic of dynamic, higher-order systems.

Figure B.2.9 (A) Top panels. Real-time imaging quantifications of Axin2 (Wnt-signalling target) and Lunatic fringe (Notch-signalling target) oscillations in *ex vivo* presomitic mesoderm (PSM) segmentation assay. Middle panel. Phase-kymographs show periodic waves that traverse PSM from posterior to anterior. (B) Period of activity waves is identical for Wnt and Notch-signaling (top panel), wave velocities are significantly different (lower panel). (C) Outcome of randomization assay: after culture of ~24 hours, several foci showing in-phase synchronized Lunatic fringe oscillations emerge. These foci correspond to miniature PSM (RNA expression of brachyury, panel top right). All foci are synchronized (middle panel) and show a collective frequency (Fourier transform shown in bottom panel), corresponding to the average of input frequencies.



Box B.2.10 Linking supracellular to subcellular regulation - the interplay between morphogenesis, membrane trafficking and actin dynamics

Development and homeostasis of multicellular organisms requires coordination among cell populations, which leads to the emergence of collective or group properties that are rarely observed in isolated cells. Such properties can be manifested at the tissue level (patterning and differentiation, see also Box B.2.8 and Box B.2.9) or at the level of intracellular organisation. Advances in high-resolution imaging and *in situ* structural biology together with the expertise in genetic manipulation of different model organisms at EMBL will provide a unique opportunity to bridge the gap between the tissue and subcellular scale. This, in combination with modelling approaches, should make it feasible to achieve an integrated understanding of morphogenetic mechanisms (Figure B.2.10).

One particular area of focus will be on the regulation of cell-shape changes and large-scale tissue remodelling that drive the morphogenesis of a multicellular organism. Cell-shape changes are of fundamental importance during embryonic development and require a complex interaction between membrane and cytoskeletal dynamics. Using the early *Drosophila* embryo as a model, multiple groups at EMBL seek to understand how coordination of cell-shape changes drive remodelling, bending and invagination of epithelial tissues and how this in turns relates to changes in membrane trafficking and cytoskeletal organisation. Using a combination of genome engineering and optogenetic approaches, key components of the membrane-transport machinery and cytoskeletal regulators will be manipulated with cellular precision *in situ* (Figure B.2.10). The impact on morphogenesis will be quantitatively analysed using in-house or commercially available light microscopes. Particularly promising in this respect is the possibility to combine optogenetic manipulation of individual cells with single plane illumination microscopy (SPIM) imaging to gain an *in toto* view of cell dynamics and tissue deformation in response to changes in behaviour of individual cells or groups of cells. Automated image analysis will be used to quantify morphological (cell-shape changes, motility, etc) and molecular parameters and dynamics. These data will form the basis for quantitative biophysical modelling, which will allow the subcellular (cytoskeleton, membrane deformations) to be integrated with the tissue scale and guide further experiments. Precise optogenetic manipulations of cytoskeleton and membrane dynamics will provide a unique opportunity to test model predictions at all relevant spatial and temporal scales.

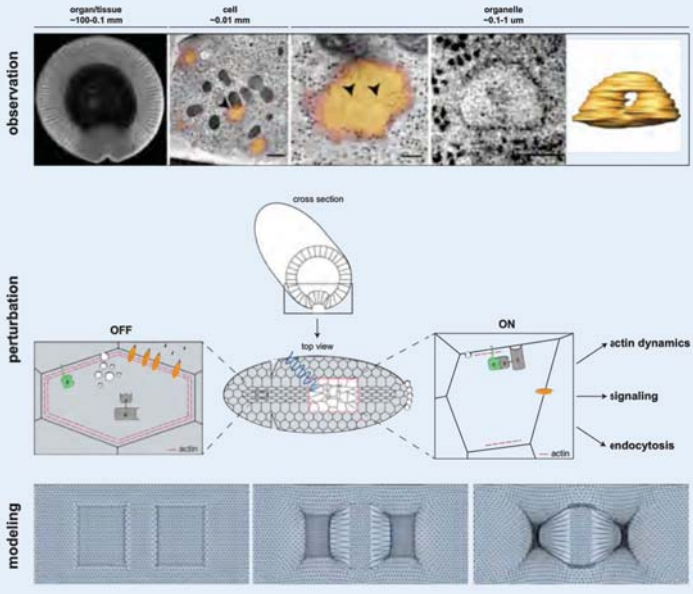


Figure B.2.10 Schematic illustrating the strategy that will be employed to study the impact of membrane trafficking and actin dynamics on large-scale tissue movement during remodeling of *Drosophila* embryonic epithelial tissues. By combining high-resolution imaging and automatic image analysis (top row) with optogenetics (middle row) and modeling (bottom row) an integrated biophysical understanding of tissue mechanics and morphogenesis will be reached. The invagination of the *Drosophila* ventral furrow tissue is shown.

2.4.4 Neurobiology

A particularly challenging aspect of organismal biology is neuroscience. Neuroscience is undergoing a significant transformation that will make it possible to link findings at the molecular and cellular level with those at the circuit and physiology level. The ability to make this link means that researchers can begin to define the molecular mechanisms of behaviour, one of the major aims of neuroscience and of direct clinical relevance. Until now, neurobiology has not been an explicit focus area at EMBL even if the Laboratory has a history of excellence in aspects of neurobiology with groups in several Units pursuing questions of relevance to this field (Sections B.1.1.2.4, B.1.3.1.1 & B.1.3.3.1). Over the next Indicative Scheme we are planning to strengthen EMBL's activities in the area of neuroscience for a number of strategic reasons.

In particular, as described in other parts of this document, EMBL offers profound expertise in three areas that are not traditionally well represented in neuroscience research but that can make critical contributions to its development: molecular imaging, structural biology, and genomics. EMBL is well placed to leverage this expertise to make a significant and unique contribution. In addition, it is increasingly apparent that there is an important interface between epigenetics and aspects of neurobiology. Most obviously, multiple monogenetic disorders that give rise to neurological disease in humans are caused by mutations in chromatin-modifying enzymes (Box B.2.11). In addition, neurodegenerative disorders are frequently accompanied by a loss of normal chromatin organisation, although whether this is causative of disease remains requires further study. Normal brain development and change during infancy, child- and adulthood is accompanied by ongoing changes in epigenetic state, suggesting that functional change in the brain during development and ageing may be driven by epigenetic mechanisms. As a result of recent turnover at EMBL Monterotondo, that has strengthened both neurobiological and epigenetic work there, and upcoming departures, we have an excellent opportunity to focus the Mouse Biology Unit in Monterotondo on the interface between these areas. We would hope to expand the Unit from the current 6 groups and two teams to 10 groups by the end of the next Indicative Scheme to bring sustainable critical mass in both areas. Due to EMBL's investigator-driven approach to research and very stringent standards of recruitment, it is difficult to make exact, detailed predictions about future research directions at this stage. We will aim to recruit group leaders who can operate at the interface between epigenetics/epigenomics and neuroscience and who can make good use of the computational, genomics, structural biology and imaging strengths available in other EMBL Units and thereby provide unique contributions to the fields of neurobiology and epigenetics. The precise research profile of the outstation will be refined with time as more new group leaders arrive. EMBL in total offers a unique interdisciplinary environment to address neurobiological questions that distinguishes it from existing national and international programmes in neuroscience.

Leveraging EMBL expertise in molecular imaging, structure, and genomics

The development of the nervous system involves a precisely coordinated programme of cell proliferation, migration, and maturation that continues from embryonic stages through adulthood. Given that synaptic connectivity is a fundamental feature of the adult brain, the establishment and regulation of dynamic cell-cell interactions are of fundamental importance in brain development. While much is known about the function of synapses in adulthood, much less is known about how synapses are formed, how they reach maturity, and how this can go wrong under pathological situations (e.g. neurodevelopmental disorders such as mental retardation, autism, schizophrenia).

Moreover, it is increasingly recognised that interactions between neurons and non-neuronal cells, including microglia and astrocytes that together with pre- and post-synaptic elements form the so-called ‘quad-partite’ synapse, are essential components for synapse formation, maturation, and function. What are the contributions of each of these elements? What is the trajectory of assembly of their molecular components? How is synaptic efficacy regulated by each component? How is their function different across brain regions and synaptic classes? How do human disease genetic variants and environmental effectors disturb this process?

Crucial to investigating this process will be the identification of the extra- and intra-cellular signalling molecules involved and their functional imaging in time and space. Integration of EMBL expertise in the areas of chemical sensors (Section E.1.1.2.1, e.g. FRET sensors, CLICK-chemistry, SNAP-tagging, molecular timers, lipid sensors), and switches (e.g. optogenetics, pharmacogenetics, Cas-tethered genomic engineering), live tissue imaging (Section E.1.1.2, e.g. light sheet microscopy, super resolution microscopy, quantitative multi-sample fluorescent imaging), optical projection tomography, high-resolution imaging (correlative light-electron microscopy, block-face EM, super-resolution microscopy, EM tomography), and structural biology (Section E.1.1.1, e.g. X-ray crystallography, SAXS, EM) will be critical for the success of this work. Investment in imaging technologies such as serial block-face scanning electron microscopy for large-volume ultrastructure imaging and circuit reconstruction and single-molecule EM and time-lapse super-resolution microscopy for membrane protein complex imaging will be necessary.

Once formed, the adult nervous system must acquire and sustain the processing of sensory information in order to control and adapt its motor and physiological outputs. This requires the coordinated activity of hundreds of cell-types with specialised functions. Moreover, adaptation of these systems in response to environmental factors requires molecular plasticity at multiple levels within an essentially post-mitotic cellular system. Capturing this cellular diversity and identifying the mechanism of plasticity it uses to drive behaviour is a major aim of current neuroscience research to which EMBL can make a distinct contribution. For example, emerging single-cell molecular profiling techniques (Section B.2.2.1) will be crucial to link cell-type specific gene expression profiles with traditional single-cell phenotypes based on morphology, connectivity, and electrophysiology. EMBL expertise in genetically encoded molecular sensors and switches can then be used to monitor and causally test the function of each component in the context of a tissue or the whole animal. With such a functional cellular parts list of the brain it will be possible to identify the molecular mechanisms by which neurons, their synapses, and their circuits undergo adaptive plasticity in response to changes in the environment. What are the relative contributions of transcriptional, translational, and post-translational modifications to cellular and synaptic plasticity? How do post-mitotic neurons reconfigure their epigenetic landscapes to adapt to new inputs or environments? How do cells reconfigure the subcellular compartmentalization of these components to redirect function? What is the role of stochastic mechanisms in neuronal diversity? How redundant are such ensembles? What features determine which neurons join ensembles? How can organisms with many or few neurons carry out the same function? What can comparative studies tell us about the evolutionary logic of neuronal ensembles? Areas for further investment will include CrispR/Cas tools for cell-type specific *in vivo* genome, epigenome, and transcriptome editing across species, single-cell genomics and proteomics, and iPS cell systems for human ‘brain in a dish’ tissue engineering (Section B.2.5).

Toward understanding behaviour

Understanding the neural control of behaviour and how it becomes adapted to an organism's environment is arguably one of the most challenging endeavours of neuroscience. Recent advances in genetically targeted neuronal sensors and switches have initiated a paradigm shift in research into the circuit-based description of behaviour. It is now possible to record activity from large numbers of identified neurons simultaneously during sophisticated behaviours in freely moving animals, including primates, and to manipulate them at will. Several key advances will take place in the near future, including the description of brain-wide connectivity and activity patterns in model species (some of this carried out by EMBL researchers, for example in the annelid model organism *P. dumerilii*), the reiterative use of experimental and computational data to elucidate circuits, the development of rodents and primates with humanized genetic variations (e.g. transgenic marmosets), and the use of unusual animal models to study exceptional behaviours (e.g. singing, hibernation).

These advances will lead in the next ten years to the emergence of several complete circuit-level descriptions of behaviour and will provide the cellular substrates to understand adaptive plasticity at the behavioural level and how such behavioural programmes are perturbed in mental illness. With this cell-to-behaviour knowledge researchers will be able to address a wide range of questions.

Strengthening neurobiology at EMBL

EMBL will take several steps during the next Indicative Scheme period to strengthen neurobiology research with the aim to attract and support new researchers in this area across several Research Units. This will involve further investment in infrastructure and expertise critical to neurobiology, particularly through linkage to imaging (single molecule EM, block-face EM, super-resolution microscopy, Section E.1.1.2) and updates to *in vivo* genome engineering facilities building on the Unit's track record in recombineering and classical mouse transgenic technology to establish an innovative Core Facility, potentially with scope for external service and outreach. Outreach will be strengthened and potential partnerships with neurobiology institutes in the member states will be explored and conferences in neuroscience (e.g. EMBL/EMBO Symposia, technology courses, an EMBL-wide neurobiology retreat) will be fostered.

As outlined at the beginning of this section, there is accumulating evidence that epigenetics plays a crucial role in many aspects of neurobiology and the plan to juxtapose neurobiologists and epigeneticists at the Monterotondo Unit will ensure an interdisciplinary environment where neurobiologists are exposed to cutting-edge research in chromatin biology and gene expression regulation. We expect that cross-fertilisation occurring within the Unit will lead to significant advances in the understanding of the development of neural cell types and the role of genomic plasticity and genetic variation in behavioural adaptation.

Box B.2.11: Chromatin engineering to understand neurodevelopmental disorders

Patient sequencing has revealed a wide spectrum of proteins involved in chromatin biology that are linked to human disease, for example mutations in genes that encode proteins that bind specific chromatin sites (*readers*) and enzymes that modify chromatin (*writers and erasers*) enzymes. As the turnover of canonical histones is replication dependent, post-mitotic tissues like differentiated neurons are very sensitive to effects of these mutations. Many neurodevelopmental disorders, for instance, involve mutations in genes encoding chromatin-readers or writers/erasers. Examples include mutations in the chromatin remodelling factor, ATRX, that causes X-linked mental retardation syndrome, the histone H3K4me3-demethylase, KDM5C, that causes epilepsy, and the histone acetyltransferase, CREB binding protein, which causes Rubinstein-Taybi syndrome. The common theme of these disorders is altered chromatin structure through the mutation of epigenetic regulators. However it is not known how the molecular phenotypes induce specific neurodevelopmental defects.

To investigate this, EMBL scientists will take diverse approaches, integrating knowledge obtained from structural and biochemical studies of re-engineered epigenetic regulators and directly reprogramming chromatin modifications by targeted epigenome editing. Targeted mutations that switch the binding specificity of chromatin readers will be used to causally link histone modifications with chromatin structure and transcriptional regulation. CrispR/Cas9-tethering of transcriptional and chromatin regulators will be used to selectively induce chromatin marks at defined genomic loci, thereby mimicking or reversing disease-related chromatin modifications and enabling tests of their function. Novel viral and transgenic delivery technologies will be developed for the reliable introduction of such modifications in selected cell types in the developing and adult brain. EMBL has significant expertise in both chromatin biology and neural circuit function and strengthening interactions between these fields is currently a major aim. The overarching goal is to use targeted manipulations of epigenetic factors and their effector modifications to causally examine their role in genome function and impact on neural circuit development and plasticity.

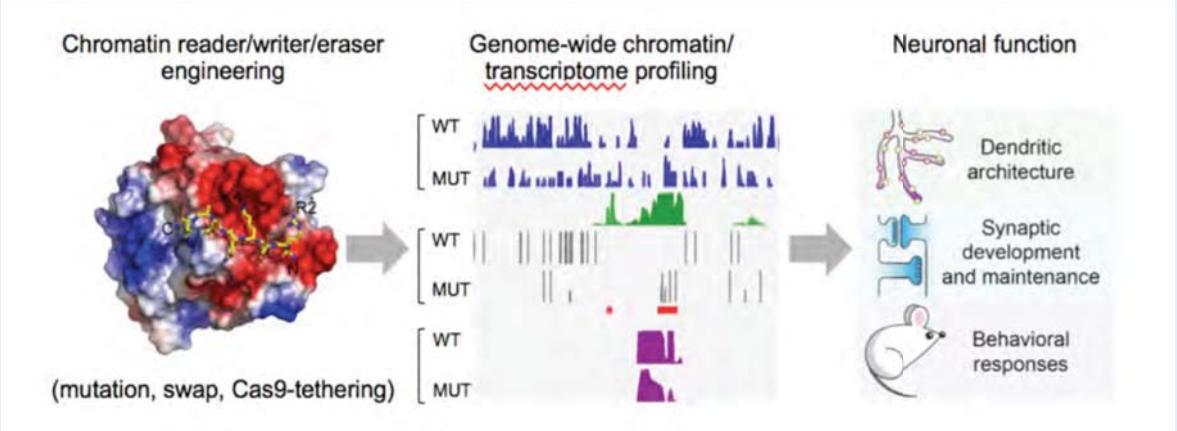


Figure B.2.11 Combining protein and epigenome engineering to understand the role of chromatin regulation in the brain. Structural and biochemical studies of epigenetic regulators will be combined with genome-wide chromatin profiling and targeted chromatin engineering *in vitro* and *in vivo* to test causal links between chromatin regulation and neural circuit development and plasticity.

2.4.5 Beyond organisms – Dissecting the context of microbial communities

As described in the previous sections, to understand how cells are organised into organs and how these then interplay to form multicellular organisms is an extremely complex undertaking. What makes it even more challenging is that the behaviour of each biological system is shaped by both intrinsic (i.e. genetically determined) and extrinsic (environmental) factors. The latter include not only physicochemical parameters but also complex biological variables such as the mix of other organisms present and their interdependences.

Microbial communities are relatively simple systems in which to study the interactions between organisms and the effect of various exogenous factors in different ecosystems. By studying microbial communities as diverse as those present in the human body or in the ocean, we can derive general principles about the functioning of the community in the context of its environment and elucidate its molecular underpinnings. In such settings, the role of various intrinsic and extrinsic properties, as well as their interplay, can be systematically analysed, and we can thereby further our understanding of biodiversity and address fundamental ecological and evolutionary problems. At the same time, we can tackle more practical questions related to, for example, drug efficacy, antibiotic resistance, disease mechanisms, dietary complications in humans and the impact of global warming on biodiversity in the oceans.

EMBL is unique in having both considerable past experience in microbiomics (that is, systemic, large-scale analysis of microbial communities) and the various technologies that facilitate the integration of theoretical and experimental approaches. Techniques to investigate the molecular crosstalk between microbes involve cultivation and genetic manipulation (both individually and as communities), large-scale screening, microfluidics, imaging mass spectrometry, various cutting-edge microscopy approaches, as well as meta-genomic, -transcriptomic, -metabolomic and -proteomic technologies. On the theoretical side, methods for the analysis of various data streams and their integration are in place (Box B.2.1) and modelling and visualisation techniques are being established. All these can be applied in combination to address various important problems in network biology, ecology or evolution. An active and collaborative research community already exists across the different EMBL Units – in particular, Structural and Computational Biology, Genome Biology and EMBL-EBI – and this is supported by the unique and customised set-ups in EMBL's Core Facilities.

To gain a community-level understanding, several layers of complexity need to be tackled. One of the first challenges has long been an issue in ecological studies: to map direct species interactions within a community under a given set of environmental conditions. These interactions can involve both co-operation and competition, and can be physical (e.g. within biofilms) or based on diffusible molecules such as toxins, immunity factors, signalling molecules or nutrients. The multi-tier metabolic exchange between individual members is particularly crucial in shaping microbial communities (Box B.2.12). Identifying these interactions, independent of whether they are binary or involve multiple species, is key for understanding community functioning. As the fluxes that comprise the interactions are determined by many factors, their investigation is a long-term project and requires the integration of a wide spectrum of experimental and theoretical tools. The unique set-up at EMBL in cutting-edge and high-throughput molecular detection and analysis techniques (described above) will allow us to establish large-scale networks of species interactions in communities, embed them into existing ecological theories, and map them with spatiotemporal resolution at different scales,

from submicrometer resolution in densely populated microbial communities all the way up to a planetary scale in global ocean sampling as exemplified by the Tara Oceans project which EMBL researchers have coordinated during the past two EMBL Programmes.

Analogous to protein networks in a cell, species networks have substructure and are often composed of functional modules. Spatiotemporal measurements are one way of detecting such subcommunities but large-scale data-driven approaches, such as co-occurrence analyses in many samples, coupled with ecological processes such as community assembly and succession, can also be very powerful in revealing fundamental community organisational principles.

A complementary layer of systematic analysis is to quantify the impact of exogenous factors on a given biological system. Through this approach, we can address the effects of drugs, natural and dietary compounds, physical parameters and host molecules (e.g. on the human microbiome) or of geochemical parameters (e.g. on ocean microbial communities) to understand the dynamics and evolution of microbial communities.

These analyses will also facilitate our understanding of population dynamics. Whereas the various cell types in a human individual contain almost identical genetic material (differing only by somatic mutations) and their genetic and epigenetic variation is predetermined by their genomes, microbial communities are genetically much more heterogeneous. Yet, until now, there has been no proper quantification of the effects of this fundamental difference on community properties. A bacterial ‘species’ is an abstraction of trillions of different strains that might only differ in a single point mutation or, alternatively, share less than 50% of their genes. We are developing new tools to address these limitations and at the same time to assess the role of strain ‘uniqueness’ within microbial communities – i.e. how the strain composition affects community structure, crosstalk and dynamics. Charting the spatial distribution and dynamics of individual microbial species within communities will help us to determine their ecological function and evolution. It will also have direct practical implications, for example in understanding how microbes acquire undesirable traits (e.g. pathogenicity, antibiotic resistance) and how to prevent their spread between individual hosts or around the globe, and in devising more active ecological intervention in the environment (e.g. optimising soil for crop growth, bioremediation or improving the digestive capacity of a human individual).

This new and exciting area in EMBL’s research portfolio will not only improve our understanding of ecosystems and how to influence them, but it might also provide guiding principles for elucidating cell communication and organ functioning in multicellular organisms (Sections B.2.4.3 & B.2.5). As effective modules and division of labour in microbial communities are thought to be prerequisites for the evolution of complex systems, their understanding might also shed light on the evolution of multicellularity per se (Section B.2.4.3). Finally, the evolution of microbes can be traced together with that of co-existing macroscopic systems, such as their human hosts, and their co-dependence can be systematically assessed (Section B.2.4.6). It is becoming clear that our lifestyle not only influences our own physiology but also contributes to the shaping of our microbiome, which has altered considerably within a few generations. More importantly, humans have an impact on many other ecosystems on earth and can impose drastic selective constraints on our environment, including the microbes that are part of it. Understanding microbial communities in selected ecosystems (including ourselves as one ecological habitat) might help us to understand and model ecosystem evolution and to apply this knowledge in regaining sustainability in our environment.

Box B.2.12: Towards a molecular understanding of species interactions in space and time

Microorganisms living in a community are dependent both on the abiotic and biotic (hosts, fellow organisms) components of their environment for the provision of essential nutrients and often compete with each other for limiting resources. Conversely, nutrient exchange between microbes can promote co-operative interactions and offer a group advantage. The resulting competing and co-operative interactions are major ecological forces that shape species composition, spatial organisation and community function. Uncovering the molecular players mediating these interactions and mapping their genetic basis is a fundamental challenge in understanding the role of microbial communities in health and disease. EMBL researchers will tackle this challenge using a multidisciplinary strategy that combines large-scale and directed experimental approaches together with bioinformatics and mathematical modelling.

Meta-genomics, meta-transcriptomics and meta-metabolomics analyses will be used to identify species co-occurring across diverse samples as candidate interacting groups. This knowledge will be complemented by microfluidic approaches to capture and co-culture species and to identify minimal subcommunities. To identify the molecular basis of their respective interactions we will: a) probe metabolic exchanges with imaging mass spectrometry; b) use a number of microscopy techniques (from light to electron microscopy) to gain insights into interactions at single-cell level; and c) profile mutant libraries and assess genetic interactions in a high-throughput manner for interactions of particular relevance, e.g. those involving probiotic or pathogenic species. In the case of human gut communities, we will supplement these studies with host cells in two-dimensional and more complex environments and assess the interplay between them and the microbes.

Building on different synergistic meta-omics studies, *in vitro* interactions and genetic screens, we will quantify inter-species metabolic exchanges using community metabolic modelling and validate the predictions using quantitative, time-resolved metabolomics experiments. Spatial organisation in communities will be similarly tackled by a combination of microscopy-based approaches and imaging mass spectrometry. For the human gut microbiome this will be complemented with data from disease or dietary intervention studies on large cohorts of individuals from many countries. EMBL scientists are already participating in various international microbiomics studies, e.g. in the areas of diabetes, obesity, colon cancer or inflammatory bowel disease, and we expect these activities to expand in the next programme period. We are also exploring ways to involve the public more in these projects, for example through crowdsourcing mechanisms.

With the described multidisciplinary approaches, we aim to uncover fundamental species interaction principles that underlie community structure, function and dynamics. These principles will be applicable to biological systems and questions beyond microbial communities, such as cell–cell communication in the context of tissues and organs and the evolution of multicellularity.

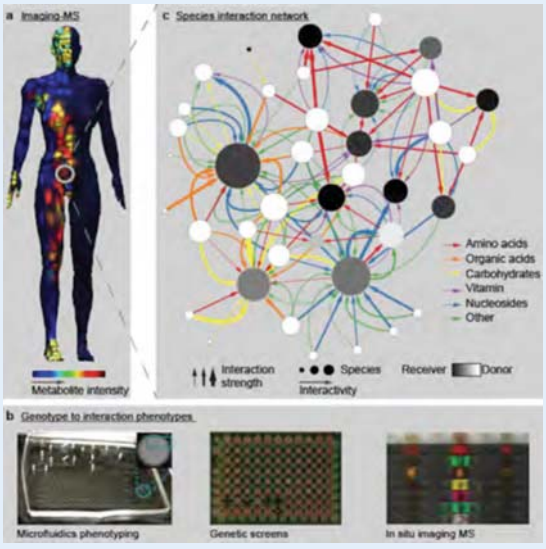


Figure B.2.12 An integrative approach towards uncovering microbial species interaction networks in the human gut. (a) Spatial distribution of metabolites (imaging mass spectrometry) associated with the microbial communities inhabiting our body. (b) Dissecting microbial communities with: (i) droplet-based microfluidic chips (left) enable high-throughput screening of single cells or small communities for particular genotypes or phenotypes; (ii) genetic interaction screens between species (middle); and (iii) using metabolite variation, captured by imaging mass spectrometry and visualised by different colours (right). (c) Reconstructed and modelled metabolic interaction network depicting the flow of metabolites in a 50-species community representing the human gut microbiota.

2.4.6 Transitions in evolution - Tracking the emergence of proteins, complexes, cells, organisms and communities

Evolution provides the ultimate context in which all life takes place and develops. It acts at all levels of life – from single proteins to complex cellular machineries, from organelles to cells, from multiple cell types to tissues and organs, and, finally, from the species to the community level. A holistic view of organismal evolution will only emerge when all these levels are understood and integrated. Complicating matters further, the genetic information that encodes life – written in the genomes of the millions of species that live today – has become more and more complex over time.

Evolutionary research faces the challenge that many of the processes that led to today's diversity took place hundreds and even thousands of million years ago. These ancient forms can be directly assessed only via the fossil record, which is often obscured by artefacts and in which many of the levels of organisation mentioned above are simply not preserved. An alternative approach is to track the historic rise of organismal life by comparative research – the inference of ancestral states via the study and comparison of extant organisms. For centuries, this approach was restricted to the use of morphological data, which has proven to be of limited informative value and has left large parts of organismal evolution unsolved.

As a consequence, current evolutionary research has a strong emphasis on the most accessible, i.e. the most recent, period of the evolutionary past – referred to as 'microevolution'. Population genetics infers principles of evolutionary change through the study of allele frequencies in different habitats, and experimental evolution investigates generations of microbes under either defined or changing conditions for the same purpose. Yet, if the whole evolution of life equals one day, these disciplines necessarily focus on what happened in the past five minutes or less and thus cannot elucidate what is commonly referred to as 'macroevolution' – the big picture of the emergence and diversification of life on Earth.

This situation has begun to change profoundly with EMBL researchers playing an important and leading role. Major triggers have been genome sequencing for an unprecedented number of species in all kingdoms of life (Section B.2.2); the sequencing of single-cell transcriptomes for various tissues and organisms (Section B.2.2.1); the revolution in imaging technology bridging scales (Sections B.2.4.1 - B.2.4.3, B.2.5, E.1.1.2); the ease of transfer of experimental techniques into new model systems; and the sampling of complex microbial communities over time and space (Section B.2.4.5). These advances have enabled a new quality of comparative research that, for the first time, has the potential to elucidate evolutionary processes at all levels simultaneously and to thus track the evolutionary changes that occurred along major lineages of organismal evolution. This will be a goal for research groups at several EMBL sites over the course of the next Indicative Scheme.

Mapping sequence data onto evolution

As a prerequisite, EMBL researchers have recently made an important contribution to unravelling the universal tree of life, i.e. the branching pattern of major evolutionary lineages. This is needed to correctly place available comparative data and to infer the interrelationships of the species under study. Most importantly, this will allow us to 'map' the plethora of sequenced genomes as well as metagenomic data (Section B.2.2 & Box B.2.12) onto this tree and, once sufficient phylogenetic coverage is obtained, to infer the origin and diversification of gene families and of the proteins they encode. The Ensembl databases at EMBL-EBI, which host genome information for various branches of the tree

of life (Section C.1), provide a strong comparative framework towards this aim. Through these databases, the incoming flood of genomic information will be transformed into an 'open book on protein evolution' (with limitless pages) that we will then have to learn to read. When did the proteins and protein complexes that shape cells first come into existence? What was their primary structure and function in the cellular context? Several groups across EMBL, in the areas of developmental, structural and computational biology and bioinformatics, are already tracking the evolution of protein structure, protein post-translational modifications and protein complexes.

Besides coding sequences and the proteins they encode, the evolution of the non-coding part of the genome is equally relevant for understanding organismal macroevolution. These sequences control gene accessibility and expression and can thus be used to put the genomic information into better temporal and/or spatial context. Evidently, the evolution of eukaryotes and especially of multicellular organisms has been accompanied by a vast increase in the complexity of the regulatory genome, as seen for instance in the genomes of the fly or mouse. Taking the bacterial operon as a starting point – does a more complex, eukaryote correlate of this operon exist and would it fit into the chromatin structure? A first view on the evolution of gene co-regulation is currently emerging with EMBL groups in the forefront. The comparative analysis of genomic regulatory architecture is an important aspect of EMBL's future plans.

Evolution of cell types and multicellular organization

In multicellular organisms, cells group into types. Tracking the evolution of such cell types represents another important branch of macroevolutionary research that will be undertaken by researchers at EMBL Monterotondo, EMBL-EBI and in Heidelberg. Cell types represent key evolutionary units that diversify in the same way as genes, protein and protein complexes. Most importantly, sequencing of single-cell transcriptomes (Section B.2.2.1) from various tissues, organ systems and organisms will enable, for the first time, a broad, large-scale comparison of cell types across large distances in the evolutionary tree. These cell types are composed of and defined by specific cellular modules – such as receptor complexes, signalling cascades, cytoskeletal elements controlling shape and contractions, cellular junctions, and more specialised units such as the presynapse for information exchange. Single-cell data allow us to track these modules across cell types and organisms and thus elucidate their step-wise emergence and diversification. Some studies carried out in the current EMBL Programme shed new light on key evolutionary events such as the evolution of neurons and of the nervous system (Box B.2.13) and these will be greatly expanded in the next Programme.

At the next higher level, cell types interact to form tissues and organs. Understanding tissue and organ evolution is pivotal for tracking the macroevolution of multicellular species and, puzzlingly, this aspect remains almost entirely unknown and unstudied. What are the principles governing the evolutionary aggregation of cell types into tissues and organs? To what extent are these processes triggered by genomic changes and what is the role of self-assembly? Mesoscopy of tissues and organs in different species, which will be the focus of EMBL's new outstation in Barcelona (Section B.2.5), has the potential to provide major insights into morphological transitions in macroevolution and eventually reveal principles of organ evolution. Similar to the interaction of cell types in multicellular organisms, unicellular pro- and eukaryotes assemble into organismal communities, triggered by both intercellular and environmental signals (Section B.2.4.5). Trying to unravel the evolution of these supra-organisms, such as oceanic plankton or microbial gut communities (Boxes B.2.1 & B.2.12) will be both extremely difficult and most rewarding.

Box B.2.13: Deciphering evolution of neurons and nervous systems

The number of human cell types is in the hundreds, with different neurons and other neural cell types representing the largest fraction. The evolution of neurons from non-neural precursors, concomitant with the emergence of the nervous system in animals, is not well understood. We will characterise and compare neural cell types between animals that possess nervous systems of different complexity (cnidarian, annelid, amphioxus and mouse, see Figure B.2.13 A) and elucidate their step-wise emergence in animal evolution.

The earliest branching metazoans, the sponges, do not have neurons but instead possess interconnected sensory and contractile cells that exhibit contractile waves moving across the body. It has been postulated that these sponge cell types are related to, and still resemble, ancient 'protoneurons' that gave rise to true neurons in the line of evolution leading to cnidarians and other animals with nervous systems. The nervous system of cnidarians consists of a nerve net (that is, a net of neurons interconnected by axon-like neurite processes; panel B). The bilaterian ancestor of annelids, amphioxus and mouse most likely already possessed a centralised nervous system, which is commonly found in extant bilaterians. Amphioxus has a neural tube but lacks an extended brain and peripheral nervous system as present in vertebrates (panel C).

We will dissociate cells from representatives of the above groups and sequence their transcriptomes. This will be complemented by proteomic studies of the same species, to reveal the neural modules present in protoneurons or neurons. In cnidarian, annelid and mouse the function of selected differentiation genes can be tested via knockout studies to assess the function of specific modules; likewise, the activity of transcription factors specific for neuronal cell types will be compared. Cross-species comparisons of 'differentiation signatures' and 'regulatory signatures' will reveal cell-type interrelationships. In particular, we will determine which, if any, of the sponge cell types is related to neural cell types in the other species; furthermore, we will test how the neuron types of the cnidarian nerve relate to those of the central and peripheral nervous system in bilaterians. Finally, we aim to identify neuron types in the amphioxus neural tube that, by their molecular signature, qualify as precursors of the specialised cell types of the vertebrate brain, such as cortical neurons.

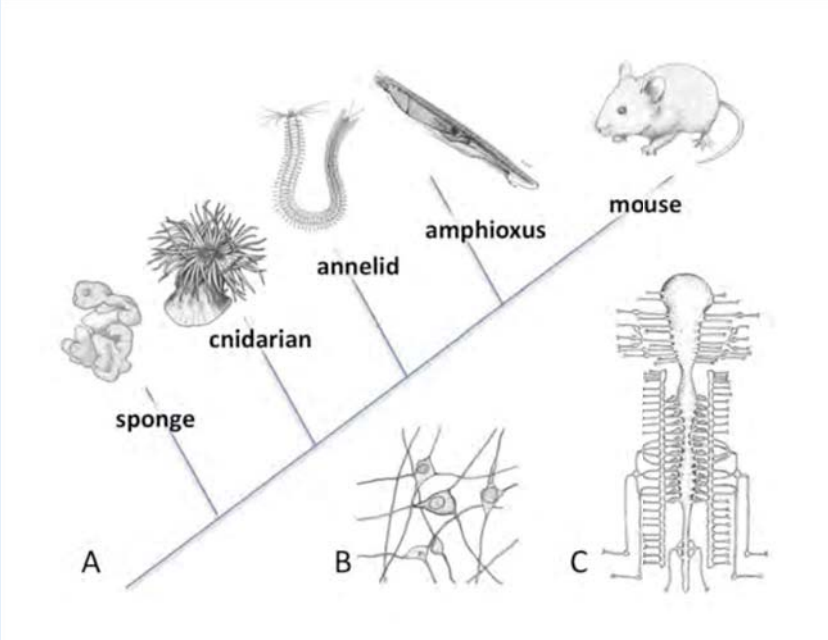


Figure B.2.13 A. Simplified animal evolutionary tree, depicting major lineages branching off the line of evolution leading to the vertebrates. B. A nerve net composed of dispersed neurons interconnected by neurites, as present in cnidarians. C. The vertebrate central and peripheral nervous system.

2.5 Tissue Biology and Disease Modelling

In November 2014, EMBL Council took the in-principle decision to create a new EMBL outstation in Barcelona, Spain. The new outstation will build on the existing partnership between EMBL and the Systems Biology Unit of the Centre for Genomic Regulation (CRG) and like all other EMBL sites, the outstation will conduct research and provide services to the scientific community. The research and service plans were described in detail in a scientific proposal that was evaluated and endorsed by EMBL's Scientific Advisory Committee in 2012. These plans and an outline budget were discussed with the member states prior to EMBL Council's in-principle decision to establish the new outstation.

In the following we outline a summary of the scientific scope and future plans of the outstation for the period of the next Indicative Scheme. These plans have to remain very broad at this stage as neither the leadership nor the research group composition of the outstation has been defined. The general scientific direction for the outstation was chosen on the basis of three main criteria: (a) choosing a scientific area which represents an exciting vision and where breakthroughs are expected from both a biological and medical perspective, (b) taking advantage of the unique and powerful expertise that has been built within the EMBL-CRG Partnership Unit, and (c) selecting a theme which is complementary to current EMBL research, which has enormous potential for synergy with existing groups and therefore adds maximal value to EMBL's member states and Europe's scientific community.

2.5.1 Future research plans at the EMBL outstation in Spain

The scientific focus of the new outstation will be on 'tissue biology and disease modelling'. Its strategy will be to address the complexity of multicellular systems mechanistically through cyclic iterations of computational modelling and experimentation. Understanding multicellular organisation represents a major scientific challenge for systems biology. Tissues and organs are under the control of molecular and genetic networks, but in a complex dynamic fashion that extends across many intercommunicating cells, and with multiple different cell types in carefully-controlled positions relative to each other. The challenge of modelling and understanding these structures derives from their multiscale nature (from molecules, through cells up to organs). Nevertheless, a true understanding of these systems, and especially the dynamic homeostatic mechanisms by which they maintain their functional organisation, is crucial for addressing many problems in health and disease.

The new outstation is uniquely poised to tackle this difficult challenge, because it can build on special expertise and skills that the EMBL-CRG Partnership Unit has acquired in the areas of multicellular tissue/organ-level imaging, image-driven modelling, computational modelling across multiple scales, and experimental genetics on model species to understand phenotypic variability.

Main themes for the scientific approach at the new EMBL outstation

- Focus on tissues and organs, and their multiscale nature from molecules upwards.
- Mesoscopic imaging to obtain quantitative data about multicellular systems.
- Incorporate both *wet-lab* experimental and *dry-lab* computational sides within research groups.
- Iterations of modelling with experimental testing of model predictions.
- *In vitro* manipulable systems, derived from both humans and model species.

Research at the new outstation will concentrate on the 'mesoscopic' scale linking molecular and cellular biology on the one hand (micro-scale), and larger-scale physiology on the other (macro-scale). This ensures complementarity with existing EMBL Research Units such as the Structural and Computational and Cell Biology and Biophysics Units (primarily focused at molecular and cellular levels), and the Developmental and Mouse Biology Units (whose interests include organism-level physiology and behaviour). It will also be complementary to the genomics, bioinformatics and modelling approaches pursued within EMBL's Genome Biology and Structural and Computational Biology and EMBL-EBI Units. As a result, there will be ample opportunities for productive synergy and collaboration with other EMBL Units.

The new outstation will investigate phenomena relevant to human health and the plan is to focus increasingly on methods applicable to human systems such as cellular co-culturing, organoids and tissue engineering of healthy and diseased tissue. In this way the outstation will aim to dissect how the properties that emerge from interactions at various levels of biological organisation contribute to health and disease. This again complements other ongoing and planned EMBL research activities aimed at unravelling molecular processes in humans (Section B.2.3).

There will be a smooth transition from the current Partnership Unit to the future EMBL outstation that will be marked by a gradual expansion of systems biology research into the area of human systems, for example by supplementing existing work in other parts of the CRG on genomic data from patients, tissue-banks and human biopsies with complex human cell culture, co-culture and tissue engineering approaches. Model systems and organisms ranging in complexity from human 3D cell (Section B.2.4.2) and organ culture to invertebrate systems (Section B.2.4.3) will retain an important role because they allow the predictions of computer models to be tested experimentally using a wide variety of perturbations (genetics, RNAi, drugs, biophysical manipulation, etc.) in material derived from normal or disease states.

The new outstation will be embedded in a vibrant scientific environment in Barcelona. Being located in the same building, we foresee intense exchange and close collaboration with the CRG. As well as a variety of core facilities that will be accessible to the research groups in the outstation, the building also houses the Barcelona Centre for Regenerative Medicine, which hosts the Spanish bank for induced pluripotent stem cells and embryonic stem cells. These research groups will be attractive partners for collaborations in the study of human multicellular systems and organs. Moreover, Barcelona is home to a variety of leading clinical research facilities and hospitals. Although translational research will not be performed within the outstation, it will be a possible focus for collaborations with neighbouring hospitals. Such collaborations might, for example, provide access to primary cell lines from patients, tissue-banks and biopsies.

Research areas

Like all other EMBL Research Units, research groups at the new outstation will cover several topics and research areas. In the following we present three exemplary areas to illustrate the scientific scope of the new EMBL site.

Disease mechanisms

Mechanisms of disease are already studied by many research organisations around the world from many different perspectives: genetic, molecular, physiological, epidemiological etc. The goal of the new outstation will be to develop a more holistic approach that explicitly adds the tissue and organ dimension to our understanding, and which combines imaging and computer modelling. Prime examples that may be developed include:

Congenital abnormalities. A large number of birth defects are present in the human population, ranging from relatively mild health impacts (such as polydactyly), to severe medical impacts (such as spina bifida or heart septal defects). Although many genetic loci have been revealed that map to these diseases, establishing a mechanistic link is extremely difficult because of the complex multi-scale nature of morphogenesis. Computer models of organogenesis are becoming more powerful and are beginning to provide predictive mechanistic links between molecular changes and macroscopic phenotypic effects. These data- and image-driven simulations will increasingly become central to our scientific understanding of congenital abnormalities. Useful models of heart and neural tube development,among others, will be developed over the next decade, and the new outstation will seek to take advantage of the existing local expertise to become a leader in this field.

Cancer: metastasis, tumour growth and angiogenesis. Constructing dynamic computer models of the processes involved in cancer is still in its infancy, but is a rapidly growing area of systems biology. In parallel, co-culture systems using primary human cells or human cell lines are allowing, for example, the morphogenesis of normal and transformed mammary ducts or lung alveoli to be studied in 3D cell-culture systems that can be manipulated with the repertoire of methods available to simple cell culture. Similarly, ‘Avatar’ mice that recapitulate the natural environment of human tumours can serve as good models, in which controlled perturbations can be applied in combination with imaging and –omics approaches to model complex cancer processes. Systems-level modelling projects, combining experimental and theoretical approaches, will be essential to understand the molecular and cellular mechanisms involved, and will depend on the advances being made in the area of quantitative 3D imaging data in the current EMBL-CRG Partnership Unit and engage in collaborations with complementary efforts in imaging technology in the Cell Biology and Biophysics Unit in Heidelberg (Section E.1.1.2).

Immune disorders. The immune system is a classic example of a complex, distributed, multiscale system, which displays non-intuitive dynamics and behaviours. A prime research area for systems biology is degenerative disease caused by auto-immune attack. Examples such as type 1 diabetes (often caused by auto-immune damage to insulin-producing beta-cells) are already benefiting from novel mesoscopic imaging technologies. Improved imaging of beta-cell mass from the entire pancreas of mouse models of diabetes using novel mesoscopic imaging techniques is revealing the temporal sequence of islet destruction in relation to the 3D spatial arrangement of the whole pancreas. This contributes to a better understanding of the mechanism by which lymphocytes mount the auto-immune attack. Quantitative data collection also lends itself

to cell-based computational modelling of the cellular dynamics so that it should be possible to build predictive models of the mechanisms involved.

Another example is the dynamics of the immune response to infection, in which lymph nodes are an essential component. They normally display a clear spatial organisation of cell types but upon infection, this complex arrangement actively and dramatically alters. Little is known about how this process occurs, but mesoscopic imaging techniques are allowing this question to be tackled for the first time (Figure C) and are providing the 3D quantitative data necessary to build cellular models of the dynamics.

Systems Genetics. An important challenge in modern medicine is how to use genetic and genomic data to make accurate predictions about the health and disease risk of individual patients, despite the low predictive power of individual gene polymorphisms and mutations for disease occurrence. Another aim of the outstation will therefore be to use model systems to develop and test methods to accurately predict phenotypic variation at the tissue and organ level among individuals, in both humans and model species. To work towards this goal, a collaborative proof-of-principle project is already in place between EMBL's Genome Biology Unit and the EMBL-CRG Partnership Unit using *Drosophila* as a model system. Another system potentially amenable to this approach is the new paradigm of *in vitro* organoids (discussed in more detail below). A systems genetics approach will also allow synergies with EMBL-EBI, which is exploring the genomic contribution to organ-scale phenotypes such as human heart morphology via bioinformatics.

Tissue engineering

Understanding how large groups of cells differentiate and geometrically arrange themselves has benefits that go beyond unravelling disease mechanisms. Perhaps the most exciting of these will be our ability to control tissues, and thus to heal damage, regenerate tissues and, one day, possibly design and build organs in a rational manner. Recently, a new perspective on these themes has emerged with the potential to revolutionise the area: *in vitro* organoids. Combined with advances in 3D scaffolds, this new research paradigm extends basic multicellular biology from understanding to actually engineering complex systems.

In vitro tissue creation. A variety of *in vitro* systems have been developed in recent years for culturing ectodermal derivatives (gut organoids and liver buds), neurectodermal tissues (optic cup and cerebral organoids), and even mesodermal tissues (nephric tubules and glomeruli). The potential exists for learning how to influence these developmental processes *in vitro*, to create tissues for transplant into damaged organs, or even to form complex structures. The multiscale complexity of this process (gene regulation coupled to cell-fate choices and multicellular architecture) will require sophisticated new approaches to be successful. But the discovery that much of this work can be done *in vitro* provides tremendous advantages. Transgenic fluorescent reporter constructs allow molecular events and cell-fate choices to be monitored within the multicellular context, and the new mesoscopic imaging techniques allow these processes to be quantified live. This is a high-quality source of quantitative data from which multidimensional computer models can be built, and the new outstation will aim to specifically recruit group leaders in this area.

Artificial Organs. As an alternative to growing 3D structures directly, the use of artificial scaffolds to construct complex cellular structures is being explored. Slowly biodegradable structures hold significant promise, for example in creating an artificial lymph node in which a suitable cocktail of manipulated lymphocytes and stromal cells are seeded, and which could be useful in boosting adaptive immunity or the efficacy of vaccines. The dynamics of lymphocyte interactions within a lymph node depend on understanding the geometry of the tissues and the changing positions of cells, and so both mesoscopic imaging and multicellular computer modelling will play an important role in studying these systems. Outstation groups using scaffolds would be an excellent complement to groups studying pure organoids, and the two approaches may synergise into a hybrid strategy.

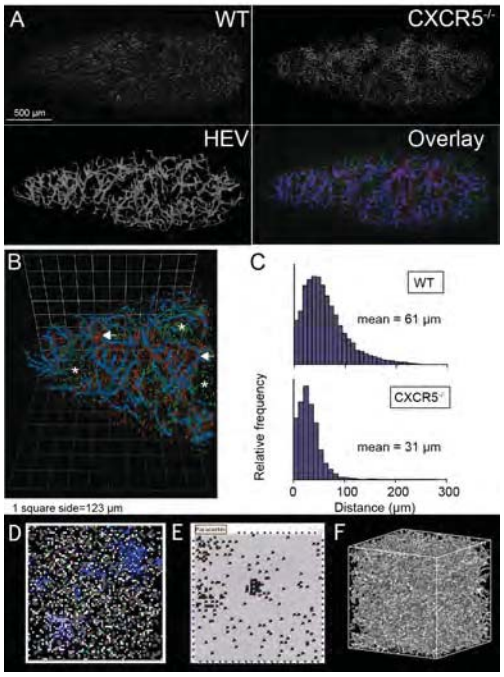


Figure C Mesoscopic imaging and computational modelling of immune dynamics. (A) Whole-organ imaging of mouse lymph nodes reveals the spatial distribution of wild type (WT) versus mutant B cells in relation to the HEV network. (B) The quantitative nature of the image processing allows accurate numerical comparisons (C) to be made. (D-E) These data will provide the basis for spatiotemporal computer modelling of the immune system during response to infections, either in 2D + time (D,E), or 3D + time (F).

Tissue technology

Like all of EMBL's Research Units, the new outstation will also engage in technology development, which directly supports the systems biology approach to tissue and organs. One or two of the new recruited groups might be selected on the basis of their technical expertise. In particular, the following areas will be highly useful for the outstation:

3D *in vitro* tissue technologies. Success in the new era of tissue engineering is likely to depend on our technical ability to manipulate small multicellular structures *in vitro*. New scaffold types allow new tissue types, and other advances are being developed such as 3D cell printing. These approaches are in their infancy. Progress for the outstation may be possible through collaboration, but if a strong group can be employed with the potential to become a leader in these technologies, it would support the planned research activities very well.

Computational reverse engineering of gene networks using imaging. So far, most effort in the field of reverse engineering has been at the level of single networks in single cells. A growing alternative is the use of quantitative images as a source of data for reverse engineering. The most advanced example of this is so far is the reverse-engineering of the gap gene network in the early *Drosophila* embryo, involved in the earliest stages of laying down the fly body plan. The current EMBL-CRG Partnership Unit, has leading expertise in this approach. As the new outstation develops, an important potential focus will be on technical/methodological developments that will allow this powerful approach to be extended to a wider range of tissues, including human samples.

Quantitative imaging. Mesoscopic imaging will be at the heart of both the research and service activities of the outstation (Section 2.5.2). These technologies are already quantitative at various levels, but the field is still young and improvements are possible and necessary. Given the high value of reliable quantitation for systems-modelling approaches, the outstation will also pursue technical improvements in quantitation for optical projection tomography (OPT, developed in the EMBL-CRG Partnership Unit), light-sheet imaging (developed in EMBL Heidelberg) and other technologies (improved algorithms, calibration with phantoms, etc.). This activity will allow collaboration with the complementary light microscopy developments pursued in Heidelberg. Particular challenges include minimally invasive imaging conditions and improved label-free imaging channels.

Image-driven modelling. The proposal to build predictive computer models of multicellular systems rests on two non-trivial computational areas. Firstly, geometric representations of tissues, organs and the cellular distributions within them need to be extracted from raw 3D images. Applying these approaches to mesoscopic samples is still in its infancy. Advanced tools for 3D segmentation, adaptive shape analysis and morphometrics are all areas with scope for development within the outstation, depending on the recruitment of suitable expertise. Secondly, algorithms and formalisms for dynamic simulations of tissue-scale dynamics still need to be improved. Current projects in the EMBL-CRG Partnership Unit already represent some of the latest state-of-the-art algorithms in this area, but further improvements could be pursued in collaboration with the modelling groups of EMBL-EBI and EMBL Heidelberg.

2.5.2 Future service plans at the EMBL outstation in Spain

The new EMBL outstation will offer a services in the areas of:

Optical mesoscopic imaging

The facility will provide access to state-of-the-art equipment and technology for *mesoscopy* – ie. imaging of samples in the millimetre-to-centimetre range (Figure D). Mesoscopy has a growing range of applications from organ development, tissue regeneration and neural connectivity, to analysis of human biopsies. As the technology is advancing rapidly, it is not possible to predict in detail which instrumentation will become available. However, the current EMBL-CRG Partnership Unit possesses particular strengths in the area of OPT and light-sheet microscopy, and is currently working on improved types of hybrid imaging technology that will be able to maximise the information extracted from each sample.

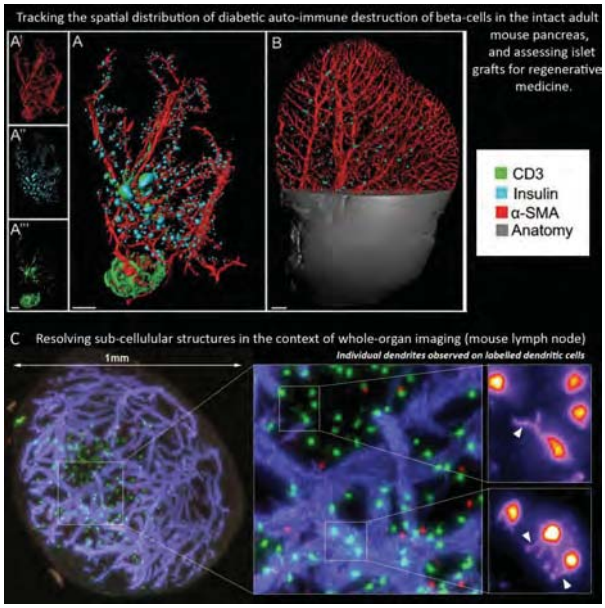


Figure D Mesoscopic imaging for biomedical research. (A,B) Optical projection tomography (OPT) is able to help researchers understand the dynamics of beta-cell destruction in type 1 diabetes, and to assess the viability of methods for regenerative cell therapy. (C) Light-sheet microscopy can provide sub-cellular resolving power within the context of intact adult mouse organs, thereby spanning the scales from organelles to organs.

Image-driven model building

As the facility matures, an increasing emphasis on model-building will be developed. So far, computational modelling has typically been a research topic, rather than a service area. However, just as sophisticated imaging techniques and methods gradually progress from experimental work into a service, so too will certain types of model building become more standardised. The facility will therefore employ skilled service staff to help visiting scientists develop their imaging data into computational representations and models of various types. It is envisaged that services will be provided in the areas of quantitative data extraction from images, the extraction of 3D atlases and the inference of lineages and networks (e.g. for spatial transcriptomics, Figure E), modelling of tissue dynamics and the user-friendly hosting of models that allows users remote access.

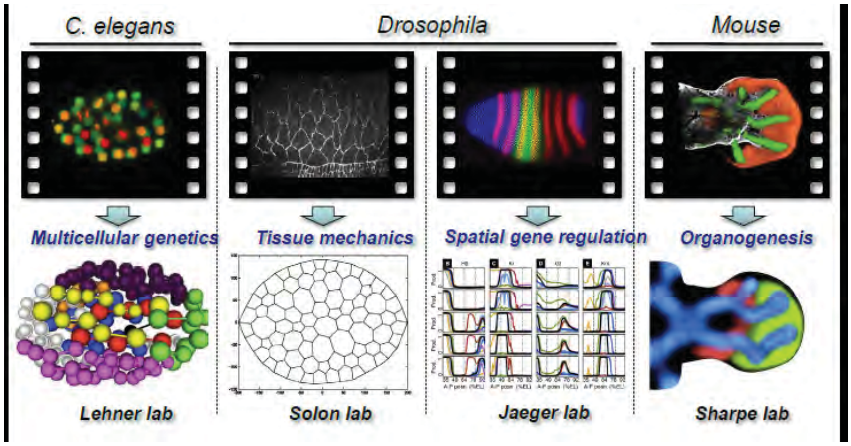


Figure E Examples of mesoscopic imaging and modelling currently performed at the EMBL-CRG Partnership Unit and at the CRG.

As at all other EMBL sites, the service facility will be closely linked to the research and technology development activities and will make new cutting-edge instrumentation and methodology developments resulting from research available to all EMBL researchers and scientists from EMBL's member states.

The suggested service is unique in Europe and provides substantial added value to EMBL and European science. Service in mesoscopic imaging technology and expertise (for samples in the *mm* to *cm* range) is currently not available at EMBL and pursued at very few research centres across Europe. The services offered will be complementary to what is currently provided by EMBL's Advanced Light Microscopy Facility (Section C.3.6). Close interaction and collaboration is foreseen between the two facilities that will both benefit from the exchange of expertise and skills and the sharing of newly developed technology.

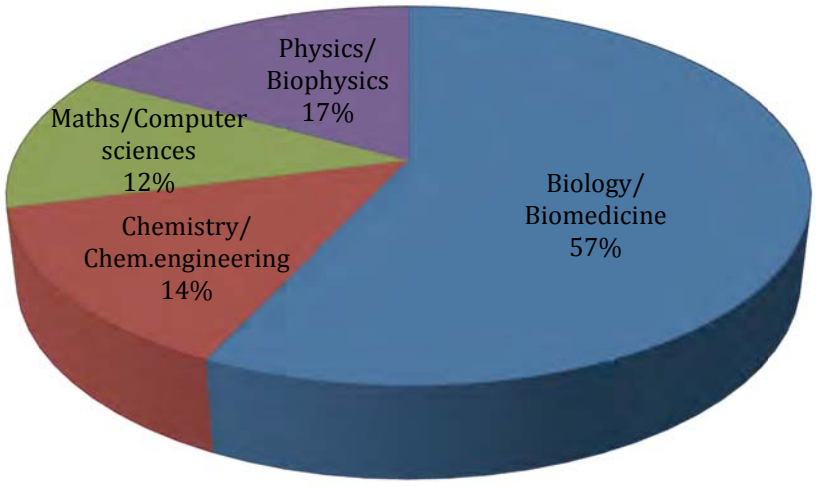
The facility will be run by a few dedicated service staff members who will assist users in operating the microscopes, offer help with data analysis and provide training in modelling software. Like all EMBL Core Facilities, a user committee will be created to enable regular feedback from users, from high-end to occasional. The facility will be open to a mixed user community consisting primarily of researchers from all EMBL sites and the CRG, but also users from Spain and EMBL's other member states. In the future, we hope that additional capacity for pan-European access to the facility will be provided in the context of Euro-BioImaging, where CRG was the proof-of-concept facility for mesoscopic imaging (Section F.1.3.2).

3. Initiatives to foster interdisciplinary collaboration

The need for interdisciplinary research in the molecular life sciences is ever-increasing. As Section B.2 Research Themes of this document illustrates, EMBL's future research activities will require the comprehensive integration of methods and insights from multiple disciplines to navigate across scales – from molecules to organisms and beyond – and thus gain a holistic understanding of biological systems.

EMBL is well prepared for this challenge. Our research community is inherently interdisciplinary as Figure B.3.1, an overview of the diverse academic backgrounds of EMBL group and team leaders, illustrates. Apart from biologists, EMBL recruits chemists, physicists, computer scientists, mathematicians, engineers and scientists with medical backgrounds.

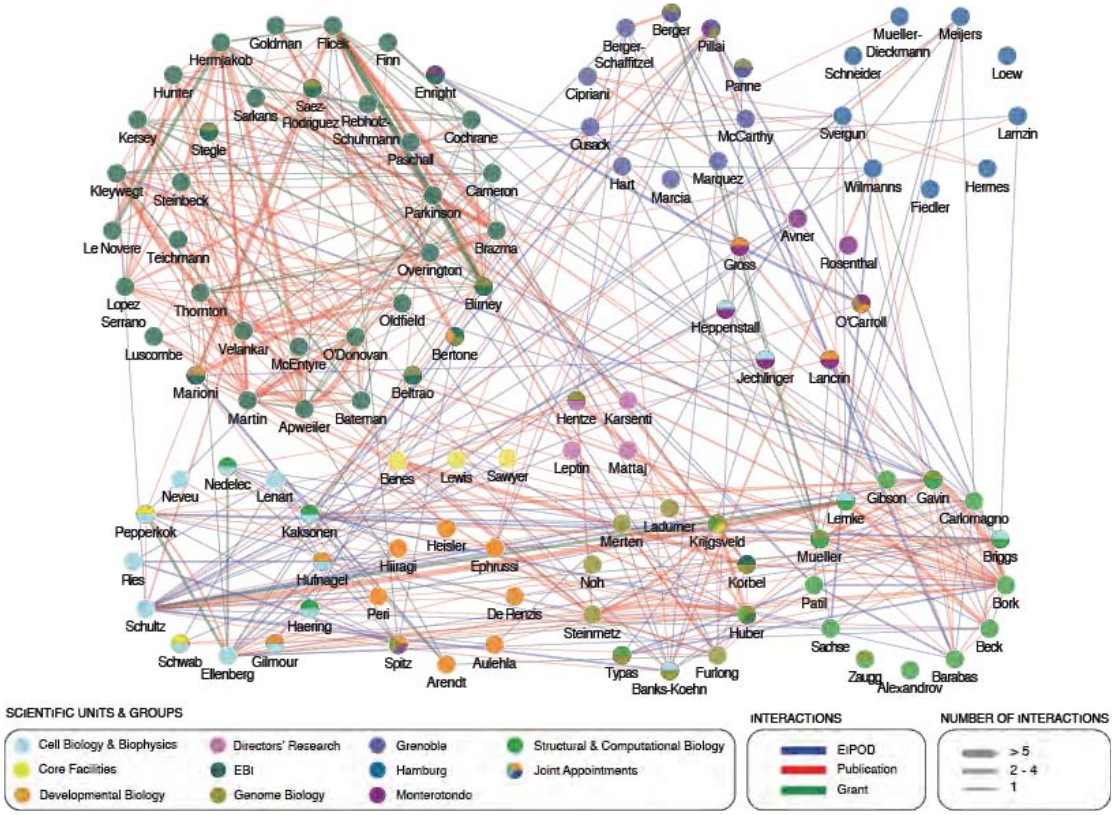
Figure B.3.1 Overview of academic backgrounds of EMBL group leaders.



However, the benefits of this breadth of available expertise can only be reaped if scientists work together across disciplines. EMBL's culture of collaboration is one of its core values and the degree to which collaboration is practised distinguishes EMBL from many other research institutions. By pooling approaches to tackle challenging biological problems that could not be addressed by one research group alone, investigators with different expertise give rise to an intricate network of internal collaborations spanning EMBL Units and outstations (Figure B.3.2). A large number of joint publications and research grants – 225 (over 12% of all publications) and 78 (over 15% of all grants), respectively – have stemmed from the cooperative efforts of interdisciplinary teams at EMBL over the first three years of the current Programme period. These numbers illustrate the success and effectiveness of such interdisciplinary collaborations and the extent to which research carried out at EMBL benefits from these interactions. This collaborative spirit not only extends throughout EMBL but also continues beyond the

Laboratory. A large majority of our publications (75% in 2012-2014) are jointly authored by researchers from outside EMBL, mostly from our member states.

Figure B.3.2 Network of internal collaborations between EMBL groups 2012-2014 as indicated by shared grants (78), publications (225) and EIPOD (95) fellows.



In addition to bilateral collaboration between groups, there have been a number of bigger, interdisciplinary projects that have involved several groups across multiple Units and have had a big impact on various fields. Recent examples include the integrated structural cell biology studies of the nuclear pore complex (Section B.1.1.1.4) and the endocytosis machinery (Section B.1.1.1.3), the detailed structural and temporal systems-based description of the simple cell of Mycoplasma pneumonia (Section B.1.1.1.6), the more complex disease agent Mycobacterium tuberculosis and the thermophilic fungus Chaetomium thermophilum (Section B.2.4.1). EMBL also participates and coordinates aspects of major international interdisciplinary projects such as ENCODE, the encyclopaedia of functional elements in the human genome (B.1.4.5), the analysis of data and specimens collected by the Tara Oceans expedition (Box B.2.1), whose first data analyses were recently published, and the Pan Cancer Analysis of Whole Genomes Analysis (PCAWG, Box B.2.4), a very ambitious attempt to comprehensively mine all the information in all sequenced cancer genomes to look for commonalities across cancer types. We will continue to encourage these efforts in future and aim at developing ways to promote further large-scale interdisciplinary projects.

To promote and nurture this interdisciplinary collaboration, EMBL has developed a variety of instruments. These range from *ad hoc* bottom-up initiatives, such as thematic retreats, journal clubs, mailing lists and meetings, to more formalised, institutional measures. Among the latter are joint faculty appointments that currently associate 34 group and team leaders (from a total of 100) with more

than one Research Unit, the EMBL Interdisciplinary Postdoc Programme (EIPOD; Section D.1.2), the EMBL Centres, and the Bio-IT Project, described later in this section.

3.1 EMBL Centres

The EMBL Centres are horizontal structures that concentrate and maintain know-how in specific topics and techniques, and make it available to user communities distributed throughout the Laboratory. Their overarching purpose is to facilitate collaboration and provide support, advice and training in interdisciplinary areas relevant to researchers within multiple EMBL Units. Through a number of activities, they encourage networking and facilitate the exchange of information as well as the sharing of resources. They also provide a platform through which EMBL scientists can interact with like-minded external communities. It is in the nature of such activities that the spectrum of Centres required by EMBL researchers will change over time and turnover has indeed occurred during the nine-year period in which Centres have been in place at EMBL. Indeed, three of the four existing Centres were established during the current EMBL Programme.

It comes as no surprise that the focus of most existing EMBL Centres lies in computational methods. In view of the vast amounts of data being generated, computational tools are essential, and commonly used by scientists all over EMBL to analyse and extract useful information from their experimental datasets. During the current Programme, the Centre for Computational Analysis, initially established in 2007, was found to be too broad in scope, so was discontinued and replaced by three more specialised computational Centres, namely for Statistical Data Analysis, Biomolecular Network Analysis, and Biological Modelling. By providing specific competence in computational approaches, these recently established EMBL Centres respond to a strong need for guidance and support, particularly for experimental researchers who may have limited skills in modelling and bioinformatics.

The expertise of the three computational EMBL Centres is highly complementary. Together they support the complete pipeline for large-scale data analysis, from ensuring statistical robustness to quantitative modelling of complex biological systems. The three Centres interact closely and to a large extent coordinate their activities, for example by organising joint training events. All three Centres also collaborate with other initiatives, most importantly the Bio-IT Project (Section B.3.2), which increases the Centres' outreach by facilitating information exchange and providing access to common resources.

All three computational Centres have proven to be extremely valuable resources and there is significant demand for their expertise. Therefore, they will be continued in the next Indicative Scheme. In addition, the need for a fourth computational Centre has been identified and thus the Centre for Integrative Structural Modelling will be established during the next Programme. This will specialise in cutting-edge computational methods for integrated structure analysis, an emerging approach that combines the strength of various structure determination techniques (X-ray crystallography, small angle X-ray scattering (SAXS), small angle neutron scattering (SANS), nuclear magnetic resonance (NMR), various high-resolution electron microscopy (EM) methods and non-traditional structural analysis techniques; Section B.2.4.1) to gain functional

understanding of molecular complexes. Although its services will be accessible to all EMBL scientists, the Centre for Integrative Structural Modelling will be of particular value for EMBL's structural biology outstations in Hamburg and Grenoble and the Structural and Computational Biology Unit in Heidelberg.

The only experimental Centre that EMBL currently operates is the Centre for Chemical Biology. This Centre was established in 2010 to provide an intellectual home to a significant number of chemistry groups dispersed throughout EMBL's Research Units. The Centre has been very successful in promoting the use of chemistry tools throughout EMBL and raising the visibility of chemical biology research at EMBL and will thus also be continued into the next Programme period.

The following sections provide a more detailed overview of the EMBL Centres' activities, highlight their successes over the first three years of the current Programme, and outline their plans for the future.

3.1.1 Centre for Chemical Biology

Over the past 10 years, chemical biology has become a significant and productive research area at EMBL, with currently seven research groups – five experimental and two computational – focusing on this area. As chemical biology is an important enabling technology for many areas of life science research, a major theme of the chemistry groups is the development of new tools and technologies that are made available first to researchers throughout EMBL and then more generally. Together with the Chemical Biology Core Facility (CBCF, see Section C.3.7) the experimental groups provide screening technologies, small molecules for manipulating biological events in intact cells and tissues, fluorescent reporters for monitoring enzyme activities and dynamic changes in molecule locations and, since late 2014, expertise in medicinal chemistry through a new dedicated Core Facility laboratory. In addition, there is a broad spectrum of chemistry expertise available, including lipid and peptide chemistry, NMR analysis of small molecules, mass spectrometry of small molecule–protein interactions, phosphatases and the development and use of artificial amino acids. The cheminformatic groups at EMBL-EBI run major databases for chemical biology: ChEMBL for data relevant to drug discovery and ChEBI and Metabolights for small molecules of biological interest.

To integrate chemistry into EMBL's research portfolio and ensure regular exchange with the biology groups, the chemistry groups are not assembled in a single Research Unit but dispersed across EMBL. In 2010, the Centre for Chemical Biology was established as a horizontal structure that provides an intellectual home and a platform for exchange for these groups. At the same time, the Centre is also a contact point for others in EMBL's scientific community who are eager to learn about how to apply chemical biology in their own research. A major focus of the Centre is therefore bringing new developments in chemical biology to the attention of the EMBL research community and providing training in various chemistry methods. In addition, the Centre also serves to increase the external visibility of EMBL's activity in chemical biology and has helped EMBL to acquire a reputation as a leading chemical biology research institute biology in Europe. This is, for example, illustrated by the fact that most groups affiliated with the Centre over the past three years have been awarded important prizes,

including the Friedmund Neumann Prize of the Schering Foundation, the Tetrahedron Prize, the KNIME Award, the MRN Innovation Prize and the Heidelberg Molecular Life Science Award. Groups from the Centre, in conjunction with the EMBL International Centre for Advanced Training (EICAT; Section D), organise the biannual EMBO Conference on Chemical Biology, which is widely regarded as the best international conference in the field.

The Centre for Chemical Biology operates without dedicated staff and depends entirely on the initiative and voluntary support of the various chemistry research groups. The Centre budget is used to organise retreats with Centre members and other EMBL researchers. Dedicated chemistry laboratory visitor space has also been created as described below.

Backward look and highlights 2012–2014

Over the first three years of the current Indicative Scheme, the Centre for Chemical Biology organised numerous internal and external training activities. Various practical courses on experimental chemical biology, screening techniques and computational aspects of high-throughput screening have taken place at EMBL. The Centre also organises a regular chemical biology retreat. The retreat is open to all EMBL group leaders and provides an opportunity to learn about new scientific developments and discuss how chemistry tools might help tackle current biological research challenges. In addition, the groups associated with the Centre gave a number of lectures on new developments in the field of chemical biology across EMBL sites.

The Centre for Chemical Biology has also contributed to the organisation of the biannual Chemical Biology conference in the EMBO Conference series. This prestigious meeting regularly attracts more than 300 participants, including leaders in the chemical biology field, to EMBL Heidelberg. Preparations for the next meeting in 2016 are already ongoing.

In the course of the current Indicative Scheme, a dedicated chemical biology visitor space has been set up at EMBL to address the growing demand for synthetic chemistry throughout the Laboratory. This has been heavily used by mostly collaborative projects, many of which are in the context of the EIPOD Programme (Section D.1.2). More recently, the new medicinal chemistry lab of the CBCF has also been hosting EMBL PhD students carrying out chemistry projects.

Future plans 2017–2021

During the next Indicative Scheme, six of the seven current chemical biology groups will come to the end of their time at EMBL and future activities of the Centre will strongly depend on the groups and expertise that EMBL recruits. New chemistry-oriented groups will need to be scientifically integrated and motivated to help the new generation of EMBL staff to embrace chemical biology.

Assuming continued support by chemistry groups across EMBL, the Centre proposes the following activities for the period of 2017–2021:

- **EMBL chemical biology retreat.** The Centre will continue to organise this one-day meeting for group leaders every 18 months. The aim is to foster the

dissemination of new techniques to EMBL group leaders and keep the Centre members up-to-date on pressing needs of the biologists.

- **External training.** Symposia and courses will be organised by Centre members for external audiences. The Centre will also continue to contribute to the organisation of the biannual EMBO Conference on Chemical Biology.

- **Internal training in chemical biology.** The Centre will provide lectures and courses on i) using the latest techniques to manipulate and monitor intracellular signalling events; ii) performing fluorescent labelling; iii) creating and using compound screens; iv) studying ligand binding by NMR; and v) using mass spectroscopy to analyse artificial biomolecules from cell sources. Finally, the bioinformatics groups of the Centre will coordinate training in software and databases.

In addition to these activities, the Centre will promote the opportunities provided by the new medicinal chemistry laboratory of the CBCF throughout EMBL. Moreover, a closer collaboration with the Protein Expression and Purification Core Facility (Section C.3.2) will be sought to implement tools from the chemistry labs for advanced protein engineering such as intein-based protein synthesis, unnatural amino-acid incorporation, and sortase (N-terminal) tagging of proteins for broad use by the EMBL community.

3.1.2 Centre for Statistical Data Analysis

Statistical analysis is a bottleneck in many research projects in molecular biology. The aim is to draw maximal insight from complex and often high-dimensional data (discovery), while making sure the conclusions are robust against systematic and stochastic sources of noise (confirmatory data analysis).

Technological progress is one of the main drivers of scientific advances in biology. The datasets are becoming larger, more complex, and often include multiple data types each with its own 'normalisation' and sampling issues. It is therefore crucial that biological researchers are able to use the most appropriate tools and methods for the analysis of their data.

The Centre for Statistical Data Analysis (CSDA) is a platform to disseminate statistical expertise and guidance throughout EMBL. It helps EMBL scientists to use adequate statistical methods for their specific technological or biological research questions. The CSDA focuses on the needs of smaller and more experimentally oriented groups, particularly groups whose primary experience is not in (high-throughput) data analysis. Its activities fall into two major categories: training and consulting.

Through its training activities, the CSDA enables biological researchers to perform many aspects of data analysis themselves. It fills skill gaps in underlying theory and computer programming that researchers require to select and execute the most suitable analysis method and to interpret the resulting output. The courses cover general statistical data analysis as well as statistical programming, mainly using the R and Bioconductor toolkits.

Consulting services provided by the Centre aim to solve researchers' specific statistical problems, which often require a specialist overview of recent advances in statistical research and good practice.

The CSDA also participates in collaborative research projects. Here, assistance is mainly provided on statistical applications that are broadly related to the analysis of high-throughput data including RNA-seq, ChIP-seq, 4C/Hi-C and DNA-seq.

All offerings of the Centre are open to all scientists working at EMBL.

Backward look and highlights 2012-2014

From its creation in February 2013 until the end of 2014, the CSDA has assisted 119 scientists in a total of 214 sessions ranging from 30 minutes to 3 hours (on average, 2 sessions per week). The consultations have dealt with a wide spectrum of problems, including responses to referees in peer review, the choice of suitable visualisations and summary statistics, experimental design, interpretation and usage of appropriate statistical tests and the interpretation and usage of bioinformatics tools for the analysis of high-throughput data.

The CSDA has offered compact courses open to EMBL researchers at all sites. In 2013 and 2014, the CSDA trained 139 students in seven courses (course duration varied from 1.5 to 5 days) at all levels, from ‘Basic R and Graphics’ for beginners, to advanced topics including end-to-end workflows for the analysis of RNA-seq data. The Centre ran a one-week course on statistical data analysis and the analysis of microarrays in Monterotondo in June 2014 and a four-day EMBO course on statistical bioinformatics, thereby reaching out to external scientists. In March 2014, the three computational Centres together with the Bio-IT Project (Section B.3.2) organised a whole week of training consisting of complementary courses in computational methods spanning the different areas of expertise of the Centres.

Additional activities of the CSDA include a biweekly Machine Learning Book Club with approximately 20 active members and a mailing list that allows more than 300 members to exchange information about topics around statistical data analysis. Moreover, the CSDA supports EMBL group and team leaders in the recruitment of scientific staff with statistical expertise.

Web resources hosted on the Bio-IT Portal (Section B.3.2) complement the information exchange catalysed by the CSDA. Teaching materials used in the courses taught are available for download by EMBL scientists and the external scientific community.

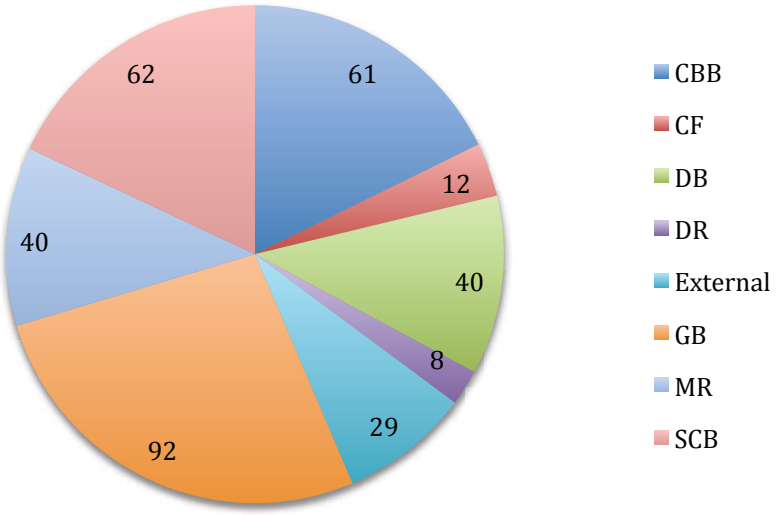


Figure B.3.3 The chart shows the number of scientists trained by the CSDA since its inception in 2013 and the distribution of users according to EMBL Research Units.

Future plans 2017–2021

The CSDA will continue to provide essential support for statistical data analysis for all scientists at EMBL through both consulting and training. This will continue to be relevant; a survey in mid-2013 revealed that 95% of the scientists who interacted with the Centre are likely to recommend it to their colleagues.

A teaching focus of the Centre will be on the dissemination of fast and efficient data manipulation methods via hands-on training and template workflows for common data analysis tasks in those high-throughput technologies that are of widest interest to EMBL researchers. Besides RNA-seq, ChIP-seq and their many variations, we anticipate a demand for data analysis tools in proteomics, single-cell sequencing analysis, third-generation sequencing and high-throughput phenotyping (Section B.2.2). This list will change over time and we will therefore monitor the demand closely.

The CSDA plans to contribute a regular column to *The EMBO Journal* on aspects of the statistical analysis of biological data and will continue to offer an annual EMBO course on statistical bioinformatics.

Computing is essential for data acquisition, analysis, exploration and result reporting and as information technologies rapidly evolve (e.g. web-based visualisation, pervasive computing, mobile devices, cloud computing) so will the modes by which we perform these tasks. The Centre will support and provide training in the effective adaptation of suitable innovations.

By maintaining and deepening its interactions with EMBL research groups working on statistical methods and bioinformatics of high-throughput data, the CSDA intends to always disseminate the most recent methods and practical experience through, for example, joint retreats, visits and small research projects. The CSDA will also intensify its efforts to provide services beyond EMBL

Heidelberg. In particular, it aims to support EMBL Monterotondo in increasing its on-site bioinformatics expertise.

3.1.3 Centre for Biomolecular Network Analysis

The construction and analysis of biological networks allows scientists to explore and integrate heterogeneous datasets resulting from the application of large-scale methods. As molecular parts are displayed in the context of their interactions, it becomes possible to predict and manipulate cellular systems *in silico.* This opens up new possibilities such as the formulation of meaningful new hypotheses regarding gene and protein function, the identification of novel protein complexes and the prediction of the crosstalk within and between signalling pathways. However, data integration and the analysis of complex biomolecular networks require extensive computational expertise.

The EMBL Centre for Biomolecular Network Analysis (CBNA) was established in April 2013 to develop new methods for the representation and visualisation of complex datasets, and to provide expertise and guidance in the field of biological network analysis throughout the Laboratory. Bioinformaticians across EMBL Research Units use the CBNA as a platform to share resources, exchange know-how, and learn about advances and new approaches in computational network analysis. In addition, experimentalists with limited computational expertise turn to the CBNA when they need help to integrate and analyse large-scale datasets or to place results obtained in small-scale mechanistic experiments in the context of existing networks.

The CBNA provides an open helpdesk, offers training courses, participates in collaborative research projects, and fosters information exchange and discussion in the field of biological network analysis.

Backward look and highlights 2012–2014

Since the Centre's inception, the CBNA helpdesk has assisted 51 scientists in a total of 145 sessions. The service has provided direction and made available tools for visualisation, integration, processing, and comparative analysis of both quantitative and qualitative datasets resulting from a variety of experimental methods (e.g. microarrays, RNA-seq, Chip-seq, mass spectrometry, protein arrays, yeast two-hybrid assays, flow cytometry, metatranscriptomics, etc.). The Centre has also been involved in six long-term collaborative research projects as a partner for systematic data analysis.

The CBNA offers training in the field of data integration and network biology for EMBL researchers across sites and at all career levels. Over the past 18 months, more than 100 participants have been trained in seven courses, spanning all levels of expertise from beginners all the way through advanced to developer levels. In 2014, the CBNA participated in the training week organised in collaboration with the other computational Centres and the Bio-IT Project (Section B.3.2), ran a collaborative course on advanced network analysis and modelling, and a developer-level advanced course for the open source visualisation software platform Cytoscape, which was open to external as well as internal scientists.

Scientific exchange is promoted by the Centre through the Network Biology Mailing List, currently hosting more than 200 members, and the Network Biology Club, with close to 30 active members who participate in monthly journal and book club meetings. Web resources, hosted on the Bio-IT Portal (Section B.3.2), complement the information exchange, helpdesk, and training activities by providing course announcements, training materials, and up-to-date software and database lists.

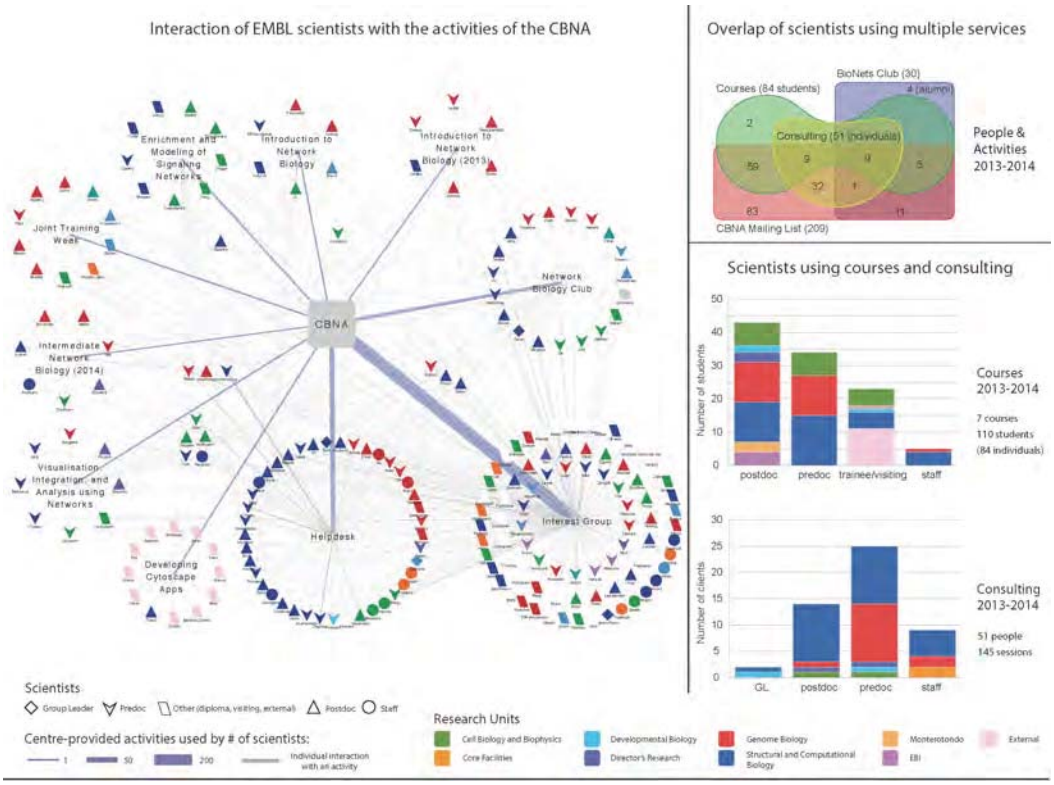


Figure B.3.4 Overview of the interactions of EMBL scientists with the CBNA since its inception in 2013.

Future plans 2017–2021

The CBNA will continue to carry out and further strengthen its activities over the period of the next EMBL Indicative Scheme. Potential for new initiatives and future developments of the Centre is seen particularly in the areas of information exchange, centralised resource management, and the application of new computational methods.

Given that biomolecular network analysis is such a rapidly evolving field, the CBNA intends to create an online blog to ensure that scientists have immediate access to up-to-date information and new methods.

Reducing the workload on local machines and outsourcing to external clusters is another highly relevant issue for large-scale network analysis, which requires considerable computational power. For this reason, the Centre will implement a centralised resource with pre-installed packages, which allows the performance of relevant tasks in a simplified manner through web interfaces on the EMBL computing cluster.

Finally, the field of network biology is rapidly progressing into investigating the dynamics of cellular systems. The molecular kinetics with which biological networks assemble and dissociate to generate specific cellular responses can now be captured by experimental techniques such as affinity purification-selected reaction monitoring. The visualisation and analysis of the resulting datasets, however, is inherently difficult and represents a major challenge. Elucidating these dynamic networks will require new approaches that allow spatiotemporal enrichment and the integration of '-omics' resources. Novel *ad hoc* visualisation techniques are therefore under development and will continue to be pursued by the CBNA over the coming years. These new approaches range from simple animations, through interactive models, to more abstract representations.

To stay abreast of the developments in the field of network visualisation and analysis, the Centre will introduce, alongside its regular activities, a yearly workshop and an annual retreat. These events will bring together scientists, identify needs and merge expertise from Research Units across EMBL sites, promote fruitful discussion, and establish close ties with scientists involved in the development of relevant new methods outside EMBL.

3.1.4 Centre for Biological Modelling

Quantitative assessment of the abundance and activities of biomolecules is central to an increasing number of research projects at EMBL. The resulting datasets are providing novel insights into the function and dynamics of various biological systems. To facilitate the process of transforming data into biological insights, the Centre for Biological Modelling (CBM) supports EMBL researchers through training, consultation, collaboration and networking activities. As a result of the expertise present at EMBL, the Centre offers support for modelling at various scales – from molecular networks to whole cells to ensembles of cells. The emphasis is on facilitating development and the use of models founded on physical/chemical first principles, and on helping in the interpretation of simulation results to generate testable hypotheses and new mechanistic insights.

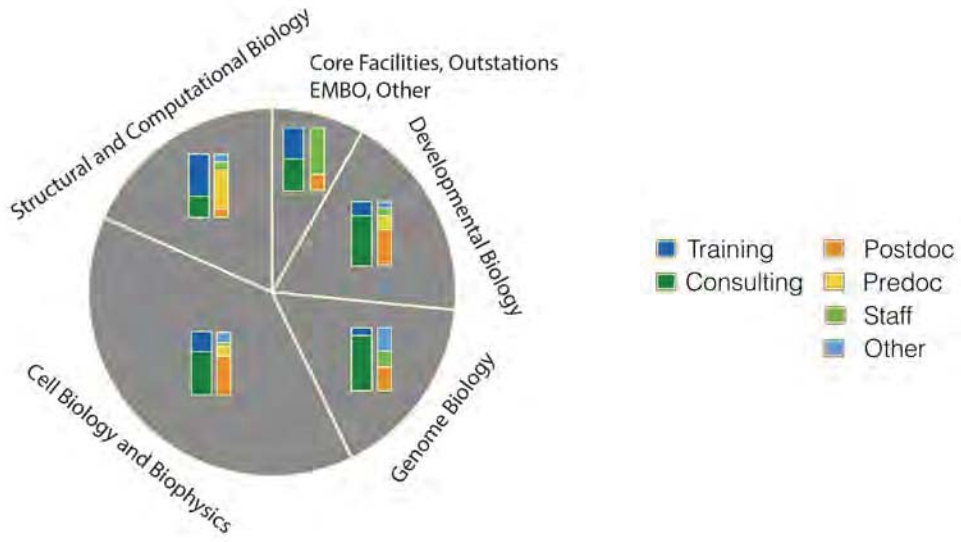
As modelling problems are often closely related with statistics and network analysis, the CBM works closely with the other two computational Centres and the Bio-IT Project (Section B.3.2) to enable efficient sharing of resources and to increase its outreach.

Backward look and highlights 2012-2014

The CBM became active in March 2013 and has since carried out a variety of training, consulting and networking activities. Training activities include introductory modelling courses based on MATLAB as well as contributions to external courses organised by EMBL scientists.

Two introductory courses on MATLAB were well received by postdoctoral and predoctoral fellows from various Units. The Centre's contribution to external courses such as an '*in silico* biology' course organised by EMBL-EBI, was also appreciated by the participants. The CBM's consulting and collaboration activities covered a wide range of topics, from symmetry breaking during development to microbial community dynamics. Overall, around 50 EMBL scientists have so far used the different levels of support offered by the CBM. The Centre continues to

seek further outreach through various networking activities and by directly approaching EMBL scientists.



Outreach of CBM activities during 2013-2014.

Figure B.3.5 The chart shows the distribution of CBM users according to EMBL Research Units and career stage.

Future plans 2017–2021

The current support activities of the Centre will be maintained with continuous revisions based on user feedback. To keep its activities in line with the evolving research at EMBL, the Centre will focus its support in three areas that, from a modelling perspective, are expected to gain momentum in the next Indicative Scheme period:

- **Emergent phenotypes in heterogeneous cell populations**. Rapid advances in single-cell analytical technologies are providing a high-resolution picture of heterogeneity in biological systems such as tumours or microbial populations. Modelling will be a powerful tool for understanding the phenotypic consequences of such heterogeneity.
- **Cell-state transitions**. The principles governing cell-state transitions, e.g. stem cell differentiation or in the context of a developing embryo, will be another area in which the CBM will expand its support for theory-based analysis.
- **Multi-scale modelling**. The CBM will strive to increase its expertise and toolbox in multi-scale modelling techniques to bridge between biological processes operating at different time-scales, e.g. translation, post-translational modifications and metabolic fluxes.

On the training front, the CBM, in consultation with EICAT (Section D), will offer a new training course in mathematical modelling during the second-year bioinformatics module of the core course for predoctoral fellows. By using the experience gained through its internal training activities, the Centre also plans to increase its contribution to external courses in various member states.

3.1.5 Centre for Integrative Structural Modelling

One of the pervasive themes in EMBL's research plans for the next Indicative Scheme is bridging the scales of biological organisation (Section B.2). An important aspect of these plans is the study of molecular complexes in action (Sections B.2.4.1 & B.2.4.2). Such complexes are dynamic ensembles of multiple, overlapping compositional and conformational states in the cell. A functional understanding of these ensembles requires a combination of various structure determination techniques such as NMR with SAXS, SANS and fluorescence resonance energy transfer (FRET) or X-ray crystallography with EM and cross-linking mass spectrometry (Section B.2.4.1 for examples). Also, the integration of these data with cellular imaging and '-omics' techniques will become increasingly important (Section B.2.1.2).

Although structure determination, quality control and validation procedures for these methods are well established, integrative structure determination is still in its infancy and far from being standardised. A variety of software solutions for data integration exist, however their application generally requires expert computational knowledge that goes beyond that of most of the predominantly experimental groups at EMBL. More importantly, integrative approaches are also prone to pitfalls that are not necessarily obvious to non-experts. As a consequence, there is a significant need for training and consultancy, bundling expertise, facilitating knowledge exchange across sites and maintaining know-how at EMBL.

To address these needs, EMBL will establish a new Centre for Integrative Structural Modelling (CISM), which will complement the existing three computational Centres by focusing on the following scientific challenges:

- **Quality control and validation**
  Due to the manifold approaches being applied and the present lack of cross-technology standards, integrated structure determination projects are challenging to implement and the results are often not straightforward to interpret. The CISM will provide *ad hoc* advice regarding which modelling approach to choose and how to interpret and validate structural models. Where appropriate, it may suggest additional experiments to independently validate models. The CISM will collaborate closely with the Centre for Statistical Data Analysis for the custom design of validation procedures. Collaboration with EMBL-EBI's Protein Data Bank in Europe (PDBe; Section C.1) deposition services and validation task force are envisaged to help position EMBL at the forefront of this rapidly developing field.
- **Bridging across length scales**
  To understand the architecture of many macromolecular assemblies, it is crucial to bridge across length scales by integrating data obtained from the same sample at various different resolution levels. The most established application is spatial restraint-assisted docking and fitting, e.g. of X-ray or NMR structures into EM maps or SAXS envelopes. One frequent task for the CISM will be to assist users in choosing and implementing the most suitable approaches given a certain data set.
- **Integration of '-omics' and imaging data with structural models**
  To understand the functional relevance of ensembles of structural states, integration with '-omics' and imaging data is needed in order to place structural models into the spatiotemporal context of the cell. In particular,

variables such as post-translational modifications, cell-type specific stoichiometries, concentration gradients imaged in live cells or structural diversity quantified by classification of images, might be projected into structural models. This aspect will become more important in the future and effective software solutions require further development. By building on existing expertise in the Structural and Computational Biology and Cell Biology and Biophysics Units regarding the integration of structural and imaging data and collaborating with the Centre for Biological Network Analysis for the integration of **'-**omics' data into structural models, the CISM will support users in addressing this challenge.

Like the other computational Centres the CISM will be operated by one dedicated staff member, who will engage in a combination of the following activities that will be available for researchers at all EMBL sites:

- **Consulting**
  The CISM will keep track of the diverse existing software solutions and provide EMBL scientists with advice and training in the context of individual consultancy sessions. This activity will facilitate knowledge exchange across sites and increase the effectiveness of structure determination at EMBL.
- **Collaborations**
  As capacity allows, the CISM will engage in scientific collaborations with EMBL groups. In these cases the service provided can go beyond consulting and, for example, include the collaborative development of novel modelling approaches.
- **Teaching, outreach and community building**
  The Centre will provide training for structural modelling by offering courses and online tutorials, by linking people with the required know-how, inviting external speakers and tutors and organising meetings and retreats within the interest group. The training courses organised will be complementary to existing courses offered by EMBL Hamburg and Grenoble in the area of integrated structural biology. In addition, depending on the interest among the EMBL community, a biannual integrative modelling course and a biannual retreat will be organised. It is also envisaged that a structural modelling club will be formed, which could regularly meet through videoconferences involving interested researchers at all sites. Know-how and software will be shared online, taking advantage of the already established Bio-IT Portal (Section B.3.2).
- **Software development**
  Existing software solutions for integrative modelling are often poorly integrated into established structure determination software packages and are challenging to use for non-experts. For this reason, several EMBL groups are developing their own custom-made software tools. The Centre plans to catalyse this activity by providing a platform that brings together developers with complementary expertise and interests. The CISM will also, where applicable and desired, help to disseminate the most valuable algorithmic solutions to a broader user community.

3.2 Bio-IT Project

Bioinformatics in its broadest sense – that is, using computers to work with biological data – is integral to science at EMBL. The increasing use of high-throughput experimental approaches is driving a growing need for skills in diverse, integrative, computational data analysis, and hence bioinformatics skills.

Whereas EMBL-EBI is entirely dedicated to computational research, at the other EMBL sites bioinformatics expertise is distributed across groups and Research Units. In particular, EMBL Heidelberg has a large number of dispersed bioinformatics users. This was highlighted in a 2012 survey that found that 40% of EMBL Heidelberg staff spend half or more of their time using computers to collect, process, or analyse biological data, an increase from the 34% reported in a similar survey in 2009. Some groups reported that bioinformatics and computational methods form 50% or more of their research activities, although most groups use it as an essential support for predominantly bench research. Notably, 74% of EMBL group leaders surveyed in 2012 predicted further growth in bioinformatics activities in their groups.

EMBL Heidelberg's bioinformatics users share many common challenges and needs, use the same tools, and carry out similar analyses, yet are dispersed across the Laboratory. In 2010, the Bio-IT Project was initiated at EMBL Heidelberg to give bioinformatics users a context to meet, identify common problems, challenges and goals, and to develop ways to address them collaboratively to minimise redundant activities, better utilise the potential of this growing discipline, and increase the effectiveness and efficiency of their research using these tools.

The Project depends on voluntary collaboration from bioinformaticians across EMBL Heidelberg, together with technical support via one part-time staff member. Members of the Project work closely with the three existing computational EMBL Centres (Section B.3.1). Although the Bio-IT Project focuses on EMBL Heidelberg, it has helped the computational Centres to connect with other EMBL sites, resulting in several visits to EMBL Monterotondo for consulting and training activities. The Bio-IT Project also draws on the computational expertise available at EMBL-EBI, for example by inviting EBI scientists as trainers for courses held in Heidelberg.

Backward look and highlights 2012–2014

Building a community amongst the EMBL Heidelberg bioinformatics users is key to the Project. Developing and improving communication between users across EMBL Heidelberg has identified common interests and goals, which has in turn facilitated the initiation and deployment of the following shared community resources:

- An internal web resource to disseminate bioinformatics-related information (the Bio-IT Portal)
- A repository of software for easy deployment on Linux computers (the Bioinformatics Computational Resource)
- An EMBL server for the remote, version-control, software-development tool git (git@embl)

- Regularly updated centralised versions of key bioinformatics databases
- Training courses and associated infrastructure (for registration, advertising, etc.)

Community collaboration is promoted via regular events, in particular bimonthly 'Bio-IT Taskforce' meetings with participants from EMBL Centres, Research Units, and Core Facilities across EMBL Heidelberg. Further opportunities for bioinformatics users to meet and exchange ideas and experience are provided at other networking events throughout the year. Bio-IT Project members also organise and teach numerous courses on computational biology topics for EMBL scientists. Additionally, several members of the Bio-IT Project volunteer to organise local bioinformatics networking activities beyond EMBL. This has helped develop stronger links between bioinformaticians working in private and public sector institutions across Heidelberg.

Future plans 2017–2021

The Bio-IT Project will continue to develop and maintain an active community of bioinformatics users at EMBL Heidelberg, and will expand to accommodate the expected increases in the number of bioinformaticians and the level of bioinformatics activity at the Laboratory.

Training courses and events have proven an excellent way to increase the bioinformatics skill of EMBL scientists, and create a community. If increased resources are available in the new Indicative Scheme, the Bio-IT Project will expand and develop the training activities it delivers and supports.

Currently, the web-based resources of the Bio-IT Project are available only behind the EMBL firewall. In the future, to increase cooperation between sites, we plan to make selected content from current internal pages available outside the firewall, to showcase Bio-IT and other related activities to others.

The Bio-IT Project does not currently offer consulting services to EMBL Heidelberg groups. If more resources become available, the Project plans to develop such services, for example in the realms of sequence analysis and pipeline development.

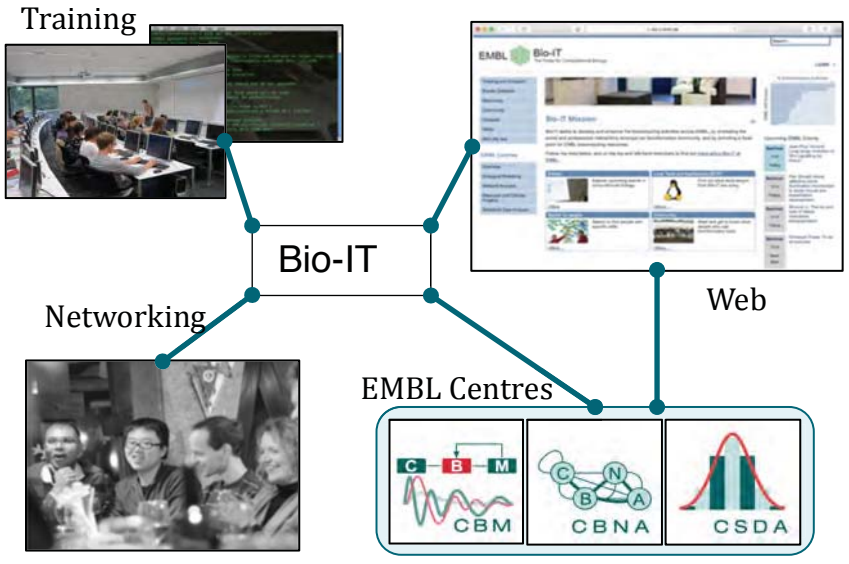


Figure B.3.6 The Bio-IT Project offers bioinformatics support, training and resources for scientists across EMBL Heidelberg.

C. Infrastructure, Services and Support

EMBL Mission: Providing world-class research infrastructures and services

1. Bioinformatics services and databases

1.1 Introduction

1.1.1 The EMBL-EBI service mission

One of EMBL's missions is the provision of services and infrastructures to the biomedical science community. An important aspect of this is the provision of data resources through EMBL's European Bioinformatics Institute. EMBL-EBI is the world's leading source of biological and biomolecular data. Its core mission is to enable life science research and its translation to medicine, agriculture, industry, and society by providing biological data, information and knowledge. EMBL-EBI's services include the provision of biological databases and the tools to explore them, and its user community includes commercial and academic researchers throughout Europe and the world, including ever-increasing usage by medical professionals and the healthcare industry. At a finer scale EMBL-EBI's service mission has three components:

1. To provide freely available data and bioinformatics services to all facets of the scientific community
2. To provide training in use of our databases and tools to scientists at all levels both in academia and industry
3. To coordinate and participate as an integral part of ELIXIR, the European Research Infrastructure for biological data, the provision of biological data throughout Europe

In addition to its service teams, EMBL-EBI also conducts excellent research in bioinformatics as described in Section B.2.1 of this Programme. The research activities at EMBL-EBI contribute synergistically to its service mission through feedback on its tools; by creating new services and analysis methods; and by giving the services early insight into scientific trends and advances.

EMBL-EBI's service mission will continue undiminished during the 2017-2021 EMBL Programme, but this constancy of mission should not hide the challenge of ensuring that the services will continuously adapt to changes in science and technology. These challenges include:

- **Exponential increases in data volumes** that challenge the resource teams to increase their efficiency and challenge the physical resources to handle the data.
- **New data types**, for instance in human variation, require new data resources.
- The **diversity of data**, particularly data produced by high-throughput (omics) technologies, means that the resources must integrate many data types and facilitate analysis to add value.
- The increasing **diversity of users**, including those from resource-poor geographic locations, as well as large numbers of medical and healthcare professionals whose needs are applied rather than research-driven.
- **Changes in information technology** also affect the services. Increasingly, users do not wish to install an expensive computational infrastructure for occasional data analyses and datasets are becoming too large to download for local use. EMBL-EBI has therefore established a cloud infrastructure that allows users to implement their own virtual environment to make use of the data resources. This solves a major problem for many users but has cost implications for the required computational infrastructure.

EMBL-EBI will continue to respond to these and other emerging changes in its service provision. The goal is not just to keep up with science by recording its findings, but also to support exploration by looking over the horizon to 'the next question', thereby contributing to tomorrow's research as much as today's.

1.1.2 Current data resources

EMBL-EBI's data resources cover the entire range of biological sciences from DNA sequences to proteins, chemicals, structures, systems, pathways, and literature. Its suite of resources is not static: the data resources are regularly expanded as data volumes evolve and when new technologies create new data types. Figure C.1.1 lists EMBL-EBI's major data resources in each category along the arrow from genes to systems. Data resources can also become obsolete and when this happens, they are closed down.

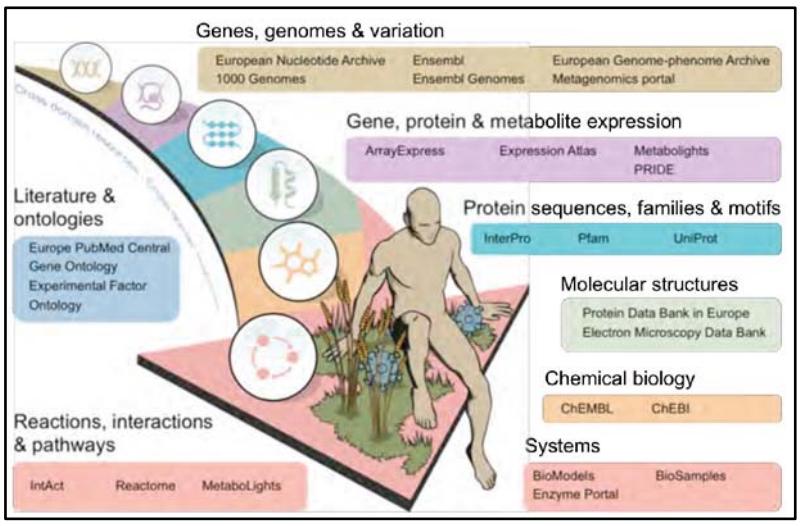
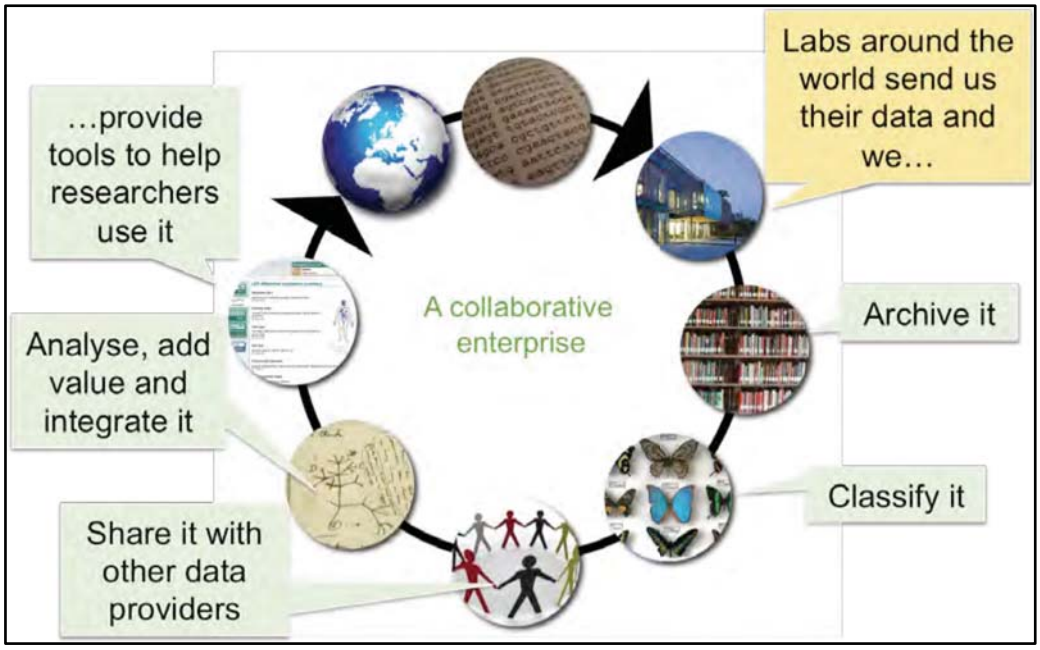


Figure C.1.1 EMBL-EBI data resources, from genes to systems.

1.1.3 User focus

The driver for our data resources activities is service to our users. The data in our resources is submitted by users, and we add value by archiving the data, sharing it with other data providers, and providing tools for analysis and integration. We continuously seek feedback from users to improve our services and to identify new data types that might require establishment of new resources. This process is diagrammed in Figure C.1.2.

Figure C.1.2 EMBL-EBI data resources cycle.



1.1.4 International activities

Biological research is an international enterprise, and like all EMBL activities the bioinformatics services take place in an international context (Section F). Many of the data resources are managed cooperatively, with partners with whom we share the tasks of accepting data submissions, data quality assurance, and data curation. The European Nucleotide Archive, for example, shares its core data with GenBank in the US and the DNA Databank of Japan, and PDBe shares data with the worldwide Protein Databank Consortium that includes three other partners. Additionally, some of our resources are jointly funded as international projects, with sites in two or three countries. Examples are UniProt and the Gene Ontology. Our users, too, are worldwide: our web pages are accessed millions of time per day from around the world.

EMBL-EBI also hosts the management hub for ELIXIR, the European Research Infrastructure for biological data and serves as an ELIXIR node. ELIXIR is tasked with coordinating the bioinformatics data infrastructure across Europe, and will make extensive use of EMBL-EBI's expertise and infrastructure to undertake this task (Section C.1.3.6).

1.2 Backward look and highlights 2012 - 2014

Over the course of the current indicative scheme EMBL-EBI's physical infrastructure and data resources have undergone substantial changes, which have been accompanied by

essential growth of staff and administrative restructuring to streamline management of the increasingly diverse portfolio of data resources. As foreseen in the last EMBL Programme, technological changes have continued to put previously difficult-to-collect data within the reach of most scientists (e.g. transcriptomes, proteomes, metabolomes) and we have where necessary established new resources to handle newly available data types. Changes in sequencing technology resulting in decreased costs and increased throughput have also resulted in exponential increases in DNA and RNA data volumes that we have responded to, and continue to plan for.

1.2.1 Improving services

In the EMBL Programme 2012-2016 we described the then nascent EB-eye search engine that allows users to easily search all data resources using a single query. EB-eye is now fully integrated and was the first step of a wide-ranging effort to improve the usability of our services. We have recruited two usability experts who have undertaken user testing to systematically provide feedback that aids many of our resources to redesign their home pages. In addition, the EMBL-EBI website was redesigned to provide a unified 'feel' to all users, regardless of which resource they access. Underneath these visual improvements we also revamped the content management system for all web pages: this has improved the user experience through increased responsiveness of the pages, and also enhances management of the web resources through use of a single unified system for all teams.

1.2.2 Data resources: growth, diversity, and usage

A major indicator of the success of our data resources is the ever-increasing volume of submissions by scientists: encouraged by our own outreach efforts and our continued effort to facilitate the process of data submission; by the requirement of funders that data is submitted to public repositories; and by advances in technology that have increased the volumes of data being produced by the world's scientists. Figure C.1.3 shows growth of six major data resources over time. In all cases growth is substantial, and for some resources exponential.

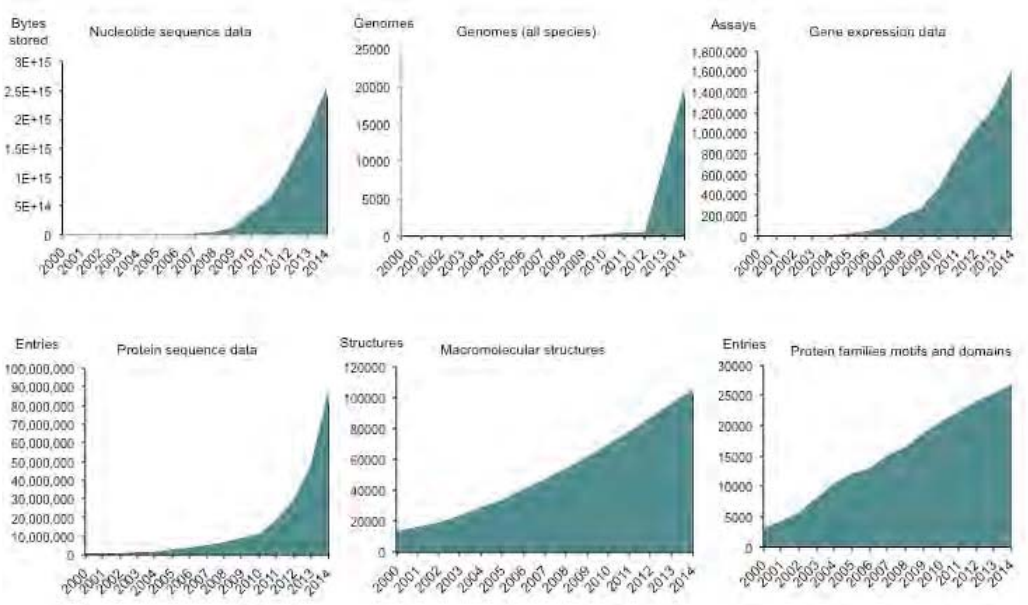
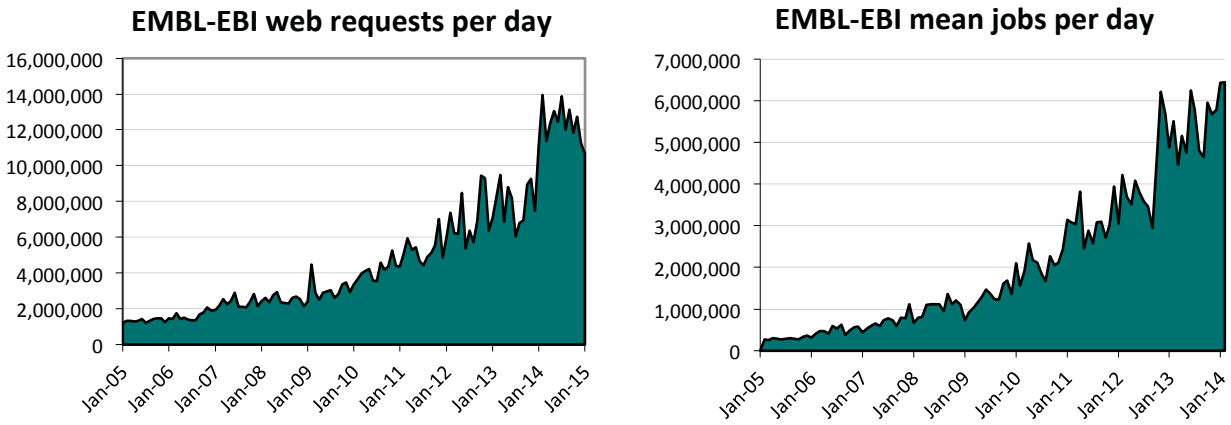


Figure C.1.3 Growth of EMBL-EBI data resources

The growth of our resources, combined with the improvements in the experience of our users, are reflected in the continued increase in usage of our resources, summarized in Figure C.1.4 as requests per day and jobs per day to the entirety of EMBL-EBI managed websites. On an average weekday at the end of 2014 we now see in excess of 11 million requests to our websites.

Figure C.1.4 Usage of EMBL-EBI resources, as measured by requests to the website (the web browser of a user requests fetching a file from the EMBL-EBI web server) and computational jobs (sequence searches and alignments, InterProScan, and many other tools) executed per day, continue to rise.



1.2.3 Infrastructure and staff

Infrastructure

EMBL-EBI successfully bid for funding from the UK government's Large Facilities Capital Fund and was awarded €100 million over a 10-year period starting in 2011. One third of this funding was earmarked to construct a new EMBL-EBI building on the Wellcome Genome Campus and the remainder for computational infrastructure — data centre space, computer cores, and storage — over 10 years.

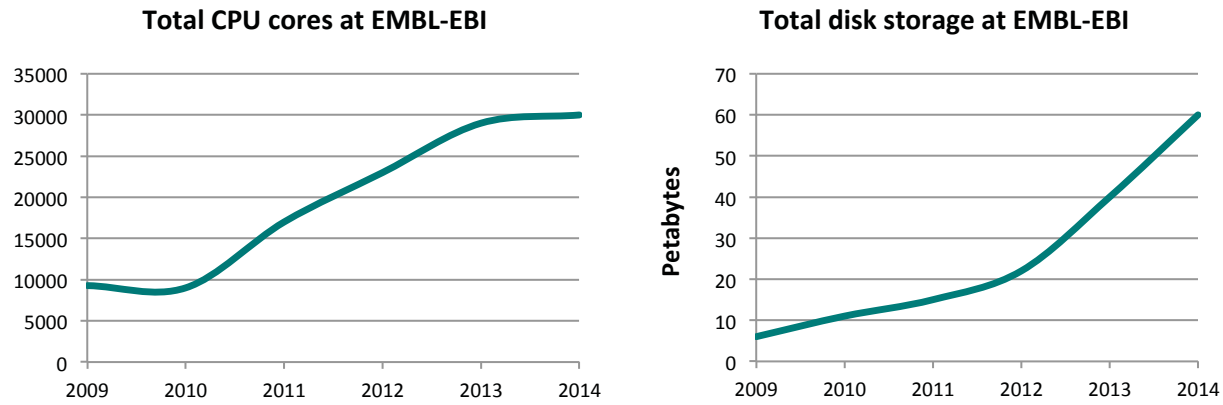
The EMBL-EBI South Building was formally opened in 2012. It hosts the ELIXIR hub offices, EMBL-EBI's industry programme, the new Innovation and Translation Suite (described further in Section E.4), as well as the web services and internal technical services (IT) teams, some research teams, and a number of data resource teams. The move of many personnel from the EMBL-EBI main buildings and from Portakabins has resulted in improved working conditions for all EMBL-EBI staff. The new building provides EMBL-EBI and ELIXIR with space to grow for the remainder of this decade.

In the last EMBL Programme we described plans to establish two data centres in London that would become the primary nodes for our services and would serve as a complete backup for all of EMBL-EBI's data resources. These two data centres were established in 2010 and successfully ran all of EMBL-EBI's outward facing web services for almost four years. The initial contract lasted until the end of 2014. Early in 2014, we sought bids from our then current host as well as other vendors to provide data centre space for the next four years. As a result of this process we chose a different vendor and have successfully moved the two self-contained facilities in two different London data centres into two self-contained facilities in one location, also in London. We will continue to serve

users from this London data centre while using the data centre on the Wellcome Genome Campus both as a backup and for development and updating of existing services.

The computational hardware in our data centres is continually growing to provide storage and analytical power for users, as shown in Figure C.1.5.

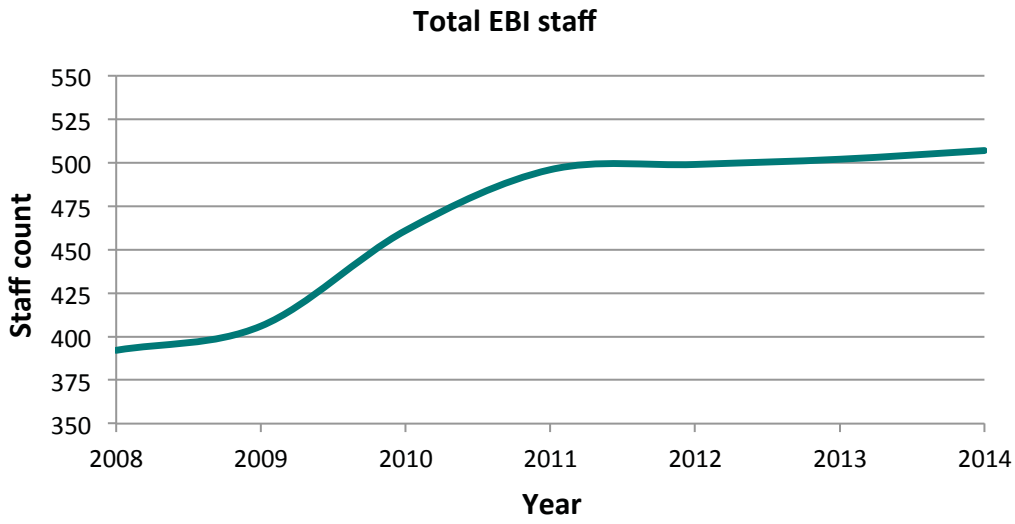
Figure C.1.5 Expansion of EMBL-EBI computational hardware 2009-2014.



Staff

Although our data resources and data infrastructure are growing rapidly, staff numbers have been relatively stable since the end of 2010: we are handling more data, and managing more data resources, with only a slow increase in staff count. However, we expect that the increasing data production and demand on bioinformatics services will lead to further growth of the EMBL-EBI staff numbers over the next programme period. Figure C.1.6 shows the total number of staff at EMBL-EBI over the past five years.

Figure C.1.6 EMBL-EBI staff count 2008-2014.



1.2.4 Data resources

New technologies create new data types and allow generation of large volumes of previously rare data. Our service teams monitor these new data types and from time to time initiate new services to handle the new data. Most new data resources begin with startup funding from a funding agency. This is used to set up data structures, establish submission pipelines, and to create some basic analytical tools for the new data. Two recent examples are MetaboLights and EBI Metagenomics, both of which were established using funding from the Bioinformatics and Biological Resources fund of the UK's Biotechnology and Biological Science Research Council (BBSRC). EMBL-EBI also occasionally takes over management of data resources that were begun elsewhere. Table C.1.1 lists data resources we established at EMBL-EBI since 2010 or that moved to EMBL-EBI since 2010.

Table C.1.1 EMBL-EBI data resources established since 2010. When two dates are listed the first is year the resource was established, the second is the year it moved to EMBL-EBI.

Resource	Year established	Description
DGVa	2011	The Database of Genomic Variants archive (DGVa) is a repository that provides archiving, accessioning and distribution of publicly available genomic structural variants, in all species.
EBI Metagenomicss	2011	The EBI Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data that aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample.
Enzyme Portal	2012	The Enzyme Portal integrates publicly available information about enzymes, such as small-molecule chemistry, biochemical pathways and drug compounds.
Europe PubMedCentral	2012	Europe PubMed Central offers free access to biomedical literature resources
MetaboLights	2012	MetaboLights is a database for Metabolomics experiments and derived information.
Pfam	1996/2012	Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. Moved from Wellcome Trust Sanger Institute.
Rfam	2003/2012	Rfam is a database containing information about non-coding RNA (ncRNA) families and other structured RNA elements. Moved from Wellcome Trust Sanger Institute.
RNAcentral	2012	RNAcentral is a resource that offers integrated access to a comprehensive and up-to-date set of ncRNA sequences that are provided by a collaborating group of expert databases and supplemented by sequences from the European Nucleotide Archive

		(ENA).
Treefam	2006/2012	TreeFam is a database of phylogenetic trees of animal genes. Moved from Wellcome Trust Sanger Institute.
SureChEMBL	2010/2014	SureChEMBL is a database of chemical structures derived from patents. Moved from Digital Science (digital-science.com).
European Variation Archive	2014	The European Variation Archive (EVA) is an open-access database of all types of genetic variation data from all species.
EMPIAR	2014	EMPIAR, the Electron Microscopy Pilot Image Archive, is a public resource for raw, 2D electron microscopy images.

Data resources have a natural lifespan, and from time to time we discontinue a data resource due to obsolescence or the end of a finite funding period. Additionally, some resources are merged into others, and some resources move with an individual when he or she leaves EMBL-EBI. Table C.1.2 lists resources that have been discontinued at EMBL-EBI for one of these reasons.

Table C.1.2 EMBL-EBI data resources disestablished since 2010.

Resource	Year disestablished	Outcome
NMRShiftDB	2010	Moved with Principal Investigator to University of Köln, Germany.
CluSTr	2011	Discontinued.
Genome Reviews	2012	Incorporated into Ensembl Genomes.
ASTD	2012	Incorporated in Ensembl and Ensembl genomes.
LGICdb	2012	Moved with Principal Investigator to Babraham Institute, UK.
IPI	2012	Discontinued and functionality replaced by UniProt Proteomes.
RESID	2012	Moved with Principal Investigator to University of Delaware, USA.

1.2.5 Summary of focal areas from last EMBL Programme

In the last EMBL Programme we highlighted four areas of focus for the current indicative scheme: genetic variation, cheminformatics, samples & phenotypes, and electronic literature. All four areas have since seen tremendous progress as described below.

Variation

Developments in DNA sequencing technology have led to an explosion in genetic variation data for humans and other species. There are thousands of completed human genomes, many thousands of cancer genomes, and tens of thousands more human genomes planned or in pipelines. Each genome has many millions of variants when compared to a reference genome.

EMBL-EBI has been addressing the challenge of how to keep track of these variants and their potential phenotypic or medical significance. The Database of Genomic Variants archive (DGVa) was established in 2011 to store variants in genomic structural information (insertions, deletions, inversions, translocations and locus copy number changes) for all species. This complements the single nucleotide polymorphism database (dbSNP) at the National Center for Biotechnology Information (NCBI) in the US, which stores non-structural variants.

To further enhance service in the area of variation we created a new Variation Team in 2012 to consolidate and enhance our ability to track genomic variation. In October 2014 EMBL-EBI announced a new service, the European Variation Archive (EVA), which is an open-access database of all types of genetic variation data from all species. This new archive consolidates and expands all of our previous variation activities.

Chemoinformatics

Our resources describing chemicals and bioactivities (ChEBI and ChEMBL) have continued to grow successfully, both in size and in users, and have been recently extended with the chemistry patent resource SureChEMBL in 2014. These resources were augmented with the MetaboLights database of metabolomics experiments and derived information, which was founded in 2010 and formally inaugurated for public access in 2012. The Chemistry cluster will be working both towards a more complete documentation of the endogenous metabolomes of organisms, with a focus on selected model species, as well as on the influence of xenobiotic compounds on molecular systems, with applications, for example in synthetic biology and compound safety. EMBL-EBI’s significant expertise in the area of chemistry and small molecules contributed heavily to the establishment of the Center of Therapeutic Target Validation (CTTV, Box E.5).

Samples & Phenotypes

The BioSamples database, inaugurated in 2009, aggregates and standardises sample information for reference samples (e.g. Coriell Cell lines) and samples for which data exist in one of EMBL-EBI's assay databases such as ArrayExpress, the European Nucleotide Archive or the proteomics data repository PRIDE. It provides links to assays on specific samples, and accepts direct submissions of sample information. The Experimental Factor Ontology supports cross platform and species queries in BioSamples. BioSamples and the Experimental Factor Ontology, in concert with an improved search engine, have greatly improved the discoverability of related datasets across all our data resources.

Literature

Over the past five years EMBL-EBI has taken over the lead role in developing Europe Pubmed Central (Europe PMC) and consolidated existing literature services, such as CiteXplore, a database of biomedical abstracts, into Europe PMC as a single database and brand. Today Europe PMC acts as the supporting repository for 26 European

funders of life science research. It contains around 30 million abstracts, including all of PubMed, and around 3 million full text research articles. The content is updated daily and supported by powerful search and retrieval mechanisms, including cloud-based full-text searches. The website is now accessed by around one million unique IP addresses every month.

Europe PMC has played a leading role in the implementation of unique author IDs (ORCIDs), both through incorporating them into article records and search mechanisms, and through the development of a tool to allow authors to link their articles to their ORCID record. EMBL adopted the use of ORCIDs for all its scientific staff in 2014. Work on the citation of articles, text-mining of data citations in articles and linking articles to funding has made Europe PMC is a key component in open scientific credit systems.

The Europe PMC team has made great progress in linking literature and data stored in other EMBL-EBI databases with the help of text-mining data accession numbers and vocabulary-based text-mining pipelines, including GO, EFO, gene/protein symbols, disease, organisms and chemicals. This forms the basis of future work to develop tools that identify relationships between scientific concepts and new search result ranking mechanisms and will contribute to improving database curation workflows.

1.3 Future Plans 2017-2021

Bioinformatics, and indeed all of biology, is in the midst of a revolution: new technologies are making it easier and cheaper to undertake large scale experiments that generate vast quantities of data, and more and more life scientists are becoming data scientists. For EMBL's Bioinformatics Services this revolution brings challenges, but also exciting opportunities. The challenges lie in how to process, store, and analyse these data, and the opportunities lie in integrating these new data to generate new knowledge, and in developing new services for an expanding user community. In this section we highlight some of these challenges and opportunities and discuss how we will respond to these over the next few years.

1.3.1 New users, increased user focus & user-driven data integration

The increasing diversity of our user base, and the complexity of the data we handle, means that we have to become even more proactive in understanding the needs of users and in engaging with new user communities. The success of Ensembl and Ensembl Genomes in engaging with the farm animal and plant communities has established those two services as focal resources with the best possible quality data for all farm animal and plant scientists worldwide. A priority over the next few years will be to engage in a similar way with other user communities, for instance in biodiversity, pathogen surveillance, and food safety. We also aim to become a major facilitator of the use of omics data for translational medicine by working with clinicians and clinical data resource providers.

With increased data volumes many user communities are interested in tailoring data delivery for their specific needs, for instance by using a dedicated portal to organise, deliver, and provide analysis tools for data that are held at EMBL-EBI and other public repositories. Current examples of such portals are VectorBase, PomBase, and WormBase, which provide community-oriented access to, respectively, data on arthropod disease vectors, single-celled yeast, and roundworms.

There is clear demand, particularly among human disease research communities, for similar portals. We have experience and expertise in creating such resources and will partner with research communities and funders to support the development of specialised portals that make use of data held at EMBL-EBI and other places in new and innovative ways.

Another extremely important and growing user group for EMBL-EBI's services is industry. The pharmaceutical industry has traditionally been the biggest non-academic consumer of EMBL-EBI services. Through the EMBL-EBI Industry Programme and other ways of interacting with industry (both described in detail in Section E.4), we have regular contact with commercial users, which helps us set priorities. Over the period of the next EMBL Programme we see our interactions with industry partners becoming stronger as we work together to address the challenges and opportunities created by big data in terms of cost effective paradigms to manage the volumes of data, methods to ensure appropriate integration of information, and models to protect the confidentiality of proprietary, licensed and personal information in a manner that promotes innovation and translation into practical benefits.

It is also important to point out that our ability to provide a strong user-training programme in bioinformatics and the use of the data resources is completely reliant on technical experts in the service teams as well as on the support provided by the training and logistics expertise of the training team. The EMBL-EBI service teams will continue to contribute to EMBL's training mission as summarised in Section D.

1.3.2 Service and research collaboration

EMBL-EBI has strong research groups in bioinformatics that are well connected to other computational and experimental researchers within EMBL and around the world. The bioinformatics service teams benefit tremendously from interacting with the research teams, which expose them to emerging scientific questions and methods. The largest research projects with which EMBL-EBI service teams are currently involved are bulleted below: all involve large-scale data generation, storage, analysis, and actively synthesizing the data to understand big questions in biology.

- 1000 Genomes project
- International Cancer Genome Consortium
- International Human Epigenome Consortium/BluePrint project
- International Mouse Phenotyping Consortium
- iPS cell characterisation and resources: Human Induced Pluripotent Stem Cells Initiative & European Bank for Induced Pluripotent Stem Cells
- Analysis and interpretation of large scale proteomics data

In addition, the bioinformatics services benefit from regular exchange with wetlab biologists at other EMBL sites. This exchange is vital for facilitating and maintaining close contact with cutting-edge research, and helping anticipate future trends in data generation and user needs. Good examples of collaboration between EMBL researchers and bioinformatics service teams are the new image repository (Box C.1.2) that EMBL-EBI will build up together with imaging experts from EMBL and Euro-BioImaging (Section

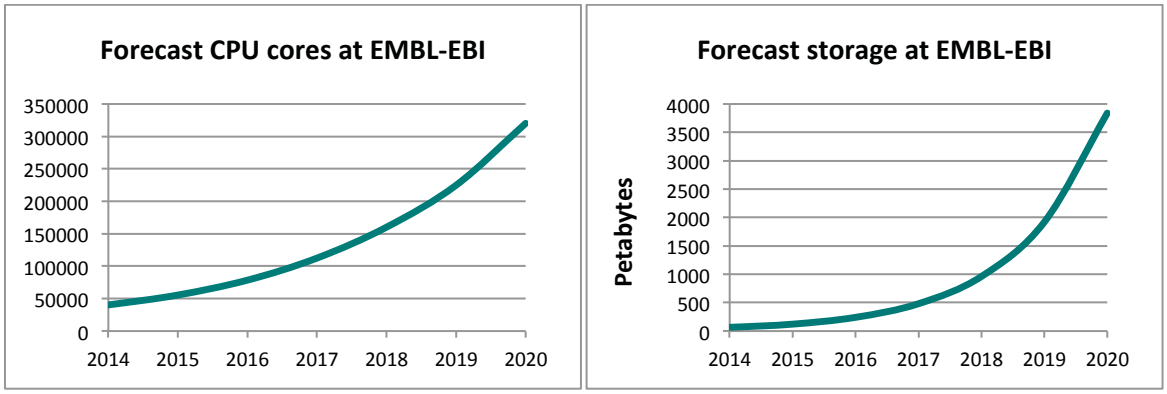
F.1.3.2) and collaborations in the area of single-cell sequencing data resources that involve wetlab groups at EMBL Heidelberg and the Wellcome Trust Sanger Centre.

In addition to collaboration with experimental groups at EMBL, the bioinformatics service teams also work closely with groups on the Wellcome Genome Campus, in Cambridge and the rest of the UK; and are involved in many large, international consortia. In the interests of the future development of EMBL-EBI services, we will continue to foster close interactions between service teams and research groups as well as active involvement in leading international research collaborations.

1.3.3 Computational requirements and Cloud services

The rapid growth of biomolecular data (Figure C.1.7) will create challenges with regard to storage and compute capacity as well as network bandwidth. It may be necessary to increase the amount spent on hardware relative to staff since our expected hardware requirements currently outpaces the drop in prices for storage and compute. To cope with bandwidth and other limitations it is important to enable users to bring their analysis pipelines to the large datasets available at EMBL-EBI. This will be achieved by expanding our Cloud services.

Figure C.1.7 EMBL-EBI computational hardware requirements 2014-2020. 2014 figures are actual, later years are forecast.



EMBL is already very active in developing and testing cloud-computing solutions for biological data. EMBL-EBI operates a cloud service, the Embassy Cloud (Figure C.1.8), that has been live for two years and currently has 10 users: one commercial and nine academic. The original rationale for the cloud service was to enable users to bring analysis platforms (as virtual machines) to very large datasets within our data centre and so obviate the need for users to download the large datasets and install services into their own data centres for local analysis. The Embassy Cloud has been very successful in this regard. However, there is arguably a more important benefit: cloud analyses are usually collaborations between different groups, and the virtual pipeline allows all partners to see all steps of the data pipeline in action and without large data transfers. This transparency of intermediate results has proven extremely useful for the users who are, as a result, working more flexibly and more efficiently. The EMBL-EBI Embassy Cloud will continue to expand as user demand for cloud services increases, and we will work to improve the usability of cloud services so that users who, for instance, do not have expertise in setting up a virtual machine can access our cloud services.

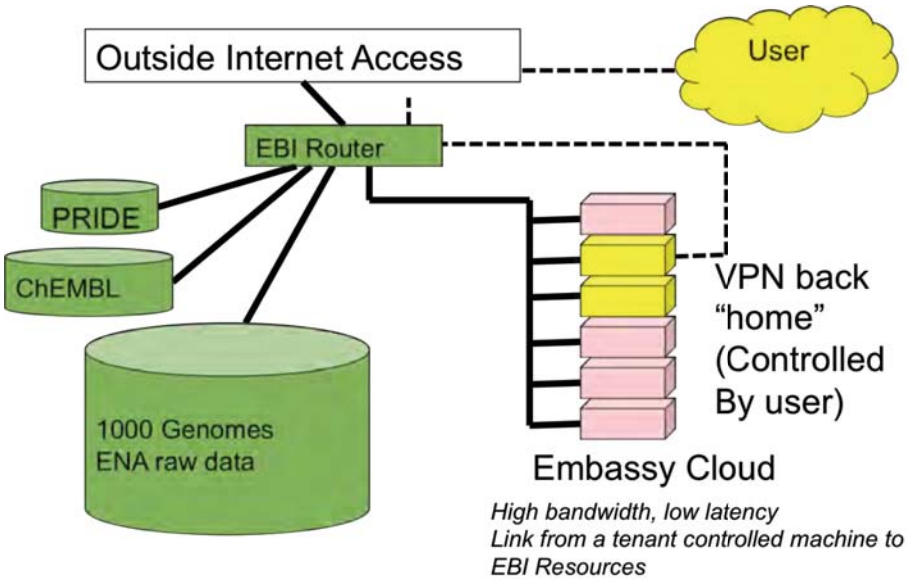


Figure C.1.8 Figure showing access to the Embassy Cloud. Users have rapid access to the cloud resources because their virtual machine(s) are inside the EMBL-EBI network.

EMBL-EBI is also engaging in discussions with commercial providers to test suitability of commercial cloud services for biomedical research and service provision. In addition, EMBL Heidelberg and EMBL-EBI participate together in 'Helix Nebula – The Science Cloud', a partnership between leading IT providers and three of Europe's leading intergovernmental research organisations (CERN, EMBL and ESA) that is supported by the European Commission. The aim is to develop cloud-computing services for academia involved in big science.

These initiatives have underlined the large potential for cloud computing in life science research, but privacy and confidentiality issues, especially regarding medical and other sensitive data, require more work in the future. We are already, with some of our EIROforum partners, involved in discussions with the European Commission on the next phase of Helix-Nebula. Building on these ongoing efforts EMBL-EBI is in a good position to take on a strategic role in cloud computing for biological data in the future and we will continue to explore this possibility.

1.3.4 New data: integration, challenges, opportunities

The volume of data and the wealth of new data types present many opportunities to create new knowledge and we will actively promote our data resources and reach out to various user communities to promote these opportunities. We highlight a few of these below.

Variation

New sequencing technologies have reduced the costs of sequencing, and those costs continue to drop. The current laboratory cost to sequence a human genome, around €1,000 (not including assembly and analysis), makes the costs of sequencing roughly comparable to most standard single-gene laboratory tests: it will not be long before clinicians request whole genome sequencing rather than a series of gene-specific tests. The number of sequenced human genomes is increasing exponentially. Over the next few years there will be a tremendous effort to catalogue the variation in these newly

sequenced genomes, to understand how these variants affect phenotype and how they affect an individual's health, and to provide reference data upon which clinical analyses of sequences is based. EMBL-EBI does not handle patient data, but we do act as a repository for reference data, and we will store reference variation data for humans (and other species) and will provide these data to scientists and healthcare professionals.

The European Variation Archive (EVA) tracks genomic variation data at all scales, from large chromosomal rearrangements to single nucleotide polymorphisms, as well as phenotypes for those variants. The EVA will grow, probably exponentially, over the next few years to incorporate variants from new publicly available sequences and to develop and implement new strategies for reporting variation.

Single cell analysis

New sequencing technologies not only enable high throughput, but also extreme sensitivity. This makes it possible to quantify DNA, RNA, and epigenetic changes in single cells, offering unprecedented access to study how genetic and/or environmental variability impacts on individual cells and cell types and their phenotypes. Sinngle-cell technologies will generate another source of big data: rather than a single genome or transcriptome from an individual there is now the potential to generate hundreds or thousands of datasets from single cells from a single individual. From a data storage perspective these data are similar to any other genomic, transcriptomic, or other omics data and can be accommodated in existing resources provided these can be sufficiently expanded. One challenge will be to integrate the various datasets from a single cell so that users can identify all data generated during a single experiment. Existing resources, in particular the BioSamples database, can already handle such cases, but we must improve efficiency of single-cell data handling and the ease of data submission. A bigger challenge, and one that is also an opportunity, is in analysis of single cell data. What are the best ways of integrating multiple different types of data from a single cell, or of studying temporal changes in the transcriptomes of multiple cells from the same organ system? We will be working, in collaboration with our users, to develop the tools to answer these questions.

Box C.1.1: The European Genome-phenome Archive

The European Genome-phenome Archive (EGA) is a permanent archive for biomolecular data generated from individual human research subjects in the course of disease, population genomics, or other biomedical research. As of 2014 data from more than 500,000 individuals and nearly 600 studies are included. These data are critical for efforts to establish genomic-based clinical care and other efforts in molecular approaches to medicine.

The EGA was founded at EMBL-EBI in 2007 and expanded in 2012 to a partner site at the Centre for Genome Regulation (CRG) in Barcelona, Spain. We envision that this partnership will grow into a fully federated network of European databases supporting the storage, access, sharing, and analysis of research data generated by biomedical researchers. A federated EGA would be a unique and sustainable resource providing robust data security, physical backup, and live service failure recovery.

A federated EGA, with multiple sites beyond the current EMBL-EBI and CRG implementations, would have tangible benefits for all EMBL member states. For example, federation expands the number of datasets and total human and computing resources available to European scientists, making datasets and technical expertise available to a wider audience.

Software and standards supporting a common EGA infrastructure would be made available to EGA partners supporting various activities including data submission, archival storage and data access. Importantly, a federated model would include a consolidated search function and support new data access and analysis methods such as secure analysis within a cloud-like computing facility.

EMBL-EBI would play a leadership and coordination role in the context of the other EGA node(s) to ensure compatibility and facilitate federated data sharing. Extensive technical and policy coordination is already underway between the CRG and EMBL-EBI.

These activities are important components of emerging global scientific efforts. For example, the EGA shares many goals with the Global Alliance for Genomics and Health (GA4GH). The EGA is already active in the GA4GH and is well positioned to take a major leading role for this effort in Europe. In addition, the coordination role and organisational model of ELIXIR is perfectly suited to a federated EGA.

Images

In the past EMBL-EBI has not dealt with imaging data. However, imaging is a central component of many of EMBL's future research plans, as outlined in Sections B.2.4 & E.1.1.2 of this document, and will drive progress in life science research also beyond EMBL. Recent technological advances have resulted in datasets that contain molecular or phenotypic information and images. For example, single cell experimental work often now includes collections of microscopy images of individual cells or groups of cells in association with generation of molecular data (genomic or RNA sequencing). A single experiment therefore has both molecular and imaging components.

EMBL-EBI was a partner in the EU funded FP7 Systems Microscopy Network of Excellence and participated in developing standards for describing images, developing ontologies to describe cellular phenotypes, and associating these data types with molecular data. On the basis of this previous work, and in collaboration with Euro-BioImaging (see Section F.1.3.2), we are now actively developing an image repository to store high-resolution light microscopy images that are associated with molecular data in

our repositories. At present the emphasis is on cellular images and single cell data, but this resource could be extended to include samples of organs, tumours, and even whole organisms. Other image types include cellular and organismal atlases, tomography, and chemical complexes. We will be actively exploring the demand for, and the technical feasibility of, developing resources for these data types.

Box C.1.2: Developing an Image Data Repository in close collaboration with Euro-BioImaging and ELIXIR

It is our aim to build an Image Data Repository (IDR) that will host image datasets for the worldwide scientific community. This will be developed in collaboration with Euro-BioImaging (Section F.1.3.2). We will be working closely with colleagues at Dundee University in the UK, and with EMBL's Cell Biology and Biophysics Unit in Heidelberg. EMBL-EBI will contribute experience in handling large and complex datasets and build on its experience in leading the database and data standards work packages in the EU funded FP7 Systems Microscopy Network of Excellence. A pilot project towards this goal is already under way, jointly with Dundee University and partly funded by the UK's BBSRC.

The need for such a resource has been highlighted in the Euro-BioImaging preparatory phase project, which performed surveys among European scientists and imaging facilities. Much of the published research in the life sciences carries with it detailed image data that are often used for quantitative measurements of biological processes, but which also typically use these data in conjunction with genomics or other molecular data. For instance molecular phenotyping uses RNAi reagents to knock down specific genes, followed by automated microscopy-based measurement and recording of the effects of each gene knockout. To represent the results from these high-throughput experiments we need images of these cells, appropriate ontologies to describe the image analysis results, links to affected genes in the genomes, and the respective molecular reagents. EMBL-EBI is in a perfect position to lead such data integration.

An image-based genome-wide high-content screen may have over a million images. New 'virtual slide' and 'light sheet' tissue imaging technologies generate individual images that contain gigapixels of data showing tissues or whole organisms at subcellular resolution. The size and complexity of image datasets and multi-dimensional images makes centralised data submission, handling and publication extremely complex. To address these challenges, the IDR will be combining centralised and distributed approaches. We have already started working towards establishing image data standards and ontologies as part of our participation in the Systems Microscopy Network of Excellence and BioMedBridges projects. Over the next years these activities will be extended to the global level. To begin with, we will use the emerging BioStudy Database at EMBL-EBI as a home for reference images but a more specialised database will be developed as the project gathers momentum.

In this way our approach will combine the experience of EMBL-EBI in managing large datasets and standards development, the expertise of the EMBL Cell Biology and Biophysics Unit in molecular imaging, and foster the close integration between EMBL-EBI and ELIXIR as well as collaboration with Euro-BioImaging.

Medical data

Molecular biology is increasingly relevant to clinical research and, in some cases, clinical medicine (see Section B.2.3). This is because the fundamental processes of living organisms are driven by molecular events, and because the costs of automated data-gathering technologies continue to fall. Coupled with cost–effective imaging techniques, a large amount of data can be gathered and integrated to help inform both clinical research and patient treatment. EMBL's bioinformatics services are increasingly being called upon to provide support for clinical researchers and practicing healthcare professionals. Our core function is to store and serve data that are in the public domain and we do not, and will not, store patient data, which are inherently private. However, there is large scope for providing reference datasets for clinical providers and to work internationally, particularly in the context of ELIXIR, to harmonise storage and analysis of medically relevant omics data across Europe. We discuss this in more detail in Section C.1.3.5 below.

Unstructured data

Large volumes of data produced as outputs of biological research fit neither in the core, structured databases at EMBL-EBI, nor in the narrative of a research article. The formats, types, sizes, and levels of detail of these data are highly variable and include images, models, videos, spreadsheets, software code, etc. These data fall between the cracks of research articles and structured research data, yet they are an important part of any scientific study.

Several publishers have launched data journals, implemented open data policies, and are becoming more diligent in handling supplemental data. Furthermore, funding agencies are beginning to mandate that data outputs from projects are made publicly available and are requiring data management plans in grant applications. The NIH's BD2K initiative is striving to create an international 'data commons' to provide the basis for open science in the future. Because of the diversity and scale of the data being generated, it is likely that any solution will be distributed, and in fact some solutions to meet this demand have recently emerged, for example, FigShare, Dryad and Zenodo.

There are significant potential benefits for EMBL-EBI to extend its services into this area. The BioStudies database (under development) will provide a study-centric view of all the components related to an experiment: housing unstructured data as required; linking to articles and structured data elsewhere; and aligning with projects such as Euro-BioImaging. This database will give EMBL-EBI the opportunity to spot emerging scientific trends and potentially provide future new services based on clusters of new data and relationships with existing data and articles. Via ELIXIR and BD2K, EMBL-EBI will also be exploring mechanisms to improve data discovery through the use of universal identifier systems such as DOIs and related mechanisms. Coupled with the continued adoption of ORCIDs, which are used to unambiguously identify the articles, grants and datasets generated (or curated) by a person, it is clear that the EMBL-EBI has the opportunity to play a significant role in future scientific credit systems and open science infrastructures though engaging in this area.

1.3.5 Opportunities in medicine and health

Other parts of the EMBL Programme (Section B.2.3) show clearly that molecular biology is increasingly relevant to clinical research and to the practice of medicine. The application of basic biological knowledge to create clinically useful treatments or drugs has been greatly enhanced by molecular data, to the point where basic research and

translational research, which have traditionally been quite distinct, now overlap. Over the next Programme period EMBL-EBI will actively engage in building bridges between biological information and clinical data to develop medically relevant data infrastructures. Within this general context we anticipate that our work will focus on three general thematic areas, summarised below.

Providing reference datasets for clinical research

EMBL-EBI, in partnership with other European institutes in an effort that may be coordinated through ELIXIR, will focus on data that is open, research-based, independent of healthcare systems, and relevant to human health. This includes:

- Annotation of variants of clinical importance in different diseases;
- Annotation of somatic variants, and their association with drug responses;
- Cataloguing and understanding small molecule/drug interactions with proteins and protein variants, including their 3-D structure interactions and pharmacokinetic parameters;
- Cataloguing and understanding human pathogens and their virulence components;
- Cataloguing and understanding biomarkers (e.g. protein, RNA, metabolites, and complete metabolomes) for diagnosis and disease monitoring;
- Capturing and making available molecular profiles that characterise differences between disease and normal states, or provide sub-classification of a disease;
- Developing resources that are fundamentally about understanding biology, in particular the molecular basis of human disease.

Working collaboratively with partners without handling clinical data

Rather than seeking to maintain all data resources we will form partnerships with other groups, initially in Europe and in collaboration with ELIXIR, but with the intent to expand the range of partners as opportunities arise. Within these partnerships, we expect to offer our engineering, algorithmic and statistical expertise in biomolecular work, as well as an in-depth understanding of all manner of biomolecular datasets to help our partners store and handle large data resources.

One area of current interest is the development of software and tools for managing clinical phenotypes and anonymised patient records for research purposes. We expect to be heavily involved in such initiatives, and would also expect that the resulting outputs would be deposited in the appropriate EMBL-EBI databases.

EMBL-EBI's core remit is to share data, which makes our participation in practicing healthcare inappropriate, as noted above. However, our participation in clinical research is both appropriate and desirable and we will actively collaborate with healthcare researchers in such areas.

As EMBL-EBI only handles data for broad research reuse, a key criterion is that the outcome of any specific project must be of utility to researchers beyond EMBL-EBI and the other partners. For non-human data, this means availability in the public domain, while for human data we will have managed access on the basis of patient privacy and informed consent.

We will actively participate in opportunities for international coordination of open human disease data initiatives worldwide. Large North America investments in health research will open up an increasing number of opportunities for clinical research collaborations, including those relating to infrastructure. Our coordination of the Locus Reference Genomics (LRG) and Genetic Testing Registry (GTR) resources are notable examples.

Promoting healthcare data coordination across Europe and the world

Healthcare, unlike research, is localized, usually within a country, and is practiced under complex national legislation, with most of the patient-specific data being private to the patient and clinician and in the local language. These data rarely leave the hospital or clinic, let alone the country.

By contrast, research is carried out internationally, often with very lightweight legal restrictions, and mainly in English (in the interests of facilitating discussion amongst international groups). Biological research also has a long history of data sharing, as illustrated by the publication of the human genome in the public domain.

EMBL-EBI is firmly international and highly supportive of data sharing, which makes the participation of our service teams in practicing healthcare inappropriate. European countries have diverse healthcare systems, each of which has different internal drivers and components. They are strikingly different in their uptake of genomics and other molecular technologies. We like to collaborate with the various systems in the context of ELIXIR to help develop expertise and data management skills for genomic medicine. We expect the speed of adoption by various national systems to be variable and we will maintain a flexible approach to capacity building. Ultimately, we hope that each separate healthcare system will develop at least one biomedical informatics institute or network with a coordinating centre that would be a natural partner for EMBL-EBI. The development of biomedical informatics networks may be coordinated through an ELIXIR node in some countries.

It is clear that his model requires effort on both sides, on the part of EMBL and of the national medical institutions, but we believe that such collaboration would be a very effective way forward. EMBL is committed to this endeavour and to establishing the required collaborations with member state institutions in the area of medical informatics. Good but non-medical examples on which such cooperations can be modelled are the collaboration between EMBL-EBI and the Centre for Genomic Regulation (CRG) in Barcelona, who jointly operate the European Genome-Phenome Archive, and the long-standing collaboration between EMBL-EBI and the Swiss Bioinformatics Institute in operating UniProt, the knowledgebase for protein sequence and functional information.

Additionally, EMBL is a founding member of the Global Alliance for Genomics and Health, an international coalition dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. The key goal of the alliance is to create data standards and strategies for storage and analysis of medically relevant genomic data, and to catalyse the creation of data sharing standards and methods to ensure worldwide interoperability of medical genomics data. EMBL-EBI, as well as other parts of EMBL, will participate fully in Global Alliance activities in order to promote interoperability of data worldwide.

1.3.6 EMBL-EBI role in ELIXIR

Hundreds of thousands of researchers in basic biology and applied disciplines actively access EMBL-EBI resources every year, many on a daily basis. Over the coming decades the challenges of providing effective bioinformatics infrastructure for these users will continue to grow, and have accordingly led to the development of a distributed European infrastructure for biological data: ELIXIR. EMBL played a very active role in establishing ELIXIR, which is described further in Section F.1.3.1. ELIXIR broadens the scientific base of data service provision, making full use of important contributors throughout Europe with EMBL-EBI hosting the central, coordinating hub. EMBL is committed to working together with all our national partners in ELIXIR to develop sustainable models and solutions for data archiving and access in the future.

In addition to EMBL being the host of the ELIXIR hub, EMBL-EBI is also an ELIXIR node, and will provide services to the ELIXIR infrastructure. One aspect of this service provision will be the designation of many existing EMBL-EBI resources as 'ELIXIR-affiliated' resources, and of some of these ELIXIR-affiliated resources as 'ELIXIR core resources'. ELIXIR affiliation, and assignment as core resources, will be contingent on these resources fulfilling certain criteria with regard to usage, reliability, and importance to their respective user communities. The exact parameters of these criteria are as yet undefined and one task in the near future will be to define the criteria by which resources should be ELIXIR-affiliated and/or designated as ELIXIR core resources. EMBL-EBI has more established data resources than any other single organisation in the ELIXIR signatory countries and will serve as the model for the affiliation of data resources from other ELIXIR nodes.

EMBL-EBI may also undertake commissioned work for ELIXIR, either by itself or in collaboration with other ELIXIR nodes. Such work will be for specific projects that make use of EMBL-EBI's expertise in data handling, analysis, and the creation of services, and that add value to ELIXIR activities. The details of each such work, including any ELIXIR funding support, will be agreed between EMBL-EBI, the ELIXIR Hub, and other involved ELIXIR nodes; and will be specific for each project. We expect that the ELIXIR hub will commission a number of projects from EMBL-EBI in the near future.

In addition to these relationships between EMBL-EBI as an institution and ELIXIR, individual EMBL-EBI service teams will also undertake activities under the ELIXIR umbrella, primarily through grant funding. We anticipate that one type of activity will be the development of data resources for healthcare and medicine across Europe, such as those described above, in which EMBL-EBI service teams develop clinically relevant resources or participate in the construction and definition of infrastructures for medical data.

2. EMBL structural biology services

2.1 EMBL's mission in the provision of scientific infrastructures in structural biology

Structural biology over multiple resolution ranges underpins the mechanistic understanding of biological organisation and processes at all levels and is crucial for translational research, notably for drug design. EMBL has often been at the forefront of the development of novel technologies to allow the increasingly sophisticated study of complex structures and its mission is to make such technologies available to a broad international scientific community. Indeed, EMBL is uniquely able to provide highly complementary, integrated and internationally leading services from its three structural biology-oriented Units in Grenoble, Hamburg and Heidelberg. EMBL Grenoble and EMBL Hamburg are situated at powerful synchrotron and the future European X-ray Free Electron Laser (European XFEL) sites and their focus is on X-ray-based research services in structural biology, which they provide in close collaboration with the European Synchrotron Radiation Facility (ESRF) and the German Electron Synchrotron (DESY), respectively. During the past few years, both sites have complemented their synchrotron-based facilities with a complete pipeline of structural biology services in sample preparation and characterisation as well as data processing and analysis. Whereas the emphasis of EMBL Grenoble is on protein expression and crystallisation technologies, the Hamburg Unit focuses on biophysical methods that permit a thorough sample quality assessment prior to subsequent X-ray data acquisition. EMBL Hamburg has also established an impressive portfolio of computational services required for automatic and rapid interpretation of different types of structural biology data.

In addition to their service activities, the two Units in Grenoble and Hamburg are not only spearheading technology and methods development not only in synchrotron instrumentation but also in the areas of sample preparation and computational data analysis (Section E.1.1.1). Technology advances are rapidly shared between the two EMBL sites and directly implemented into the service platforms so as to benefit the large external user community. They are also frequently adopted by other synchrotron sites, sometimes after technology transfer to companies. Regular bilateral meetings of the synchrotron instrumentation groups in Grenoble and Hamburg ensure exchange of information and often result in joint technology development projects.

In Heidelberg, the Structural and Computational Biology (SCB) Unit is integrated into the interdisciplinary research environment of the EMBL headquarters. To date, the Unit has mainly focused on internal research services and collaborations in structural biology, centered largely on advanced electron microscopy (EM) and imaging techniques at the interface with cell biology. The local infrastructures and leading expertise in various EM applications, however, put the SCB Unit in Heidelberg in an excellent position to offer its state-of-the-art services to a broader external research community than in the past.

Transnational access to most of EMBL's structural biology service activities has been for many years supported by Integrated I3 Projects from the European Commission. During the current Indicative Scheme, this has been via P-Cube for protein production and crystallisation (2009–2013) involving all three Units, ELISA (2009–2011) for synchrotron access (EMBL Hamburg and EMBL Grenoble) and BioStruct-X (2011–2015, co-ordinated by EMBL Hamburg and including EMBL Grenoble), which covers protein

production, crystallisation and synchrotron access. A new I3 Project, iNEXT, has recently been funded by the EU but, unfortunately, overall European funding for user access and related infrastructures RTD projects has diminished in recent years and is now inadequate to meet the needs of the structural biology community, particularly for access to new technologies such as advanced EM. A dedicated EMBL scheme to fund access to some of the facilities offered should be considered by the member states.

2.2 Backward look and highlights 2012–2014

2.1 Synchrotron radiation services

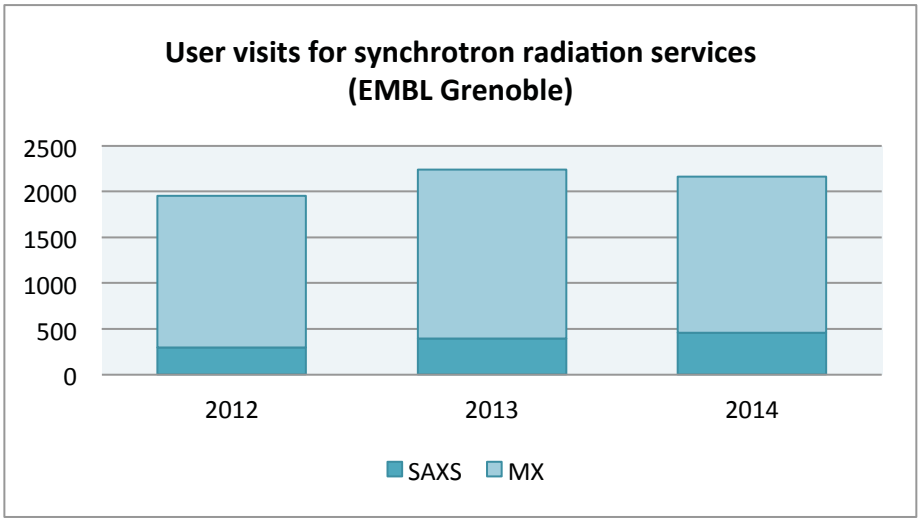
The construction and provision of synchrotron beamlines has revolutionised structural biology during the past decades. The Protein Data Bank now hosts more than 100,000 high-resolution structures of biological macromolecules and more than 85% of them have been determined by X-ray crystallography, which is now almost exclusively carried out at leading synchrotron sites around the world. Taken together, the beamlines operated by EMBL Hamburg and EMBL Grenoble–ESRF, despite the transition from DORIS III to PETRA III and temporary shutdowns at the ESRF, have produced close to 40% of all Protein Data Bank entries originating from Europe over the first three years of the current Indicative Scheme (2012–2014).

EMBL Grenoble

In Grenoble, the EMBL–ESRF Joint Structural Biology Group provides user access to five (six in the future) macromolecular crystallography (MX) and one biological small angle X-ray scattering (bioSAXS) high-performance beamlines at the ESRF. In addition, EMBL manages the Collaborative Research Group beamline, BM14, on behalf of a consortium comprising EMBL, ESRF and the Indian research community funded by the Indian Ministry of Science and Technology.

Over the first three years of the current Indicative Scheme, the number of user visits in Grenoble has remained steady despite closure of the first generation of beamlines on ID14 and a five-month closure of the ESRF for the phase I upgrade (from December 2011 to May 2012). This level of access has been achieved by a substantial investment in fast pixel array X-ray detectors and important instrument and software developments. The construction and commissioning of a new suite of modern beamlines on ID30 at the ESRF is in the final stages and user operation on the Massively Automated Sample Selection Integrated Facility (MASSIF) beamlines, which for the first time allow fully automated sample evaluation and data collection, started in 2014. The tunable, high-intensity variable focus beamline, ID30B, will commence user operation in 2015 and will expand the current data collection capabilities for challenging crystals.

Figure C.2.1 Number of user visits to ESRF beamlines operated in collaboration with EMBL Grenoble. Note that there was a five-month shutdown of the ESRF at the beginning of 2012.



EMBL Hamburg

EMBL has always had full responsibility for financing, constructing and operating the structural biology beamlines in Hamburg. During the present Indicative Scheme, EMBL Hamburg completed the construction, commissioning and early operation of the new research infrastructures at the PETRA III storage ring. The integrated facility, EMBL@PETRA3, includes three state-of-the-art undulator synchrotron radiation beamlines, P12 for applications in bioSAXS, beamlines P13 and P14 for complementary applications in MX, an associated facility for biological sample preparation and characterisation (SPC), and a module for structural biology software services. To begin user operation at the three new beamlines P12, P13, and P14, the work required during the present Indicative Scheme mainly addressed three areas: beamline optics, experimental end-stations, and the overall beamline infrastructure.

EMBL Hamburg provides 80% of the available beamtime to the external user community and user selection is based on peer review and scientific excellence. Owing to the DORIS III / PETRA III transition phase and the long PETRA III shutdown in 2014, the annual user statistics during the first three years of the current Indicative Scheme have been variable and inevitably lower than normal. Whereas in 2012 parts of the facilities at DORIS III and PETRA III could still be used in parallel, by 2013 only the new beamlines at PETRA III were available, still in their commissioning and early operation phase. Despite this, in all categories of use, the year 2013 had the highest number of users during the reporting period. This demonstrates the attractiveness of the new beamlines. The MX facilities were still in the ramping-up phase during this period and the SAXS statistics demonstrate the international leadership of the SAXS user support provided in Hamburg. We expect the numbers for both SAXS and MX to increase considerably in the coming years as all beamlines and end-stations move into routine operation.

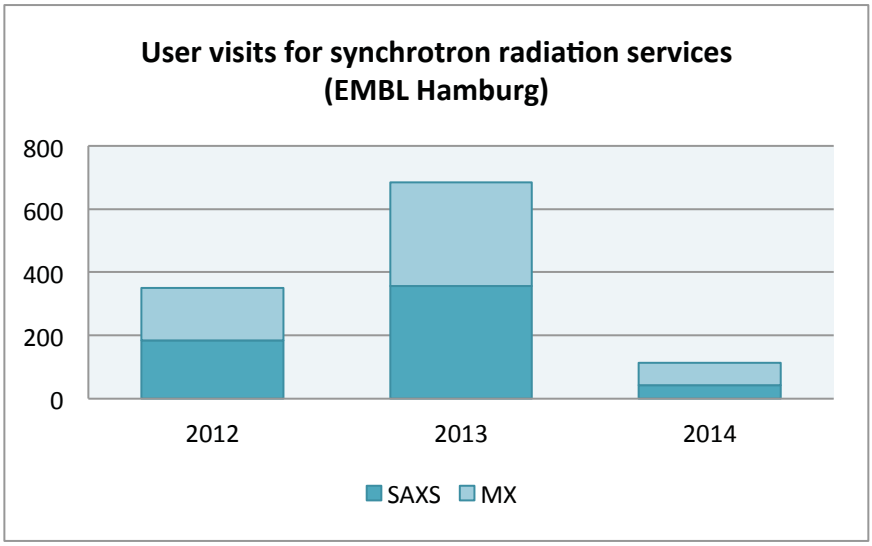


Figure C.2.2 User statistics for EMBL Hamburg. Note that whereas in 2012 parts of the facilities at DORIS III and PETRA III could still be used in parallel, by 2013 only the new beamlines at PETRA III were available in their commissioning and early operation phase. In 2014, user service was only operational for the month of January due to the PETRAIII shutdown from February 2014 until March 2015.

2.2 Sample production, characterisation and crystallisation services

Technological advancements over the past years have made it possible for structural biology to tackle targets of ever-increasing complexity. The quality of the target material is crucial for successful crystallisation. However, complex, multi-component samples are difficult to produce and often unstable. Therefore, access to specialised production facilities in these areas near to the beamlines is often essential to a successful outcome.

EMBL Grenoble and Hamburg have therefore extended the portfolio of services available to European scientists through the introduction of advanced facilities for the identification of soluble protein domains (ESPRIT in GR), eukaryotic cell expression (EEF in GR), sample characterization (SPC in HH) and high throughput crystallisation (HTX in HH and GR) (Figure C.2.3). During the current Indicative Scheme, EMBL developed new technologies, including second generations of the ESPRIT and Multibac systems for protein production, the CrystalDirect instrument for automated crystal harvesting and processing and an integrated crystallisation facility management software system (CRIMS), which has now been installed in several other structural biology facilities in Europe (Section E.1.1.1.2).

The sample preparation and characterisation facilities in Hamburg and Grenoble are highly integrated and located in close proximity to the beamlines. They provide complete structural biology pipelines by connecting work required for sample purification and preparation to the beamlines at the ESRF and PETRA III. The uptake of these new services has been impressive, with over 600 projects processed by these facilities in the period 2012–2014.

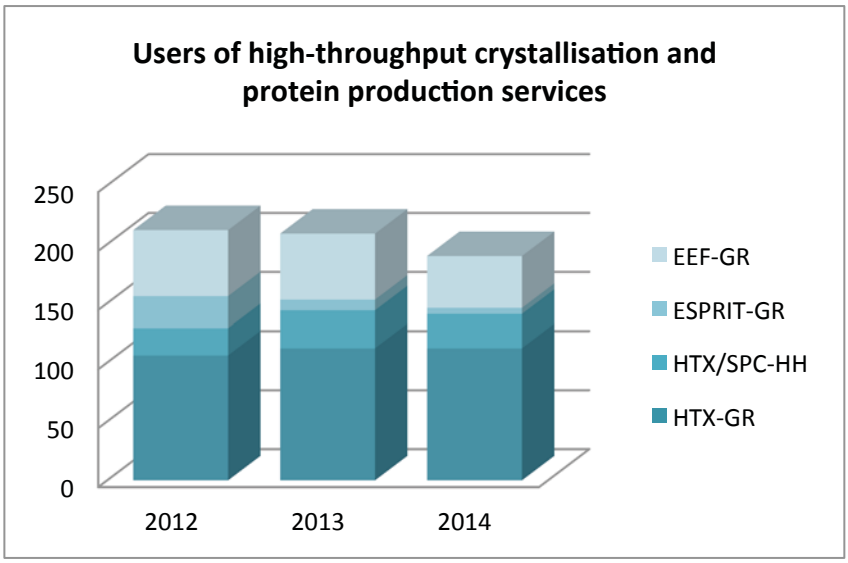
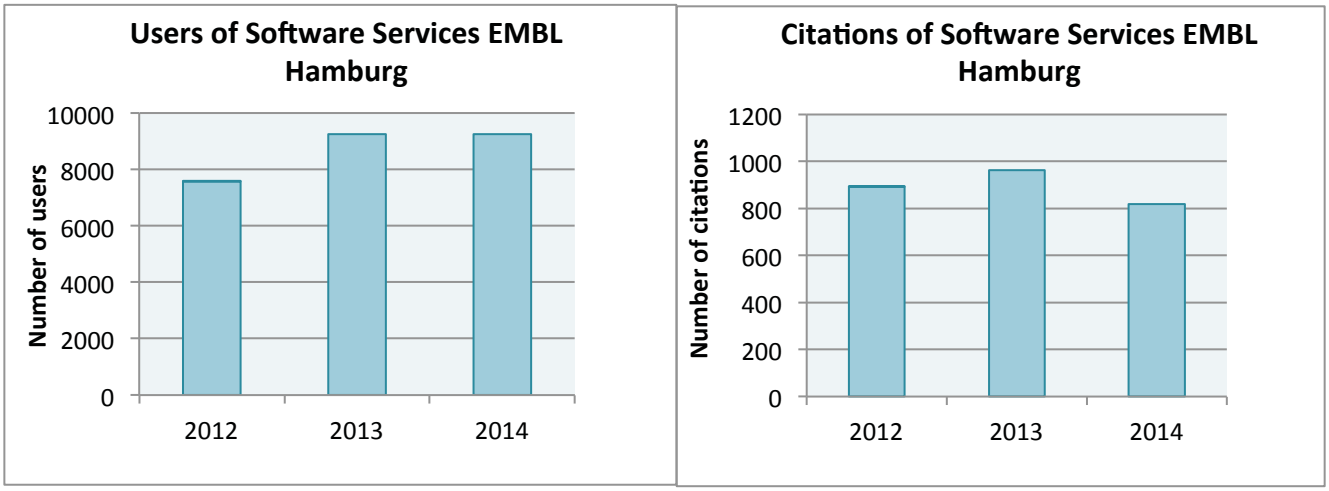


Figure C.2.3 Number of users of sample preparation and crystallisation facilities in Hamburg (HH) and Grenoble (GR). HTX, high-throughput crystallisation; SPC, sample preparation and characterization; EEF, eukaryotic expression facility; ESPRIT, high-throughput expression.

2.3 Software services in structural biology

EMBL Hamburg has world-leading expertise in developing software for structural biology data analysis, which is made available to a large number of users. A total of eight different software packages have been offered during the present Indicative Scheme. By far the most heavily used services are the ARP/wARP package for the automatic assembly of X-ray crystallography data into structure models and the ATSAS suite for interpretation and modelling of SAXS data. Both packages are available for downloading and as remote services. Other widely used software packages developed by scientists at EMBL Hamburg are HKL2MAP and AutoRickshaw, both of which aid in automating computational structure determination following X-ray data acquisition. Altogether, the EMBL Hamburg software packages were used by more than 25,000 users and more than 9,000 laboratories between 2012 and 2014. The overall number of citations of the software packages from that period exceeds 2,400.

Figure C.2.4 Number of users and citations of EMBL Hamburg's remote software services.



2.3 Towards a new generation of highly integrated facilities for structural biology (2017–2021)

Future plans and challenges for EMBL's structural biology services will be to a large extent defined by the availability of new major infrastructure or infrastructure upgrades in Hamburg, Grenoble and Heidelberg. In Hamburg, the European XFEL will begin operation in 2017 and holds tremendous potential for structural biology innovations, ranging from time-resolved femto-second serial crystallography to the analysis of single or small clusters of biological particles. In Grenoble, the ESRF phase II upgrade will convert the ESRF synchrotron ring into a nearly diffraction-limited synchrotron facility with unprecedented low-emittance by 2020. Similar plans by DESY for the PETRA III ring are in an early phase of discussion. The EMBL Grenoble and Hamburg Units will cooperate towards exploiting the scientific possibilities created by these new developments at both sites and to increase access to these facilities through the development of new scientific services.

In Heidelberg, the EM facilities have benefited from recent novel detector and image processing developments, which have greatly improved resolution and have consequently created an enormous demand for access to cutting-edge EM technology among European scientists. Depending on the availability of funds, EMBL is considering broadening its structural biology service portfolio to provide access to EM-based methods to external users to help meet this new demand.

In the future, integrative approaches that use complementary structural biology techniques and hybrid modelling will be important for projects addressing increasingly large and complex systems (Section B.2.4.1). This means that EMBL's structural biology services need to extend beyond single disciplines. We are aiming to achieve this by gradually integrating X-ray-based structural biology methods in Hamburg and Grenoble specifically with EM and light microscopy approaches, chemical biology, diverse '-omics' methods, and computational biology. Integrative approaches will also benefit from local consortia, the Partnership for Structural Biology (PSB, Section F.1.1.4) in Grenoble and the Centre for Structural Systems Biology (CSBB, Section F.1.1.1) in Hamburg. We will foster and develop these partnerships, which we expect to play an increasingly important role in leading structural biology projects at the interface with other key research areas in life sciences.

Taking all these planned activities together, the three EMBL Units will be in an excellent position to further broaden their portfolio of state-of-the-art research services and to maintain future leadership in structural biology service provision in the future. Below we outline specific developments planned for the period of the next Indicative Scheme.

Ultimate storage rings

The ESRF phase II upgrade will allow many new scientific goals to be pursued by enabling room-temperature data collection and *ab initio* phasing of nanometer-sized crystals to time-resolved SAXS measurements. Such techniques will allow molecular dynamics and time-resolved pump-probe studies to be performed and will benefit drug discovery programmes by significantly reducing material and sample handling. Other possibilities include the use of intrinsic elements for nanometer X-ray fluorescence imaging and coherent diffraction imaging, both still in their infancy. Many of these developments can only be pursued on the most modern synchrotron beamlines. EMBL Grenoble and ESRF will actively contribute to fully exploiting the scientific possibilities of ultimate storage sources by developing new methods and software (Section E.1.1.1).

Depending on demand and scientific merit, the PETRA III or ESRF phase II beamlines could also contribute to the optimal selection of X-ray laser experiments, maximizing their chance of success. In the future these new generation sources will be routinely used for understanding protein dynamics in a biological context and enabling new medical discoveries.

Fully automated protein-to-structure pipelines

The introduction of fast pixel array detectors at synchrotron beamlines both at the ESRF and PETRA III has opened new avenues for structural biology and will have a strong impact during the next Indicative Scheme. The new detectors are rapidly changing the way X-ray diffraction experiments are performed. The CrystalDirect technology for automated crystal handling and processing, developed at EMBL Grenoble (Section E.1.1.1.2), offers a unique opportunity to exploit the full potential of fast pixel array detectors. It enables the preparation of crystals for X-ray data collection at a scale that matches the capacity of future MX beamlines and bridges the automation gap between crystallisation and X-ray data collection. These developments underpin a new generation of integrated crystallography services based on fully automated protein-to-structure pipelines. One example is the recently inaugurated MASSIF facility at the ESRF that supports a fully automated service for sample evaluation and data collection capable of processing 300 crystals per day. In the future, this will be complemented by MASSIF-3, a high-intensity micro-focus beamline. Scientists at EMBL Grenoble will also aim to develop a new generation of sample changers and tools to allow higher screening capability (Section E.1.1.1.1). A significant effort will be required to upgrade and integrate multiple data management resources along the crystallography pipeline to support the flux of samples and provide rapid and convenient access to results by the users.

Ambitious developments are also planned for the EMBL@PETRA III beamlines in Hamburg, which in the coming years will be transformed into stable, reliable and state-of-the-art operational facilities. All three EMBL@PETRA III synchrotron radiation beamlines have already demonstrated their suitability for the pursuit of highly challenging experiments that are currently either very difficult to do or not possible elsewhere. The aim is now to build on this and make all three beamlines as productive as possible. This will require ongoing upgrades to the beamlines regarding state-of-the-art instrumentation, additional recruitment of experienced service staff, and the expansion of all beamlines for complete remote use for experiments (as described above). Specific upgrade plans include two major investments within the next two years: the implementation of the CrystalDirect system in the Hamburg high-throughput crystallisation facility, next to the PETRA III beamlines, and the introduction of a state-of-the-art-detector with rapid readout with a target frequency of 750 Hz. We anticipate the number of MX user projects to significantly increase from the early phase records (Figure C.2.2) whereas SAXS applications will have a smaller increase because of their already high popularity during the first user period in 2013.

The new suite of structural biology stations at PETRA III and the ESRF will provide sample analyses at unprecedented scale, which will support highly challenging structural projects in the future and facilitate the use of crystallography infrastructure for non-expert users. The problem of funding future transnational access for academic users, caused by reductions in the funding available to structural biologists from EU I3 programmes, was highlighted in the opening section. Users from industry will continue to obtain access via the existing commercial schemes.

New modes of data collection

The Hamburg Unit is actively implementing serial synchrotron crystallography (SSX), which is based on multi-crystal data collection strategies and makes use of a combination of recent developments from XFEL diffraction experiments and new hardware developed by the EMBL Grenoble and Hamburg instrumentation teams. Scientists in Grenoble will develop a complementary range of services from automated data collection of routine samples (for ligand screening) to more advanced data collection protocols (Section E.1.1.1). In the future, these approaches will be made available to external users. Methods developed at PETRA III in collaboration with EMBL Grenoble will be ready for implementation at the ESRF after its phase II upgrade.

New frontiers for biological SAXS

EMBL Hamburg will capitalise on its leading position in bioSAXS and work towards an integrative SAXS-driven hybrid structural biology portfolio. SAXS can be combined with structural, biophysical and biochemical methods for comprehensive characterisation of biological structures, including their dynamics and flexibility. The portfolio will include SAXS coupled online with advanced purification and characterisation by static and dynamic light scattering to improve the sample quality and facilitate subsequent structural modelling. To study membrane proteins, nanodiscs combined with the separation of empty lipid micelles by in-line purification will be used. Labelling techniques using naturally bound metals and tagged fluorescent dyes will additionally aid future shape determination by SAXS. Combinations of these approaches are particularly useful for the study of the flexible or even unstructured parts of protein samples that are not accessible to high-resolution structural analysis.

In Grenoble, the focus will be on combining SAXS and small-angle neutron scattering (SANS) – in collaboration with the Institut Laue-Langevin (ILL) – as complementary methods for structural biology. These developments will benefit from the ESRF phase II upgrade, enabling smaller sample volumes to be accurately measured using sub-second exposures. On the bioSAXS beamline BM29 at the ESRF a new sample exposure unit will be implemented for complementary biophysical measurements (Section E.1.1.1) and a new high-performance liquid chromatography (HPLC) system will be installed to improve the quality of online bioSAXS measurements. The use of microfluidic technology for more complicated sample measurements and characterisation will be continued and exploited for MX applications.

High throughput pipelines for compound and fragment screening through macromolecular crystallography

The screening of small molecule compounds and fragment libraries through MX is currently employed by scientists in industry and academia for structure-guided drug design and to generate chemical tools. However, it requires the analysis of large numbers of crystals and can consume considerable human resources, which limits the size of the chemical libraries that can be analysed and raises the access barrier for academic laboratories. New services specifically dedicated to support screening of compound and fragment libraries for targets of high biomedical relevance will be developed at EMBL Grenoble based on the new automated protein-to-structure pipelines. Similar experiments have started at EMBL Hamburg and will be implemented and provided to the external research community via the CrystalDirect instrument, next to the MX beamlines at PETRA III. These activities will be pursued in collaboration with the EMBL Chemical Biology Core Facility in Heidelberg (Section C.3.7), which will provide fragment and compound libraries and expertise in chemical biology. This will

require an investment to upgrade the high-throughput crystallisation (HTX) laboratory in Grenoble with new robotic equipment for the manipulation and delivery of large chemical libraries. The HTX lab and the Chemical Biology Core facility in Heidelberg will cooperate to provide coordinated access to a series of complementary services from assay-based and MX library screening, to structural characterisation of small molecule protein complexes and chemical engineering. The new services should lower the access barrier to these techniques, facilitate the analysis of larger small-molecule collections and stimulate faster translation of results from research in structural biology into applications in biomedicine.

Exploiting the X-ray Free Electron Laser in Hamburg for life science research

The European XFEL in Hamburg is constructing the most powerful XFEL in the world with unmatched peak intensity and pulse frequency. There will be six instruments, one of which is dedicated to applications on single particles, clusters, and biomolecules (SPB). The operation of the SPB instrument will be coordinated by the European XFEL and three user consortia: SFX to provide a serial femtosecond crystallography instrument; XBI, which was initiated by EMBL to offer advanced biological sample preparation laboratory tailored to the experimental needs of XFEL life science experiments; and DataXpress, to provide IT hardware and a data analysis software toolkit to interpret the XFEL data acquired.

The strategic aims of the EMBL Hamburg Unit in respect to future XFEL applications in the life sciences are threefold. The first is to establish a research portfolio with the aspiration of providing leadership in specific areas, such as combining serial synchrotron crystallography and serial femtosecond crystallography, the use of cellular components as crystallisation tools and subsequent 3D structure determination. Preliminary efforts along these lines have been made and some data have already been published. The second aim is to take on well-defined and focused responsibilities in providing dedicated research infrastructures to the international research community. Finally, EMBL Hamburg intends to gain direct access to limited XFEL beamtime, in exchange for a commitment to community-oriented responsibilities.

At present, within the framework of the XBI project, EMBL will lead a future sample preparation and characterisation laboratory with the future European XFEL headquarters building. The lab will be about 560 m^2 in size and will allow parallel access to about 20 users. EMBL is currently recruiting a group leader to head this facility. External funding has already been raised by one of the XBI user consortium partners, Uppsala University, from the Swedish Research Council and by EMBL from the German Ministry of Science and Education (BMBF). More resources will have to be provided by the other XBI consortium partners to fully equip the laboratory and to ensure appropriate, stable staffing. A second direction the EMBL Hamburg Unit is keen to pursue is formally joining the SFX user consortium and to provide dedicated staff in the future operation of the SFX instrument, which is expected to host thousands of international users from 2017 onwards. Participation in both of these exciting new ventures will depend on our ability to recruit additional user support staff in Hamburg.

Cryo-electron microscopy service provision at EMBL Heidelberg

Recent technical developments (i.e. direct electron detectors, phase plates, automated data collection, improved processing software) are revolutionising EM by significantly improving the resolution it can achieve. Cryo-EM is increasingly becoming complementary to X-ray based methods as a high-resolution structure determination technique for macromolecular complexes for which crystallisation is difficult or

impossible. Cryo-EM can now reach similar resolution to X-ray crystallography, but without the need to obtain well-diffracting crystals. In parallel, the improvement in data quality obtained by cryo-EM has reduced the size limit of complexes that can be productively analysed, meaning that smaller complexes will in future be increasingly analysed by cryo-EM, while electron tomography and tomography averaging techniques will allow higher resolution information to be obtained *in situ*.

As a consequence, there is an enormous demand in European structural biology community for access to high-end electron microscopes and support in sample preparation, data acquisition and image processing. There is little doubt that this will result in high-end cryo-EM being made available to the community in dedicated facilities analogous to synchrotron sites for X-ray crystallography. EMBL, with its broad experience in providing structural biology services, is ideally positioned to address this demand. Provided sufficient funds become available, the existing EM facility in Heidelberg could be expanded to allow us to cater for external users from EMBL member states. The facility would be fully integrated with and is complementary to the other structure determination and imaging techniques available at EMBL (i.e. X-ray, SAXS, NMR, mass spectrometry and light microscopy). In order to create a multiplier effect that goes beyond simply providing access to a technology that is highly in demand, the proposed facility will have the mission to train users in the operation of the technology as well as in how to set up and operate similar facilities in their home countries.

3. Core Facilities

To successfully address challenging biological questions at the forefront of today's life sciences, easy access to diverse cutting-edge technologies and the accompanying expertise is essential. To support its young group leaders and small research groups and to avoid unnecessary duplication of expensive equipment, EMBL has shared facilities that are open to all its scientists. If capacity allows, EMBL's Core Facilities also provide services to visiting scientists from EMBL member and associate member states, EMBL partner institutions (Section F.1.1.4) and members of the EMBO Young Investigator Programme. Every year there are on average 1,000 users across all the Core Facilities, about 350 of whom are external users or visitors who in the period from 2012-2014 came from 25 different countries.

EMBL currently operates seven Core Facilities at the main Laboratory in Heidelberg in genomics, protein expression and purification, proteomics, electron microscopy, flow cytometry, advanced light microscopy and chemical biology. In addition, there are specialised facilities available at EMBL Monterotondo for flow cytometry and microscopy, which support the local research groups and are tailored specifically towards their needs. The staff members there, however, interact closely with the Core Facilities in Heidelberg to share both expertise and the implementation of best practices.

The technological focus of EMBL's Core Facilities and the services offered by individual facilities depend on the needs of EMBL's scientific community and change with time as new scientific areas emerge and technologies advance. To ensure the Core Facilities meet user demand, each facility has a dedicated committee consisting of users that represent the various EMBL Units. The committees make recommendations on new technology requirements, facility operation and strategy. Further valuable information is obtained by recurrent user surveys, which have consistently shown a very high degree of user satisfaction. All Core Facilities are subject to regular reviews by EMBL's Scientific Advisory Committee. In the 2014 Core Facilities review, their overall performance was rated as outstanding.

To ensure the Core Facilities offer cutting-edge technology, the staff collaborate closely with leading industry partners and instrumentation providers (Section E.4). These collaborations sometimes entail joint technology development and frequently enable access to the latest equipment in its testing phase, sometimes long before it becomes commercially available.

Essentially, there is a cycle in which advanced users work with the Core Facilities to adapt new technologies for a service environment. Once this has been robustly achieved, the new technologies can be made available to the broader community of non-expert users, and this drives EMBL's research on a wide front.

The Core Facilities are also strongly committed to EMBL's training mission (Chapter D). They organise and run advanced courses – an average of 17 per year in the period 2012-2014 – and teach internal and external scientists specific methods and approaches. Many of these courses are organised in close collaboration with industry partners that provide technology, valuable know-how and qualified trainers.

By providing advice on the successful operation of high-level service facilities to laboratories in member states and associate member states, the Core Facilities also contribute to EMBL's mission of international integration (Chapter F). In addition, EMBL's Core Facilities are active members of international networks such as the Protein Production and Purification Partnership in Europe (P4EU), the European Cytometry Network (ECN), the European Light Microscopy Initiative (ELMI) and Core4Life, a pan-European excellence alliance for core facilities. These networks enable the exchange of best practices and information on new and emerging technologies. EMBL's Core Facilities are also active in new pan-European research infrastructures such as Euro-BioImaging (Section F.1.3.2).

The Core Facilities provide the required technological framework for many of the future research plans outlined in Section B.2 of this document and must evolve in parallel with the research activities to adapt to changing technology needs. To illustrate the central role of the Core Facilities in EMBL's research, we provide an Appendix (Appendix 2) that lists a selection of research projects that have been enabled by the Core Facilities over the course of the current Indicative Scheme. Close alignment between research strategy and the Core Facilities is maintained through regular discussions in both the user committees and in senior faculty committee meetings. Equipment purchases are also discussed in these fora and several Research Units usually contribute, together with the Core Facilities, to the funding of new equipment.

Future trends and plans for EMBL's Core Facilities

As the services offered by the Core Facilities develops flexibly in response to changing demands and requirements in the scientific community, detailed developments over the course of the next Programme can only be partly foreseen. Nonetheless, in the following we will outline three major developments that we expect to impact on the Core Facility services over the period 2017-2021.

We expect that growing possibilities to acquire large image datasets – for example, by three-dimensional (3D) high-throughput microscopy, single plane illumination microscopy (SPIM) imaging, and volume imaging by automated electron microscopy methods – accompanied by an increasing need to quantify imaging data will translate into a higher demand for automated image analysis. Consistent with these expectations, the 2013 EMBL Core Facility user survey revealed that users feel limited by the current image analysis capacity. To address this shortcoming, we are planning to strengthen the Core Facilities' image analysis and computing hardware capacity. We will also explore possibilities to hire additional Core Facility staff with exceptional skills in image data analysis to further improve user-tailored services in this area.

As Section B.2.2 of this Programme illustrates, the study of metabolism and metabolites will be of growing interest to EMBL scientists in the next period. Many groups are aiming to visualise, measure, and annotate metabolic activity in various contexts. To support these needs, we are planning to establish a Metabolomics Core Facility that will provide services in quantitative analyses of small molecules including lipids, drugs and natural products based on targeted and untargeted mass spectrometry. In 2014, we recruited an expert in imaging mass spectrometry who, in his position as a team leader in the Computational and Structural Biology Unit, will work on visualising spatiotemporal dynamics of metabolites and metabolic activities. In the next Indicative Scheme period, his

expertise will be leveraged to build up a Metabolomics Core Facility. The plan is to initially build up and offer core services for untargeted and targeted metabolomics. Depending on technical feasibility, user demand and capacity, in the future these services may be complemented by more specialist services in imaging mass spectrometry for the spatial analysis of tissues.

Over the course of the current Indicative Scheme, correlative approaches integrating light and electron microscopy analyses have become heavily used throughout EMBL (Section B.1) and we confidently expect them to become even more popular and powerful in the coming years (Section B.2). We predict that correlative approaches combining single-cell analyses by imaging or flow cytometry with single-cell genome analyses will also become equally important. Therefore, protocols and workflows that enable integrated cross-facility and cross-technology approaches correlating phenotypic with genotypic analyses will need to be developed. We do not expect these to be limited to the correlation of imaging and genome analyses but also anticipate a demand for workflows combining imaging or flow cytometry with, for example, proteomic and metabolomic analyses. We will work together with facility users and industry partners to develop and implement these emerging technology trends in the EMBL Core Facilities.

In the following sections, we outline broad trends and developments that we see emerging within individual Core Facilities over the course of the next Indicative Scheme. These should be considered in the context of the research plans outlined in Section B.2.

3.1 Genomics Core Facility

The Genomics Core Facility (also known as GeneCore) offers state-of-the-art technologies and expertise for cutting-edge functional genomics analyses. The facility provides its services to a broad community of users ranging from individual scientists in small research groups to large consortia (e.g. the International Cancer Genome Consortium). GeneCore is an integral component of many user projects as illustrated by co-authorships in 45 articles and many acknowledgements between 2010 and 2014.

The implementation of ground-breaking massively parallel sequencing (MPS) technology over the past years has strengthened EMBL's position at the forefront of European research by enabling its researchers to start connecting genotype and phenotype. As a result of its broad adoption and high user demand, GeneCore's MPS suite has been upgraded and significantly expanded. In addition to standard MPS applications, GeneCore has been involved in method development and optimisation for single-cell analyses, particularly single-cell transcriptomic and metagenomic studies.

Future plans

Single-cell genomics is a promising new approach with the potential to greatly advance our understanding of the molecular processes of health and disease. As Section B.2.2 shows, single-cell genomics is an area in which EMBL plans to be active in the period of the next Programme. However, this approach is technically very challenging, largely due to high levels of technical noise and the current lack

of methods to quantitatively access the complement of nucleic acids in a single cell and analyse DNA and RNA simultaneously. A further bottleneck is the throughput of the technology. To address these challenges, GeneCore will continue to collaborate with researchers in the Genome Biology Unit and EMBL-EBI as well as with industry partners to develop and implement workflows enabling seamless 'from-sample-to-sequence' processing of single-cell experiments using microfluidics.

Commercial systems with the potential to overcome current shortcomings in sensitivity and feature detection have recently been launched. GeneCore has started a collaboration with Oxford Nanopore Technology, one of the leading technology providers in this area, to explore the possibility of using their systems to provide single-cell genomic services in the future.

As many of the research plans outlined in Section B.2 rely on the integration of various experimental techniques, we anticipate a major future user need to be the integration of genomic data with information obtained by light microscopy, high-content screening, and proteomic analyses. In collaboration with other Core Facilities, their users and industrial partners, GeneCore aims to develop workflows that integrate these complementary technologies. This will require not only the optimisation of methods used for sample harvesting and data generation, but also the development of new analysis workflows enabling data integration.

3.2 Protein Expression and Purification Core Facility

The expression, purification and characterisation of recombinant proteins are frequently rate-limiting steps in projects that require biochemical, biophysical or structural analyses of proteins. The Protein Expression and Purification Core Facility (PEPCF) offers technology, support and advice to its users at each step of the protein expression, purification and biophysical characterisation process. It provides a large collection of vectors and host strains for protein expression and produces high-quality reagents such as enzymes or cytokines that are in common use in EMBL's research groups, thereby helping researchers to save time and significant costs. The facility also supports and trains users in performing biophysical experiments to characterise proteins and their interactions.

As protein expression is an essential component of most areas of molecular biology research, the user community of the PEPCF covers almost all wet lab-based research groups at EMBL. Although the facility has standardised many steps in protein expression and purification, each project is essentially a new challenge as no two proteins behave identically. To define the best conditions for a particular protein, the PEPCF has established small-scale expression and purification screening of multiple expression samples in parallel. New developments and improvements in the tools and techniques used for protein expression and purification are constantly being evaluated and, if useful, implemented by the PEPCF.

In collaboration with the Genomics Core Facility, the PEPCF has initiated and participated in a community project to sequence the genome and transcriptome of the Sf21 cell line, an insect cell line that is used for protein expression by many laboratories worldwide. This project has been coordinated by the Protein Production and Purification Partnership in Europe (P4EU), a network of European protein production and purification facilities that has been active since 2010.

Future plans

The proteins selected for expression and purification will continue to become more challenging: membrane proteins, multiprotein complexes and proteins whose detailed biochemical function is unknown are likely future targets of facility users. Helping users to obtain high-quality reagents (e.g. antibodies) will require expert advice and support by the PEPCF.

The PEPCF will evaluate and implement new expression technologies such as cell-free expression, which could become relevant for areas that require smaller amounts of protein but a high level of automation or miniaturisation (e.g. screening of mutant libraries, single-molecule analysis).

By improving information exchange with other experts in the P4EU network, the identification, development and implementation of useful methods is likely to become faster and lead to a better service for PEPCF users.

3.3 Proteomics Core Facility

The Proteomics Core Facility (PCF) offers services in mass spectrometry-based analysis of peptides and proteins. The main application areas are proteomics and structural biology, in which mass spectrometry is used as an intermediate step towards protein structure elucidation by electron microscopy (EM), nuclear magnetic resonance (NMR) or X-ray crystallography. The PCF supports these activities by offering full proteomic workflows for the identification and quantification of proteins in simple and complex mixtures, and by analysing intact proteins mainly for molecular weight determination. Services range from sample preparation to data collection and bioinformatic data analysis. In addition, the Core Facility staff members maintain all mass-spectrometric platforms at EMBL to ensure stable instrument performance.

Investments in the PCF, with regard to both the installation of new equipment and the implementation of experimental workflows, have aimed at improving performance for in-depth and quantitative proteome analysis. In particular, this has entailed the acquisition of high-resolution mass spectrometers using orbitrap technology, in conjunction with nano-flow liquid chromatography systems for peptide separation. Procedures for miniaturised sample preparation, developed by the research group of the PCF team leader, have been adopted. In addition, methods for stable isotope labelling of proteins and peptides for proteome quantification have been implemented, as have various software tools for data analysis. The PCF now offers a comprehensive service for quantitative proteome profiling, ranging from sample preparation to data analysis.

The application of mass spectrometry in the area of structural biology has focused on determining the molecular weight of intact target proteins. A major

development in the PCF has been the installation of a high-mass Q-TOF in 2014, which overcomes the limitations in sensitivity of the preceding instrument and permits the analysis of large proteins (70 kDa–1 MDa). The analysis of intact proteins has been extended by implementing procedures for determining N- and C-terminal protein sequences as an accurate measure of protein integrity.

Future plans

Future developments will focus on increasing mass spectrometric capacity to meet increasing demand, and on adopting a number of advanced proteomic approaches. These goals have been made more achievable with the recent acquisition of an Orbitrap Fusion instrument, which combines several unique features such as multiplexed proteome analysis by tandem mass tag labelling for the simultaneous analysis of 10 samples in a single experiment. This will allow much more sophisticated sample analysis than before and support more complex biological projects. For instance, several replicates can be combined in one experiment, thereby boosting the statistical power that can be achieved in data interpretation. In addition, samples taken at multiple time points can be simultaneously analysed, facilitating the refined time-resolved investigation of developmental processes or cellular perturbations. It is expected that EMBL scientists will pioneer high-throughput applications of proteomics by combining this mass spectrometric technology with the magnetic bead-based sample preparation protocols recently developed at EMBL. This should appeal to prospective users who have thus far not been served by the PCF, such as those in the Developmental Biology Unit or in the Monterotondo outstation. In addition, the Orbitrap Fusion allows complementary modes of peptide fragmentation and thus provides an important novel tool for the detailed characterisation of proteins and their modifications, in the areas of both proteomics and structural biology.

3.4 Electron Microscopy Core Facility

The Electron Microscopy Core Facility (EMCF) specialises in several techniques for a wide variety of specimens, ranging from molecules to organisms. It provides a full service, including consultations and access to routine and high-end methods and equipment.

Electron microscopy is undergoing a renaissance, partly owing to improvements in correlative and 3D techniques (Section B.2.4.1). Through the association of the EMCF with research groups and teams across its Units, EMBL now performs at the cutting edge of both technologies.

The EMCF offers access to a unique combination of correlative light and electron microscopy (CLEM) techniques that have been developed at EMBL and elsewhere. Similarly, the EMCF is one of the few facilities worldwide to provide routine access to transmission electron microscopy (TEM) tomography and serial TEM tomography. With the recent implementation of automated serial imaging by scanning electron microscopy (SEM) and focused ion beam scanning electron microscopy (FIBSEM) in 2014, ultrastructural analyses can now be achieved on larger volumes, with unprecedented results having already been obtained in cultured cells, small model organisms and mouse tissues.

Future plans

Considering the expected evolution of EMBL research during the next Indicative Scheme (Section B.2), the EMCF will strive to contribute to systematic approaches such as the construction of molecular and ultrastructural cell atlases, and subcellular analysis from cell biology experiments *in toto*. However, in order to do so, data throughput will need to be significantly improved for quantitative data analyses, and high-resolution imaging will need to be performed on multicellular specimens. To be successful, the EMCF will therefore focus specifically on three areas: the automation of data collection; the correlation with other imaging modalities; and the imaging of large volumes. As experienced in the past, the close collaboration between the EMCF and the EMBL Units should continue to guarantee the successful implementation of cutting-edge techniques.

We expect to be engaged in the automation of EM workflows, especially for data acquisition and data analysis. Our efforts towards this will include renewing our portfolio of TEMs and deploying the necessary computing resources. In parallel, the newly introduced SEM-based automated serial imaging techniques will be further developed to increase data acquisition throughput.

Linked to CLEM, these automated strategies will enhance our capacity to focus on and measure specific phenotypes in heterogeneous cell populations and to select and analyse rare or even unique structures in model organisms. CLEM, particularly when involving super-resolution and cryo-EM, will also become an efficient bridge between molecular/cellular biology and structural biology. The EMCF is expected to become a hub for such interdisciplinary approaches.

Large-volume imaging in EM is still a tedious and time-consuming process because of a lack of targeting strategies during image acquisition and the inefficiency of automated image analysis. Nevertheless, the targeting of individual cells or objects in complex samples is becoming accessible due to the combination of 3D light microscopy, micro-computed tomography (CT) imaging and large-scale 3D EM. The preliminary work undertaken by the EMCF and some of its partners (e.g. the Advanced Light Microscopy Facility and research teams in the Cell Biology and Biophysics Unit) indicates that we can make progress in the next period through a combination of automated workflows and correlative approaches.

3.5 Flow Cytometry Core Facility

Since its establishment in 2004, the Flow Cytometry Core Facility **(**FCCF) in Heidelberg has grown in terms of its equipment and the volume of services provided, which reflects the increase in demand from a diverse and expanding user base. The facility has three modular high-speed cell sorters that allow full customisation in terms of laser lines, laser output power and optical layout. This open configuration guarantees optimal instrument-to-sample matching, which is in line with recent trends in terms of sensitivity, sorting recovery and fluorochromes. The analytical capacity of the facility's services recently improved substantially with the introduction of a SORP LSR-Fortessa bench-top cytometer (5 laser lines and 19 detectors). Although particle sorting continues to dominate the services being performed, an increasing number of users profit from the

state-of-the art multicolour flexibility and analytical power provided by this user-friendly piece of equipment.

The facility has been instrumental in supporting several projects that require sorting as a preparative step towards *in vitro* characterisation, including genome, transcriptome and proteome studies. FCCF staff work closely with facility users to develop new techniques, adapting the equipment and identifying new components that can be fitted onto existing platforms to meet their needs. The facility has also been involved in the development of new methods to study DNA repair, fluorescent protein-based molecular clocks, and the assessment of chemical biology-based protein labelling technologies. In addition, there has been an increased demand for multiparametric sorting and analysis, in particular in immunophenotyping.

The FCCF is also engaged in training in-house researchers and member state scientists. This includes participation in basic courses and hands-on training sessions for external visitors, collaborators and pre- and postdoctoral fellows in the early phases of their stay at EMBL.

Due to an increasing demand for flow cytometric characterisation of mouse models, an additional facility was established at EMBL Monterotondo in 2007, tailored to the specific needs of mouse biologists. Services there include multicolour cell immunophenotyping, haematological analysis, cytokine production assays, gene-reporter assays, single-cell cloning and analyses of the cell cycle, intracellular signalling and gene expression. The Heidelberg and Monterotondo facilities collaborate closely and support each other to provide a wide range of services.

Future plans

Large-scale sorting of rare or weakly fluorescent cell populations has become a major user demand. Sorting is used as a preparative step towards genome, transcriptome and proteome characterisation in several biological models. Although the sorting capacity of the EMBL facilities has been able to accommodate the demand for this service so far – as a result of significant time dedicated to method development – further increases will lead to facility saturation and waiting time extensions and a subsequent need to expand capacities. At the Core Facility review in 2014, it was highlighted that this world-leading service requires further support to remain stable and the decision was made to recruit an additional staff member to the facility.

The FCCF now provides a service for new model systems including mouse embryo-derived cells, zebrafish embryo cells, plant protoplasts, salt water sponge (*Tethya wilhelma*) cells, marine annelids (*Platynereis dumerilii*), small vesicles, mosquito haemocytes and yeast colonies trapped in alginate. The challenges that these systems impose arise from limitations in sample size, very low sort target frequencies, dim fluorescence intensities, high background fluorescence, extreme particle dimensions or the lack of available specific fluorescent tags. To aid their analysis and sorting, the facility plans to set up and establish systems combining imaging and standard cytometric measurements. Again, each such service requires a customised protocol design and often machine upgrades.

The facility in Heidelberg is limited in its capacity for multicolour sorting, as the configuration of current sorters only allows up to three laser excitation lines. With

the increase in user interest in this approach, expanding the facility from three to five or more laser-based sorters will be considered.

The facility in Monterotondo is also facing challenges regarding the need for higher multidimensional analysis to address the complexity and heterogeneity of the cell populations investigated, and to complement gene-expression analysis at the single-cell level. The introduction of emerging new technologies such as mass cytometry that circumvent the multiparametric limitations of traditional fluorescence-based cytometry will be evaluated.

3.6 Advanced Light Microscopy Facility

Light microscopy is a key tool for all experimental Units at EMBL and is a central enabling technology for many of the future research plans outlined in this Programme (Section B.2). The Advanced Light Microscopy Facility (ALMF) offers a full spectrum of basic services in microscopic imaging, and has particular strengths in functional and live cell imaging as well as high-throughput microscopy. In addition, it supports users from EMBL and its member states in image analysis. The service activities are complemented by a series of regular internal and external training courses. ALMF staff develop and teach training modules on basic and advanced imaging techniques as part of both internal and external training courses.

Over the past three years, the ALMF, together with its industrial partners, has developed high-throughput feedback microscopy, which allows rapid imaging of samples at low resolution combined with online image analysis. The latter allows the identification of objects of interest that can then be subsequently subjected to detailed imaging at high temporal, spatial and spectral resolution, including, for example, fluorescence recovery after photobleaching (FRAP) or fluorescence resonance energy transfer (FRET) measurements.

This high-throughput microscopy capability of the ALMF has been used in several genome-scale small interfering RNA (siRNA) screening and follow-up projects in areas such as disease mechanisms, membrane trafficking and organelle biogenesis, DNA repair, signal transduction and mitosis (Section B1).

Recent developments in super-resolution microscopy offer as yet unmatched possibilities to collect information on molecular complexes at a resolution beyond the limits of conventional light microscopy. In collaboration with industry partners and researchers in the Cell Biology and Biophysics Unit, the ALMF has developed robust protocols for multicolour super-resolution microscopy. These protocols have been used, for example, to gain insights into the molecular structure of cellular organelles such as the nuclear pore complex (Section B.1).

Future plans

The ALMF is committed to providing advanced light microscopy services that are focused on user needs. Due to the high turnover of EMBL scientists, these needs change rapidly and therefore developments are difficult to foresee in detail.

The facility will continue to invite industrial partners to present new technology developments to EMBL scientists. If there is sufficient demand across EMBL, the

new equipment will first undergo in-depth testing and will be tailored to user needs before we decide whether it is made available to the ALMF user community. Only those technologies that can be made robust in a service environment will be incorporated into the ALMF.

In the past, automation of complex imaging protocols (e.g. FRAP or FRET measurements) has been of great value to ALMF users as it allows them to concentrate on their biological questions rather than the technical hurdles involved. We plan to work towards automation of fluorescence (cross)-correlation spectroscopy, and the correlation of live cell imaging with super-resolution or electron microscopy (in collaboration with the EMCF). These technology advances will be instrumental for some of the future research plans outlined in Section B.2. Automation will require cross-platform software solutions, which will be developed in collaboration with industrial providers of advanced technology in light, super-resolution and electron microscopy.

We also foresee super-resolution microscopy becoming more important for EMBL scientists and visitors over the course of the next Programme. In response to this growing demand, the ALMF will develop labelling and imaging protocols optimised for the gated STED 3X microscope to enable multicolour live cell imaging at sub-nanometer resolution.

The services provided by the ALMF will in future be complemented by services in multicellular imaging and whole-organ analysis at the mesoscopic level and services in image-based modelling at the new EMBL outstation for Tissue Biology and Disease Modelling in Spain (Section B.2.5). Both facilities will greatly benefit from the diverse activities in imaging technology development carried out by the various EMBL Units (Section E.1.1.2). For example, new developments in the area of SPIM and multiview SPIM are being realised in the Cell and Developmental Biology Units in Heidelberg that will improve the usefulness of SPIM for the complex multicellular systems that will be studied in the new outstation. To ensure knowledge exchange and collaboration, regular meetings between Core Facility staff and related research teams at both sites will be organised following the model of the bilateral instrumentation group meetings involving synchrotron beamline experts and other service staff who support structural biology beamline users from Hamburg and Grenoble. The two complementary imaging facilities will reinforce EMBL’s leading role in biological imaging.

3.7 Chemical Biology Core Facility

The Chemical Biology Core Facility (CBCF) is operated jointly by EMBL, Heidelberg University and the German Cancer Research Centre (DKFZ). It supports research groups from all three organisations through the discovery and development of small molecules that either serve as starting points for drug development or as biotools. Biotools are small molecules that can be used to modulate or inhibit biological processes through targeting specific molecules or molecular complexes and thereby provide a better understanding of the underlying mechanisms and pathways.

The CBCF provides all the necessary infrastructure and expertise for small-molecule screening, including an extensive diversity compound library and smaller annotated libraries, and support for computational chemistry approaches such as *in silico* docking.

One major limitation of the facility has been the lack of systematic access to the synthetic chemical optimisation of active compounds that have been identified in screens. Over the course of the current Indicative Scheme, this has been addressed by the addition of a state-of-the-art medicinal chemistry laboratory to the Core Facility. This is staffed by highly experienced scientists recruited from industry and has a focus on automation, productivity and chemical novelty. Through understanding the structure–activity relationships of small molecules, the potency or the chemical properties – for example, solubility or stability – of active compounds can be improved, resulting in better biotools and starting points for drug discovery.

In recent years, there has been a dramatic increase in the demand for screening based on the production of a specific cell phenotype, which has caused severe bottlenecks in the facility. This issue has been addressed through the integration of a 210-plate tissue culture incubator that allows two screens to be executed in parallel, significantly enhancing capacity.

Future plans

A number of projects involving the CBCF concern the disruption of protein–protein interactions. In specific target cases, the facility has been able to find a small number of active compounds in the low micromolar range of activity. However, it has become clear that classical small-molecule libraries may not be the best starting point for the identification of such inhibitors and that alternative chemistries may be more appropriate. To increase the chances of success, the CBCF has implemented a computational structure-based evaluation for cases in which the structures of the target proteins are available. Building blocks for a new chemistry platform technology that may provide a better starting point for this important but diverse target class have been assembled and we will enter a proof-of-principle phase over the next few years. If this is successful by 2017, the new platform technology will become part of the CBCF's service portfolio.

There is a significant and increasing demand for tool compounds that have been described in the literature but which are not commercially available. Synthesis is often multistep, demanding and requiring of a high level of chemistry expertise and significant time investment. Current staffing levels do not allow the CBCF to meet this demand and an additional chemist to work specifically on tool compound synthesis may be recruited in the future.

Synergies can be achieved by the integration of computational chemistry, medicinal chemistry know-how and the structural biology platforms available at EMBL Grenoble and Hamburg. This would provide structural information on how multiple ligands bind to a target protein in the medicinal chemistry optimisation process. Such an iterative pipeline would increase the chances of success and allow the facility to provide improved and more integrated small-molecule development, thereby enabling more rational design of tool compounds and early leads.

3.8 Mouse Transgenesis Facilities

Genetically engineered mice are crucial to many research groups across EMBL and specialised facilities exist at both the Heidelberg and Monterotondo sites for their design and creation. These are closely integrated with extensive facilities at both sites for the breeding, housing, and testing of genetically modified mice, and together they provide a major common resource for the support and training of both EMBL and external researchers in the use of transgenic mice. The Mouse Transgenesis Facilities are involved in the routine production of genetically modified mice as well as in the implementation, development, and dissemination of new technologies associated with mouse embryology and transgenesis. Methods for the derivation of genetically engineered mice are primarily focused on two techniques – pronuclear injection of DNA to obtain random genomic integration, and embryo injection of genetically modified embryonic stem (ES) cells to produce targeted modification of genes. A series of diverse technologies are routinely offered, including embryo transfer, sperm and embryo cryopreservation, *in vitro* fertilisation (IVF), tetraploid complementation, morula aggregation, intra-cytoplasmic sperm injection (ICSI), and blastocyst and morula injection of ES cells. In the past few years, we have seen a considerable increase in the use of ES cell lines created by EUCOMM, the EU-funded part of the International Mouse Knockout Consortium, in which EMBL has participated. We expect to see more such synergies in the next Indicative Scheme, and greater use of these ES resources for purposes other than the creation of transgenic mice.

A major innovation during the present Programme has been the implementation of gene-targeting in hybrid ES cells followed by piezo-mediated delivery of modified ES cells into morula-stage embryos. This method was developed in collaboration with groups at the Mouse Biology Unit in Monterotondo and has the advantage over earlier protocols in that the resulting offspring are essentially pure ES-derived rather than chimeric mice. The routine use of this approach has significantly accelerated the production of gene-targeted mice from approximately 12 months down to 6 months. A second major development has been the establishment of an ES cell gene-targeting service as part of the Transgenic Facility at the Monterotondo site. This has allowed EMBL groups with little experience in embryo manipulation and ES-cell handling and targeting to produce genetically modified mice. The groups at EMBL continue to benefit from the services of the Genome Engineering Facility at the Monterotondo site, which closely coordinates its activities with the Transgenic Facility to offer recombineering-based production of custom gene-targeting constructs. Together, these services make it now routine for groups to design and produce genetically engineered mice from start to finish entirely supported by Facility staff. This resource has opened up the use of customised mouse transgenic tools to groups at EMBL that otherwise would not have envisioned them as part of their research programme.

Future plans 2017–2021

The transgenesis field is poised for major changes due to the introduction of CRISPR/Cas-mediated nucleic-acid editing. The use of this technology to produce knockout and point mutations directly in embryos has been recently established at EMBL and we foresee a rapid switchover from ES-cell gene targeting to CRISPR/Cas targeting for these types of genome targeting in the

near future. Efforts are underway to extend the technology to the introduction of larger targeting cassettes, which are currently inefficient. We expect that a growing number of mouse transgenesis projects will use CRISPR/Cas during the new Scientific Programme. Potential areas of new technology development will entail the application of CRISPR/Cas gene editing to perform high-throughput or multi-site mutagenesis or epigenetic modifications in mouse embryos and an increase in the use of CRISPR/Cas technology in other species.

As before, EMBL expertise and services will be made available to the external scientific community as capacity allows and we expect this demand to increase as CRISPR/Cas gene editing becomes routine. As part of this external service effort, the EMBL Mouse Transgenesis Facilities offer annual training courses in basic and advanced transgenic methods to external researchers. In addition, they also organise courses on laboratory animal science accredited by the Federation of European Laboratory Animal Science Associations (FELASA).

During the next Scientific Programme, we aim to couple our investment in the development of emerging transgenic technologies, especially CRISPR/Cas-based methods, with an expansion of training activities so as to become a reference point in Europe for advanced mouse transgenesis and genome engineering. Moreover, we will leverage EMBL expertise and resources in mouse transgenesis to make a series of unique transgenic mouse lines carrying SNAP-tagged transmembrane proteins available to the community. These will allow innovative, live, super-resolution and electron microscopy-based molecular imaging.

4. IT infrastructure and services

At EMBL, data-driven research involving big data analytics and large-scale bioinformatics service provision have been a reality for over a decade. The EMBL IT infrastructures are at the heart of these initiatives and critically underpin almost all of EMBL's research activities.

IT Infrastructure across the five sites of EMBL is designed to serve local needs. The two biggest EMBL IT facilities are in Heidelberg and Hinxton, both operating advanced IT services (e.g. high-end compute clusters, storage, web service hosting and cloud services). The IT capacities in Heidelberg support the needs of the Scientific Research Units and the Core Facilities as well as providing central administrative IT services locally and to the outstations (e.g. SAP finance, HR, purchasing, reporting & statistics). The Hinxton IT capacity needs are driven by EMBL-EBI's very large external services component, the need to produce the data resources behind these services, as well as to support the computational research groups in Hinxton. Local IT services at the Grenoble, Hamburg and Monterotondo sites are more focused on core infrastructure (e.g. networking, desktop, email, mid-range computing & storage, etc.) and to support the specialised scientific demands of the respective sites.

Table C.4.1: Overview of the five IT services a the different EMBL sites

EMBL Site	IT Service Provision	Central Data Storage	Central Computing	Internet Link
Heidelberg	Basic, Advanced & Administrative	10 PB	10.000 Cores	2x0.5 Gbs
Hinxton	Basic & Advanced	50 PB	50.000 Cores	2x10 Gbs
Grenoble	Basic	0,5 PB	1.400 Cores	3x10 Gbs
Hamburg	Basic	0,3 PB	1.000 Cores	0.1 Gbs
Monterotondo	Basic	0,1 PB		0.1 Gbs

With the expected advances in higher throughput sequencing, imaging and detector technologies described elsewhere in this document and requirements to share large data quantities across EMBL or within international scientific consortia, the challenge of exponentially growing data volumes will continue to pose enormous demands on the laboratory's infrastructure for data storage, compute power and networking. These demands are both technical and financial in nature. Given the current data growth rates, EMBL expects to store scientific data in the order of ExaBytes by 2021, requiring central computing power for downstream analysis or service provision equivalent to tens to hundreds of thousands of CPU cores (Section C.1, EMBL-EBI Services). The IT infrastructures at all sites will be challenged by exponential data growth and the need to develop robust yet rapidly scalable high-performance IT infrastructures which at the same time must remain affordable.

As research across EMBL becomes more and more collaborative and integrated, sufficient external network linking capacities for each of the EMBL sites will be a critical aspect of supporting EMBL's missions into the next Programme period. The sites across EMBL need to be able to access a wide range of internal and external IT resources available through the Internet. The ability of scientists to use remotely hosted scientific or clinical datasets and to exploit services provided at other EMBL sites or externally is heavily dependent on their network link. All sites have been able to maintain satisfactory internal connectivity, but the limitations in external and especially transnational network connectivity in Heidelberg are starting to negatively impact research activities and thus EMBL's international competitiveness.

4.1. Backward look and highlights 2012-2014

EMBL Heidelberg IT

The IT Services in EMBL Heidelberg operate a state-of-the-art 0,8 MW on-campus data centre which is flexibly upgradable in terms of power and cooling to satisfy future capacity demands. This facility hosts the IT infrastructure to serve the needs of 2000+ local, outstation and remote users. During the past years EMBL Heidelberg has succeeded to maintain a lean IT strategy avoiding complexity to the benefit of general robustness and reducing the total cost of operation by standardising on a limited number of high-performance storage or compute platforms to provide central services, while at the same timing giving the research groups the required flexibility for research and development IT solutions. Thanks to significant increases in the IT budget granted as part of the last two Indicative Scheme decisions, this concept has allowed a successful continuous development and growth of this critical IT infrastructure scaled to the exponential pace of data growth.

Today, EMBL Heidelberg hosts a total storage capacity of 10 Petabytes (2 PB in 2012) providing secure access for research instruments, high-performance computing-based data analysis or to a heterogeneous landscape of user computers. The EMBL storage platform offers a 3-tiered storage on-demand architecture based on disk and tape technology. This has generated significant cost savings and allowed EMBL groups to choose storage categories according to their needs in terms of compute performance and data preservation for the most economic solution.

Alongside the exponential rise in research data production, high-performance computing (HPC) has gained huge momentum and today almost a third of all scientists from the Heidelberg Research Units and a number of outstation users use the EMBL HPC cluster, the major 'number-crunching' resource to support large-scale experimental data analysis, bioinformatics or modelling. This high-density compute system was upgraded in a stepwise process during the past years and currently offers access to 10,000 CPU cores, 50+TB of total on-board memory (RAM), an new ultra-fast shared parallel file system (4000 cores, 4 TB RAM in 2010); controlled by a recently installed and much better performing unified workload scheduler.

Virtualization of physical hardware in the data centre is another key asset in the EMBL IT strategy. It helps rationalise the costs for hardware, energy and space while improving the flexibility and the resilience of the server landscape. Today, EMBL

Heidelberg IT Services operate about 300 virtual machine (VM) instances (for comparison there were 80 in 2010) on the VM clusters, fully integrated with the high performance Tier-1 storage. Many of the hosted web services are public and in total serve more than 200,000 internal and external users per month.

To evaluate Cloud computing, specifically the integration of external / commercial IT capacity with the local IT strategy, EMBL is a co-founder and life science flagship of the "Helix Nebula – the Science Cloud" initiative. This EC-funded consortium of 45 partners brings together several EIROforum organisations and other research institutions with leading partners from the European IT industry, small and medium enterprises and publicly funded e-infrastructures. Helix Nebula aims to establish a sustainable federated pan-European cloud infrastructure for science. To study how cloud technology could be used in future on-demand IT provision to handle massive data volumes generated by technologies such as next-generation sequencing or imaging, EMBL has conducted one of the cloud flagship use cases (others were led by CERN and ESA) during a two-year pilot phase of Helix Nebula. This involves collaborative expertise from the Genomics Core Facility, EMBL-EBI research and service groups, and EMBL Heidelberg IT Services.

EMBL-EBI IT

EMBL-EBI manages an IT infrastructure that is spread over three locations: a data centre on the campus at Hinxton provides EBI's staff with the basic and advanced IT services needed to support their administrative, production and research needs; a nearby disaster recovery site on which the data stored on campus is replicated; and a Tier III+ data centre in London that provides the infrastructure for operating the public-facing web services and data resources. The IT infrastructure managed by EMBL-EBI encompasses over 50,000 compute cores and 50PB of storage with the London data centre currently supporting roughly 5 million unique users each year accessing its services from around the world.

A major focus over the last 4 years has been operating and developing the London Data Centre infrastructure. Following an award from the UK Biotechnology and Biological Sciences Research Council (BBSRC) through the Large Facility Capital Fund, EMBL-EBI procured an off-site data centre capability that was delivered through two Tier III data centres in London. These data centres were operated independently in an active/active configuration that allowed incoming requests for data search, retrieval and analysis to be load-balanced across both facilities, protecting users from any temporary service outage.

In recent years, EMBL-EBI has developed and operated an 'Infrastructure as a Service' cloud within its major data centre - the EMBL-EBI Embassy Cloud. This has allowed research collaborators to bring their computational environment and data to an infrastructure where they can undertake their computational analysis activity close to the public and managed datasets and services provided by EMBL-EBI. This facility is now supporting large-scale international collaborations such as Pan Cancer Analysis of Whole Genomes (Box B.2.4) and Tara Oceans (Box B.2.1), and external pharmaceutical companies.

The ever-increasing growth of EMBL-EBI's data archive, which continues to double approximately every 12 months, continues to pose many infrastructure challenges. Recently, the technologies behind the two archive copies held at EMBL-EBI has

changed. One archive copy is now held on a commodity disk-based object store distributed across three sites, while the second copy is stored across two tape systems.

EMBL-EBI's web presence was completely overhauled in 2013 with the launch of a new website and intranet. The design guidelines and templates used to establish these sites were developed after broad consultation across the institute and with external users so that they could also provide a common look and feel across the pages used by individual data resources. These guidelines have now been adopted with minor changes by ELIXIR for the Hub and associated Node websites.

4.2 Future plans 2017-2021

All EMBL sites have been successful to date in keeping pace with the enormous demands of data-driven science and to accordingly scale the individual IT infrastructures. Given the future needs for much higher levels of integration and capacity it will be essential to create extensive synergies by further alignment of infrastructures towards an on-demand IT that allows enhanced and shared provision of IT capacity and services across all sites of the laboratory.

As already mentioned above, one of the most critical IT challenges common across all sites is the need for reliable high-bandwidth external network connectivity. Data intensive science, such as the work undertaken by EMBL scientists and their collaborators from member states, requires large datasets to be accessed, moved and analysed regardless of the data location. Once all of the EMBL sites are interconnected with such links then further synergistic collaboration around the storage and analysis of data across the laboratory becomes possible. The same prerequisite applies for EMBL to leverage the benefits of cloud computing. The technicalities of introducing external high-bandwidth network links are trivial. However, the current incoherent pricing strategies of national research network providers (NREN) are not designed to support multi-sited international organisations and present a major hurdle to ensure progress and retention of EMBL's leading international role in this area. We will need the help of our host countries to solve this problem.

In addition to connectivity, all EMBL sites face challenges around the ever-increasing size of the datasets that they have to handle. A hierarchy of data storage solutions is needed across EMBL, ranging from the high-performance parallel file systems needed to serve data for intensive computing, through less perfomant network-attached high-capacity file systems, to long-term tape-based storage systems. Given the growth of EMBL's data footprint, which we expect to maintain duplication rates of between 12 and 18 months in the next Programme period, the recent developments involving the use of object-based disk and tape stores promises a reasonably-priced solution where both the performance and cost will scale.

As the data volumes continue to grow so does the need for more analysis capacity. While large-scale compute clusters deliver the main compute capability for Heidelberg and Hinxton, the potential of internal and external clouds will continue to be actively explored. Based on the EMBL-EBI Embassy Cloud (significant hardware expansion in 2015) and the integration of commercial cloud services (e.g. through Helix Nebula)

into the scientific process, both sites expect to make greater use of cloud infrastructures in the next Programme period to deliver more flexible use of either internal or external resources and to more rapidly and elastically be able to respond to short-term peaks in compute and storage needs. Such on-demand provision of cloud services likewise promises to benefit the smaller outstations, who will be able, networking-permitting, to use available cloud capacities in Heidelberg, Hinxton or those provided by external service providers.

Further synergies are expected from fostering project-specific collaboration among the IT teams and will for example allow IT solutions previously tested in Heidelberg or Hinxton at larger scale to be transferred to other sites. Monterotondo's IT is a success story, where the IT services such as networking, storage, backup, archiving, virtualisation and email are all achieved by implementing the solutions tested in Heidelberg on a different scale.

Investment in physical infrastructure, to be fully exploited, needs to be matched by an investment in the software, training, support and consultancy across EMBL. A broader use of software frameworks promises general improvements, for example building on Hinxton's success with the integration of myriads of heterogeneous data resources (EBI-Search), high-performance compute job execution (JDispatcher; 100 million jobs/year) or high-end reporting of service usage. Similarly, the planned implementation of single-sign-on across EMBL IT services will simplify access to central IT resources and will be key to allow cross-site use of cloud resources. Further development of new and easy-to-use tools and software platforms for large-scale data movement will be required to support users in transferring bulk data to and from EMBL outstations or outside the institute. Providing secure access and data sharing such tools will also be essential to support correlative scientific workflows involving small interdisciplinary teams or large consortia. Where possible, these efforts will build on established technologies from e.g. high-energy physics, the grid community or from the commercial world.

The challenge facing IT at EMBL is to not only strive for greater integration and use of services within EMBL, but to see how service providers from elsewhere in the research and commercial sectors (e.g. EIROforum organisations, Helix Nebula) can be used to deliver the infrastructure needed to support data-intensive science. Future IT budgets will continue to be ever more critical to supporting EMBL's mission and we will continue to follow the strategy of attempting to supplement the available EMBL resources with external funding where possible to meet the financial challenge. Big data in the Life Sciences, with the genomics and imaging data tsunami, in particular, will be at the heart of all the exciting research and new service opportunities outlined in this scientific programme. EMBL is in a unique position to exploit those and to meet the associated enormous IT challenges provided it succeeds in building the key element of the future EMBL IT strategy - an on-demand IT infrastructure integrating the whole laboratory.

5. Library Services

Apart from access to the latest technologies and instrumentation, staying at the cutting edge of life science research also requires access to the scientific literature. The Szilard Library at EMBL Heidelberg, together with smaller libraries in each of the outstations, provides access to the widest possible range of high-quality scientific publications and related information resources for all EMBL researchers. Selecting specialist scientific literature and tailoring the monograph and serials collection to the needs of EMBL scientists are core functions of the Szilard Library as are choosing the right offers and packages, negotiating and arranging for swift and durable access to digital resources – also in cooperation with national consortia – and securing fair and economical terms. In addition, the library also addresses researchers' needs concerning the many issues related to literature handling and publishing: developing, structuring and promoting the organisation's position towards the growing field of open access; providing advice concerning license and copyright matters; collating and monitoring the metadata of EMBL publications and offering training opportunities for literature databases and literature management tools.

Backward look and highlights 2012-2014

An EMBL-wide survey carried out in 2012 indicated a high level of satisfaction with the library services among EMBL scientists. The survey also revealed a need for more information regarding the library services at the outstations. To address this, the Szilard Library staff visited all outstations, delivered training in literature management tools and provided support for the refurbishment and restructuring of the EMBL Grenoble library.

Over the course of the current Indicative Scheme, EMBL reviewed and streamlined its internal processes for publication payment and introduced a comprehensive research information system, which allows the integrated management of data (and metadata) on EMBL's scientific publications, grants and other types of information.

Future plans

EMBL believes that the rapid and unrestricted sharing of knowledge is a key driver of progress and ensures maximal impact of its research on the global scientific community, medicine, industry and society at large. For this reason, EMBL will renew its open access publishing policy in 2015. This policy follows developments towards open access scientific publication in some of the EMBL member states and is in line with the open access mandates of several major external funders of EMBL such as the European Commission, the Wellcome Trust and Research Councils UK. EMBL uses Europe PMC (PubMedCentral), the database of life science research articles developed and hosted at EMBL-EBI, as its public repository for publications. Europe PMC already supports the mandates of 26 European funding bodies of life science research and hosts content from thousands of international scientific journals.

The open access policy will have an impact on EMBL's internal processes for the publication of scientific papers and in the coming years, resources will probably have to be reallocated from the traditional, subscription-based journals purchase model to open access costs. The Szilard Library will have a key role in implementing,

monitoring and reviewing the development of open access publishing at EMBL and act as a central port of call for all technical questions relating to publications.

The movement towards open access is not restricted to scientific publications. There is already a trend towards open access to research data, which may eventually be made compulsory by research funders. The Szilard Library aims to contribute to the organisation and development of a coherent approach to the collection, standardisation and provision of access to research data at EMBL.

Although the library is well positioned in terms of access to electronic journals, developing and expanding the e-book services will become a necessity in the coming years, and trying to find workable, affordable solutions with publishers in this area will remain a challenge.

Finally, the Szilard Library aims to expand its training efforts, particularly in the outstations, and at the same time raise awareness of the available library services among all EMBL staff.

D. Training

EMBL Mission: Training and inspiring the next generation of stellar scientists

Introduction to EMBL's training mission

Due to its turnover system, EMBL constantly produces high-quality scientists at all career levels from predoctoral students through postdoctoral fellows to independent group leaders and also experienced administrative staff. After their time at EMBL, more than 80% of our alumni return to one of the EMBL member or associate member states. In this way EMBL enriches and fertilises the national systems through a constant flow of highly skilled personnel.

Our turnover system also means that EMBL has a responsibility to provide its scientists and other staff members with the best possible training throughout their stay to equip them for their career after EMBL. In addition, one of EMBL's missions is to provide training to external researchers from member states and beyond. For these reasons, advanced training is core to EMBL's operations and its service to the member states.

Advanced training at EMBL is organised under the umbrella of the EMBL International Centre for Advanced Training (EICAT) and comprises both internal and external training activities. Internal training focuses mainly on the PhD and Postdoctoral Programmes, but also manages undergraduate project students and provides leadership training to new group and team leaders. Furthermore, internal training works closely with Human Resources to ensure that the General Training and Development Programme (GTDP, Section D.1.3), which is available to all members of personnel at EMBL, synergises with training developed specifically for students and postdocs. External training activities span the Courses and Conferences Programme and the Visitor and Scholar Programme and involve close collaboration with the outreach initiative the European Learning Lab for the Life Sciences (ELLS). This initiative targets high school teachers across Europe with the aim of bridging the gap between the latest findings in research and the classroom. The overarching theme for training is to pursue an integrative approach across all five EMBL sites, providing a centre of excellence for training the member states' most promising researchers and enabling broad access to EMBL's expertise and unique ethos.

1. Internal Training

1.1 EMBL International PhD Programme

The EMBL International PhD Programme (EIPP) was founded in 1983. Since its inception, several hundred PhD students have successfully defended their PhD thesis at more than 90 different European universities. In its capacity as one of the first structured PhD programmes in continental Europe, the EIPP has a three-decade track record of serving as a role model for other major institutions across Europe (and beyond). Today, the EIPP welcomes over 50 new PhD students each year and hosts a total of about 230 PhD students at steady state. It remains a very attractive programme, receiving a constant annual number of member state applicants and increasing numbers from non-member states year on year. Roughly 70% of the PhD students come from our member states.

Backward look and highlights 2012–2014

In the past few years, the EIPP has concentrated its activities in four major areas:

- visibility and recruitment
- mentoring, guidance and performance
- vocational training
- integration

Visibility and recruitment

After a phase of adapting the recruitment procedures for predoctoral fellows to an ever-growing number of applicants from all over the globe (Figure D.1), the years 2012–2014 were characterised by establishing an improved e-recruitment database that can efficiently handle more than 2,000 applications per year. Moreover, the EIPP developed strategies to specifically attract the type of young talent EMBL seeks to recruit. This includes new electronic advertisement strategies to target EMBL member state students and a major revamp of the EIPP web pages.

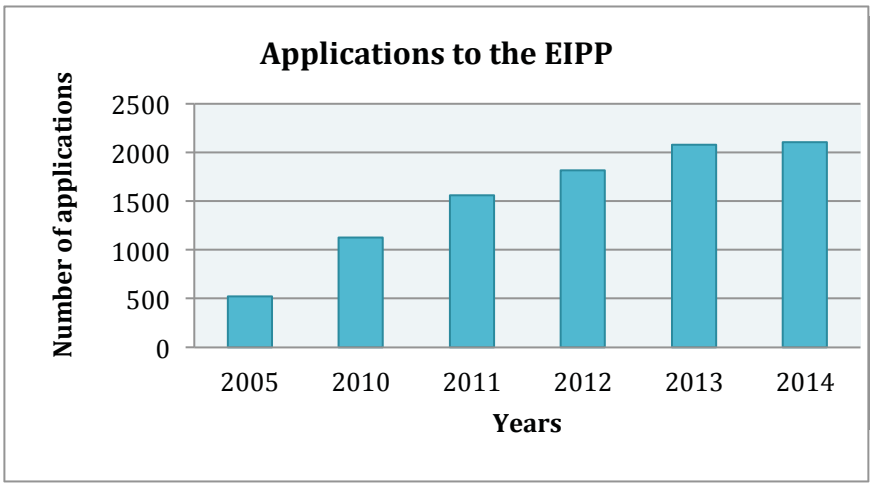


Figure D.1
Applications to the EIPP doubled from 2005–2010 and again from 2010–2014 to reach a total of 2,105 applications in 2014.

Mentoring, guidance and performance

In 2012, a new remuneration scheme for EIPP fellows was implemented that couples payment to performance. This measure was put in place in response to a tendency towards prolonged PhD thesis research (on average 3.9 years until submission by 2009/10) combined with the aim to keep the time to submission within four years. Since January 2012, all students joining the EIPP benefit from:

- a start-up payment to help them settle into their new environment
- a substantial increase in stipend in month 10 of their fellowship pending the successful completion of a first Thesis Advisory Committee (TAC) meeting.
- a stable stipend rate up until timely submission of the thesis and up to a maximum of four years.

The new payment scheme is designed to be cost neutral to the EMBL budget and has been complemented by the introduction of pensions (in 2013) and unemployment benefits (in 2014) for fellows. These financial measures were accompanied by the introduction of in-depth mentoring throughout the thesis and the means to ensure the timely development of an adequate publication strategy for the thesis results. In summary, these new structures put the EIPP at the forefront of scientific employers by providing a sustainable contractual and social infrastructure that encourages young talent from across Europe to embark on a scientific career.

Vocational training

The core course that all EMBL fellows complete in their first year is one of the crucial ingredients for the success of the EIPP. EMBL-wide efforts to nurture, challenge and inspire each new class of students mean that the course is continuously updated. The selection of scientific and non-scientific vocational skills training themes within the course (and beyond) is constantly being optimised in order to stay abreast of the changes seen in the BSc and MSc curricula inspired by the Bologna reforms and to allow for a smooth integration with the national scientific systems across Europe.

Integration

Collaboration, exchange and integration are key to EMBL's success. To implement these principles in the EIPP, all bodies that take decisions or provide feedback on the Programme always include representatives from all EMBL sites. The Graduate Committee, composed of representatives of EMBL's eight Research Units and including two student representatives, is the body that defines the EIPP's strategy and takes decisions on its organisation and operation. To gather feedback from a broader student body, the Dean of Graduate Studies also has regular meetings with 16 student representatives, two from each Unit.

Collaboration and integration extends beyond EMBL. Although EMBL has the right to award its own PhD degrees, it chooses to do so in collaboration with a network of 25 partner universities across 16 EMBL member states. In this way, students of the EIPP establish or maintain connections with national scientific communities, which often help shape their careers beyond EMBL. Reflecting EMBL's interdisciplinary profile, contracts for joint degrees have recently been extended to encompass other faculties other than those in the life sciences.

Table D.1. The EMBL International PhD Programme at a glance
230 PhD students from over 40 nationalities are enrolled in the EIPP
50-plus new students join the EIPP on average every year
1 out of 40 applicants is currently admitted to the EIPP
25 partner universities in 17 countries
2:1 is the average student to supervisor ratio
3.8 years is the average time to thesis submission
>50 annual graduations on average
>95% of students successfully submit their thesis
>90% of EIPP students achieve a first-author publication as a result of their PhD studies
Two first-author papers is the average number of publications of EIPP students
80% of EIPP graduates take up positions in the member states

Future plans 2017–2021

Innovation foreseen for the coming years will target the four major areas described in the previous section.

Visibility and recruitment

The EIPP's aim is to provide excellent training to the most outstanding young talent, primarily coming from EMBL's member states. Consequently, our focus will be on adapting EIPP recruitment strategies to ensure that we attract the best students. We especially aim to increase our visibility in the new member states as well as those with poorer representation in our current student body. Although the EIPP has continued to attract high numbers of member state applications – in contrast to most other programmes across Europe that have seen decreasing numbers of European applicants – we are aware that this challenge is likely to grow and we intend to pro-actively manage our recruitment strategies and provide equal visibility and opportunities at all EMBL sites.

Mentoring, guidance and performance

The high standards of the EIPP for mentoring, guidance and performance will be maintained by establishing enhanced career development support. To better address the needs of today's students, we will collect student feedback through satisfaction and exit questionnaires. The aim is to better support EIPP graduates on their path to both scientific research-based and non-research-based leadership positions.

Vocational training

Efforts will be placed on enhanced research ethics training in two directions:

- enhancing awareness and knowledge regarding ethical scientific conduct, (experimental) research ethics and bio-ethics
- preparing young scientists for exposure to clinical settings in the context of translational research involving patient samples and disease data

Very few of our early-stage researchers have received previous training in these areas.

Integration

We will seek to integrate the predoctoral training even more:

- across EMBL sites
- with EMBL's network of partner universities in the member states
- with the training activities offered by the EMBL Centres (Section B.3)
- between the predoctoral- and postdoctoral communities with their different but overlapping training, mentoring and guidance needs
- within EICAT's broad range of activities
- with EMBL's scientific goals

1.2 EMBL Postdoctoral Programme

EMBL hosts a total of about 250 postdocs from all over the world at steady state. Postdoctoral researchers can normally stay at EMBL for a maximum of five years. However, we currently see an annual turnover of about 60 and an average stay of about four years.

Postdocs come to EMBL via four different entry routes:

Classical Postdoctoral Programme: The classical postdoctoral stream has an annual intake of about 35 researchers who either join with their own personal merit fellowship or are recruited to a position funded by a grant available to an EMBL researcher.

EIPOD Programme: The EMBL Interdisciplinary Postdoc (EIPOD) Programme has an intake of about 20 fellows per year. Following a successful pilot phase starting in 2007, it was fully implemented in the current Indicative Scheme as an initiative to allow postdoctoral researchers to benefit from the interdisciplinary expertise and training available at EMBL and to promote collaboration between EMBL research groups and sites. EIPOD fellows are associated with two, in rare cases three, EMBL groups and work on a collaborative project connecting the expertise and research directions of the groups in a synergistic fashion. The EIPOD Programme has been partially funded by two generous awards from the European Commission's Marie-Curie COFUND Action covering the years 2009–2014 and 2012–2017.

ESPOD Programme: The EMBL-EBI & Sanger Postdoctoral Programme (ESPOD) is shared between the EMBL-EBI and the Wellcome Trust Sanger Institute. It is modelled after the principles of the EIPOD Programme, with fellows being affiliated simultaneously with a computational group at EMBL-EBI and an

experimental lab at the Sanger Institute, and accommodates an intake of two postdoctoral researchers per year.

EBPOD Programme: The latest addition to the suite of interdisciplinary programmes is the EBPOD (EMBL-EBI & NIHR Cambridge Biomedical Research Centre Postdoctoral Programme) initiative, which is also modelled after the EIPOD scheme and promotes projects that apply computational approaches to translational clinical research involving human subjects. The EBPOD Programme has an annual intake of two postdoctoral fellows.

Additional small bilateral programmes are planned at other EMBL outstations, namely in Grenoble and in Hamburg, building on the local Partnership for Structural Biology (Section F.1.1.4) and the Centre for Structural Systems Biology (CSSB, Section F.1.1.1).

The common denominator for all postdoctoral schemes at EMBL is to foster early independence while providing individual mentoring and career development support. Positions at EMBL are attractive to the postdoctoral community first and foremost because of the scientific challenges and freedom available, but the provision of competitive salaries and social benefits is clearly also a factor in the decision of postdocs on where to apply.

Backward look and highlights 2012–2014

The EIPOD Programme is the flagship among EMBL's postdoctoral activities. It has not only been successful in stimulating internal collaboration at EMBL (Section B.3), but also in attracting outstanding candidates.

Since 2012, recruitment and advertising processes for the EIPOD Programme have been substantially improved, for example by implementing a new application database, relaunching the EIPOD web pages, generating a brochure for the EMBL Postdoctoral Programme and establishing standardised interview forms for better comparability between candidates in different subject areas. Moreover, since 2012 all candidates are required to develop a research proposal as part of the application process.

As a result, the number of applicants per call to the EIPOD Programme dropped considerably (Figure D.2) but the overall quality of applicants is perceived to have improved significantly.

.The publication data collected on the EIPOD fellows since the inception of the Programme in 2007, shows that 65% of the EIPODs of a given year of intake publish on average three papers within their five-year contract. The majority of papers are published within the first three to four years of their time as a postdoc.

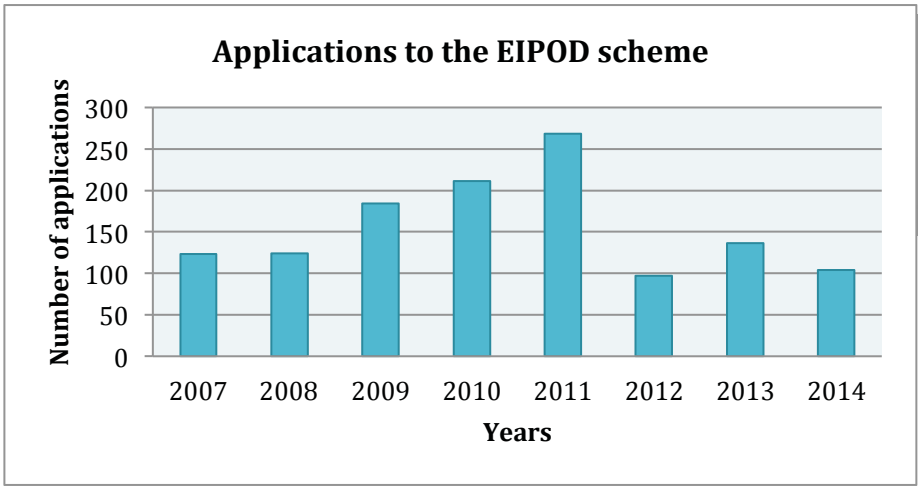


Figure D.2 Applicants to the EIPOD scheme since its pilot started in 2007

The EIPOD Programme serves as an innovation motor for postdoctoral training at EMBL overall. The second-mentor scheme established in the early days of the EIPOD Programme has been made available to all postdoctoral researchers as has a workshop from young EMBL faculty for postdoctoral researchers called 'Preparing for the academic job market'. Further training items and formats have been developed in the area of finance, budgeting and research ethics and are being evaluated in the EIPOD cohort with the goal of making them available to the entire postdoctoral community.

Based on the positive experiences within the EIPP, the postdoctoral researchers at EMBL were encouraged to establish a postdoc representatives' body consisting of two representatives per Research Unit. This helps improve information flow between postdoctoral researchers across sites and identify best practices for postdoctoral training and career development.

Future plans 2017–2021

The following major areas have been identified for further improvement of the Postdoctoral Programme at EMBL:

Enhancing the profile of postdoctoral researchers at EMBL

Many of the postdoctoral researchers at EMBL intend to become independent researchers and leaders but their professional goals are highly individual and therefore difficult to provide guidance on. Nevertheless, we think it is important to create more visibility for researchers at this career stage and we aim to better integrate postdoctoral researchers across Units and sites and within the EMBL training activities.

Enhancing networking opportunities for postdoctoral researchers

To broaden the career perspectives for postdoctoral researchers we plan to establish a broader array of collaborative opportunities to foster relations with EMBL's industrial and academic partners.

Preparing for entrepreneurial opportunities in the life sciences

We will establish a concise curriculum on entrepreneurship-oriented skills to better prepare postdoctoral researchers for their future careers. These skills will

either enable young researchers on an academic career path to more effectively realise the commercial potential of their work or facilitate their smooth transition into the private sector. The course curriculum will be complemented by a 'Corporate Summer School' focusing on commercial development pipelines that connect research with the market. All these activities will be organised in close collaboration with EMBL's industry partners via the EBI Industry Programme (Section E.4.1.1) and the EMBL ATC Corporate Partnership Programme (Section E.4.1.2).

Improving the knowledge base on postdoctoral researchers' needs and expectations

We will establish adapted versions of the EIPP questionnaires for postdoctoral researchers to find out how to better support them on their way to scientific leadership positions, both research-and non-research-based.

Providing meaningful and individualised career development support for postdoctoral researchers

Formalised, professional career development support will be provided through two complementary routes: by establishing a career development advisor for the Programme and by subscribing to the Vitae RDF (Researcher Development Framework) planner. The tools established by Vitae are designed to facilitate independent, professional career management at the postdoctoral stage and beyond and thus prepare young researchers for lifelong learning.

While working towards improved career development support and a curriculum for postdoctoral researchers focusing on intersectorial skills, the primary target of our postdoctoral training will clearly remain the acquisition of in-depth specialist research experience to educate the next generation of scientific leaders for the EMBL member states.

1.3 General Training & Development Programme

The General Training and Development Programme (GTDP) is provided by the Human Resources department and is open to all scientific and non-scientific staff at EMBL. The GTDP offers a broad range of non-scientific vocational skills courses to train staff for their roles at EMBL as well as for future positions. A limited choice of scientific vocational skills courses is made available in collaboration with EMBL's internal training programme. A great strength of the GTDP is its flexibility in terms of tailoring the courses offered in response to requests from EMBL staff and fellows.

In addition, the GTDP offers the 'sponsored study' scheme enabling staff members to enroll in part-time studies to deepen their knowledge in their own field or to prepare them for the next step in their career.

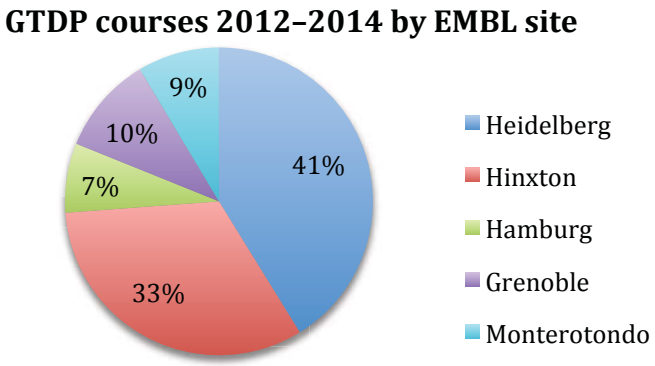
Backward look and highlights 2012–2014

From 2012–2014, about one-third of EMBL staff members participated in at least one of the approximately 80 GTDP courses offered per year (Table D.2). These numbers do not include the language courses in the three official EMBL languages, which account for another 262 participants. During the 2012–2014 reporting period, six new pilot courses were evaluated, three of which took place in 2014, covering the areas of assertiveness, scientific poster design and survey writing. Around 10 staff members have enrolled in the 'sponsored study' scheme per year with courses ranging from Masters degrees in scientific areas, such as immunology and immunogenetics, to accounting qualifications and certificates in marketing.

Table D.2 Participation in the GTDP.

	2012	2013	2014
Total no. of courses	76	76	81
Total no. of participants	608	584	575

Figure D.3 A breakdown of the 231 courses the GTDP organised in the period 2012–2014 by EMBL site.



Future plans 2017–2021

In an EMBL-wide effort involving all stakeholders in training at EMBL we will develop a training portal that allows for simplified access to all scientific and non-scientific (vocational) skills training activities across all sites. Besides creating more transparency for the training opportunities available at EMBL we hope to enhance levels of efficiency and efficacy. The training portal is expected to enable us to:

- maximise synergies regarding the administration of training logistics
- make individual training items more readily available at all EMBL sites
- provide opportunities to collaborate with local partners
- access funding opportunities for EMBL's vocational training portfolio currently not available to any of the individual stakeholder groups

2. External Training

2.1 EMBL Course and Conference Programme

EMBL's Course and Conference Programme enables us to share the very best of EMBL with the research community – particularly within but also beyond our member states. Scientists from all five of EMBL's sites work with their collaborators worldwide to deliver a rich and dynamic schedule representing the full spectrum of EMBL's activities. The Programme is developed in close collaboration with, and is to a significant extent funded by, EMBO, which is by far our most important partner in delivering external scientific training. Other important collaborators are the companies of the EMBL ATC Corporate Partnership Programme (Section F.4.1.2), which also provide vital funding for the events. Further events are organised in collaboration with and sponsored by, for example, the Wellcome Trust or within projects of the EU Framework Programmes. Together with the Visitor and Scholar Programme (Section D.2.2), our courses and conferences form the backbone of the external training offered by EICAT. Roughly 70 % of the course and conference participants come from EMBL member states. Member state scientists also frequently benefit from the approximately 150 Corporate Partnership Programme fellowships that cover registration fees and travel costs to support young scientists.

Guided by the following principles, we aim to provide our member state community with a world-class forum for sharing, discussing and learning cutting-edge biomolecular science:

- **Comprehensive**: We reflect EMBL's full spectrum of activities but in addition aim to stimulate new science by bringing researchers together from a broad range of disciplines, including the physical and social sciences as well as mathematics and computer science.

- **Responsive:** We tailor our content and our delivery mechanism to meet the changing needs of our audience, which includes scientists at all career stages from academia, industry and the healthcare services.

- **User-centric:** EMBL is a global leader in technology development to support research, from beamlines and high-resolution microscopy to data resources. Training our users to make the most of this wealth of supportive infrastructure is a natural extension of user access. Naturally, these opportunities are mostly used by member state scientists.

- **Open:** Our courses and conferences are open to delegates from across the world; we share our own experiences of developing training, learn from others and capitalise on synergies.

Backward look and highlights 2012–2014

Following the opening of the Advanced Training Centre (ATC) at EMBL Heidelberg in 2010 and the South Building at EMBL-EBI in Hinxton in 2013, the current Indicative Scheme saw an expansion in the Course and Conference Programme. The consequently increasing participant numbers necessitated the

continued professionalisation of EMBL's training programme. Taken together, our five sites host thousands of attendees each year, and we foster relationships with many more at off-site training courses and workshops or online. The Course and Conference Programme is continually expanding with 21 conferences and 51 courses in 2014 and even more planned for 2015.

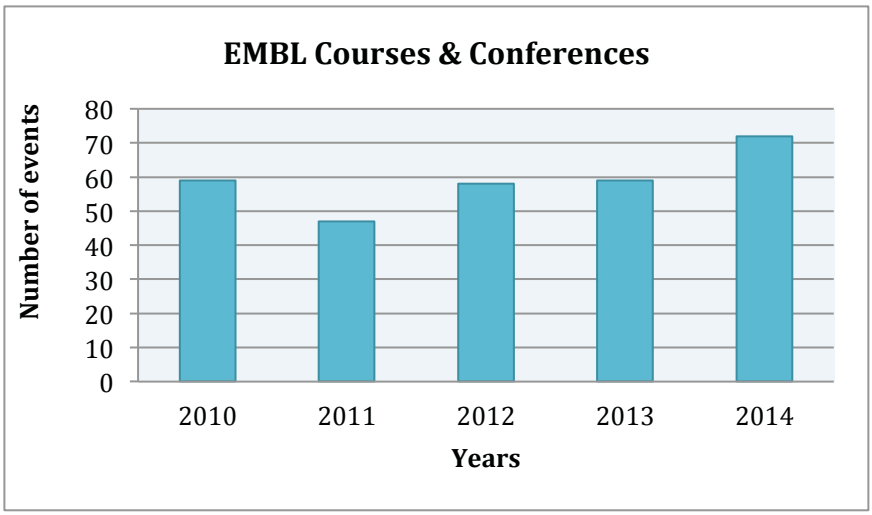


Figure D.4
Numbers of courses and conferences organised at EMBL from 2010–2014

As one of the major highlights, the EMBO | EMBL Symposia have developed extremely well and are among the flagship meetings of the Course & Conference Programme. This results from and also illustrates the excellent collaboration between the two sister organisations.

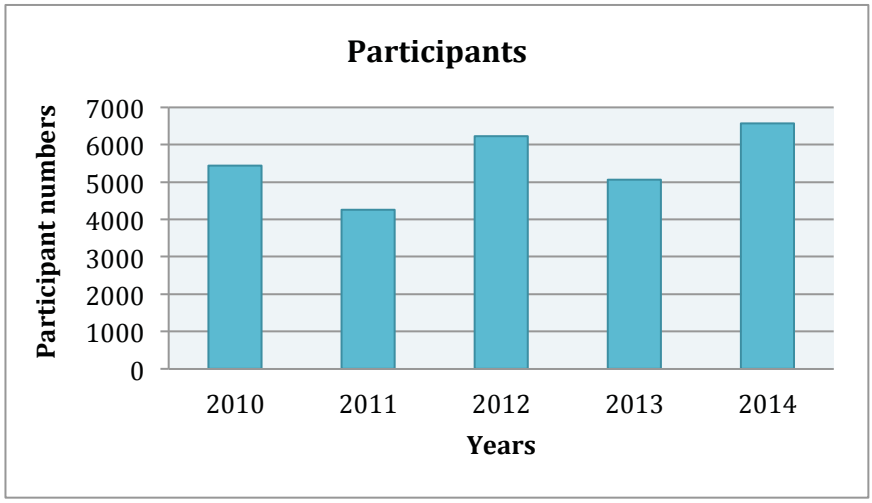


Figure D.5
Numbers of participants for EMBL courses and conferences from 2010–2014

We have also built new and extended existing alliances. Our long-standing partnership with the Cold Spring Harbor Meetings and Courses Programme and our new collaboration with the Wellcome Trust Scientific Conferences are prominent examples. These can involve the joint organisation of events or, for example, the joint organisation of a conference series that is held at Cold Spring Harbor one year and at EMBL the next, thus serving the international community

better. In addition, we involve many partners from industry in our training activities. EMBL operates a Corporate Partnership Programme (Section E.4.1.2) in which many world-leading companies are members, and a variety of advanced training courses are held in close collaboration with partner companies such as Olympus and Eppendorf. In 2014, for example, we launched a new training course series on next-generation sequencing technologies with Illumina, currently the leading provider of this technology. We generally discuss opportunities for the joint development of courses and other potential synergies in training with our industry partners in our EICAT External Training Consultation Panel, to which the companies are invited twice per year. This forum is one example of how EMBL engages at the interface between basic research and industry, and the meeting is highly valued by our corporate partners. Likewise, EMBL-EBI's Industry Programme (Section E.4.1.1) serves large multinational companies, which are significant users of EMBL-EBI's bioinformatics resources and benefit from a bespoke programme of workshops. These companies also collaborate with us to develop new training events that are of special relevance to industry. Finally, we coordinate our training activities with European research infrastructures, including ELIXIR, Europe's new distributed infrastructure for biological information (Sections C.1.3.6 & F.1.3.1), and EuroBioImaging, the emerging research infrastructure for biological and medical imaging (Section F.1.3.1).

Future plans 2017–2021

Broadening our coverage

The EMBL Course and Conference Programme positions itself among the best scientific training programmes worldwide, alongside Cold Spring Harbor, Keystone and Gordon conferences. Led by the EICAT External Training Team and with involvement from scientists across all five sites, we will continue to develop and improve the quality of the Programme, ensuring that it remains at the forefront of scientific and technological development and contributes to new scientific discovery. To this end, resources permitting, we aim to expand the Programme by 50% (in terms of attendee numbers) over the period of the next Indicative Scheme. Although the scientific topics covered will be representative of EMBL's science, our repertoire will reach beyond EMBL. Guided by EMBL group leaders and external advisors, we are constantly developing new topics for conferences and courses; this will enable us to continue to extend our coverage of important scientific areas such as immunology and neurobiology and to explore the boundaries between disciplines. We will pay particular attention to actively engaging with medical and industrial researchers, as well as other communities that have interdisciplinary overlap with molecular biology and the life sciences such as chemists (for drug discovery and for chemical biology, in which we already organise what is regarded as the top international annual conference), physicists (microscopy), computer scientists (bioinformatics, simulation, modelling) and engineers (robotics, microfluidics). We believe that this will add value to our member states by providing a forum for the exploration of new research fields, an incubator for new collaborations and a European centre for cutting-edge technical development.

We will continue to welcome the best scientists from Europe and the world to our Advanced Training Centre in Heidelberg, our state-of-the-art IT training facilities

at EMBL-EBI in Hinxton, and our specialised facilities dedicated to structural biology in Hamburg and Grenoble and to mouse biology in Monterotondo.

To increase the interaction with our member state communities even further, we will encourage EMBL alumni to hold conferences, workshops and courses in our training facilities and we also aim to involve the EMBL Partnerships in contributing to the Programme.

As part of our commitment to EICAT's role as an EMBL-wide instrument, we will increase the involvement of and support for the other EMBL sites, especially Hamburg, Grenoble and Monterotondo, ensuring that training throughout EMBL is of a consistently high quality. These outstations are also planning an expansion of their training contributions. The staff at EMBL Monterotondo, for example, intend to increase their training activities in advanced mouse transgenesis and genome engineering (Section C.3.8), as they aim to become a reference point for these technologies in Europe. There are already very clear mechanisms for exchange and cooperation between the External Training Teams of EMBL Heidelberg and EMBL-EBI in Hinxton. These teams will actively consider how EMBL staff can both contribute to and benefit from training activities.

Training in the member states

EMBL has provided training within and beyond its member states for some years. These activities have grown organically, led by scientists keen to host training focused on a topic of emerging importance to their local research community or by EMBL's International Relations team in an effort to strengthen links with EMBL's member states. Many of these activities have been funded by the hosts (for example, EMBL-EBI off-site workshops) with much of the remainder being made possible through funding provided by EMBO under the auspices of their Courses and Workshops Programme. Despite this already reasonable number of successful off-site courses, there is still considerable demand for such training, which is highly valued by its recipients. Based on our experience of providing such courses, we intend to develop this programme further in close collaboration with EMBO, motivated by the conviction that a more systematic approach to off-site training could be of great benefit to our member states. In collaboration with EMBO we would like to implement a mechanism to meet reasonable requests for courses in the EMBL member states throughout the 2017–2021 budget period, with hosts choosing from a menu of lecture- and computer-practical-based modules. This initiative would benefit the member states by raising awareness of the support that EMBO and EMBL provide to the scientific community; it would help to catalyse new collaborations between EMBL and national research networks in our member states; and finally, it would nucleate communities of scientists with a shared interest within the hosting country. However, this service cannot be fulfilled with the current budget and would be contingent upon additional resources.

E-learning

We intend to establish an EMBL-wide e-portal for training. The portal will provide a single access point for training course videos, conference summary videos, and other training materials amenable to online dissemination. Our intention is to make it easier for event participants to access training material during and after our conferences and courses, and to improve the visibility of the EMBL Course and Conference Programme through linkage of the portal to other EMBL

dissemination tools such as the EMBL YouTube channel, Facebook pages and Twitter feeds. EMBL-EBI has already developed an e-learning portal, which is very actively used, and it would be timely to extend this service across all EMBL sites and training activities. Delegates now expect and look for an online solution to complement face-to-face courses and conferences, and we are aware that other organisations similar to ours are actively developing e-learning solutions.

We will also explore opportunities to develop collaborative solutions. EMBO and EMBL, for example, already collaboratively contribute content to iBiology, a web-platform launched in 2006 by the University of California-San Francisco and Howard Hughes Medical Institute investigator Ron Vale (www.ibiology.org). The platform provides open-access video seminars, magazine articles and educational features by world-leading life scientists with the intention to make this exciting content available to a broader public. Such strategic partnerships will be included in our e-portal activities and constitute an excellent basis for the development of peer-reviewed e-learning content.

EMBL-wide monitoring and integration of external training activities

Beyond the organisation of training, we provide the general acquisition and curation of data for training activities inside and outside of EMBL. The systematic collection of these data allows us to coordinate such training activities when appropriate and enables us to better inform our member states about these activities.

2.2 EMBL Visitor and Scholar Programmes

Backward look and highlights 2012–2014

Through the established Visitor and Scholar Programme, EMBL welcomes approximately 500 visiting scientists per year. The Programme gives researchers from our member states and beyond access to the state-of-the-art Core Facilities and technology platforms at EMBL and allows visiting scientists to carry out experiments and experience the collaborative scientific atmosphere of the Laboratory. More than 60 % of our visitors originate from EMBL member states. For many visiting scientists, their stay at EMBL has had a very positive impact on their research and has therefore significantly improved their future career prospects. In the majority of cases, EMBL visitors apply directly to the group leader of interest and the application process is almost purely bottom-up.

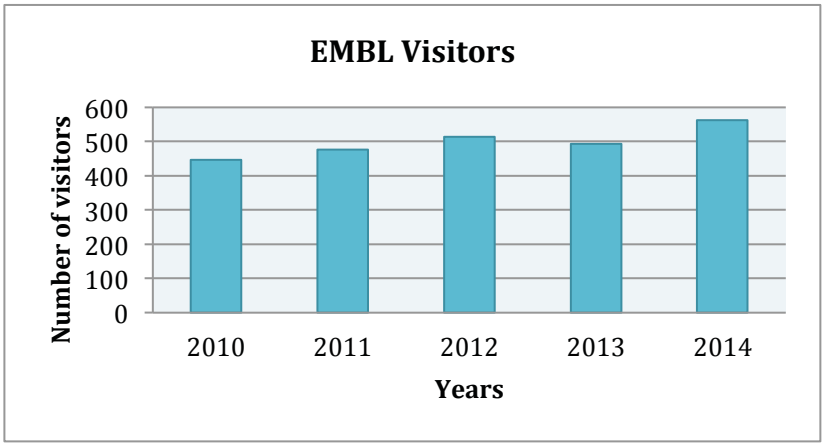


Figure D.6 Numbers of visitors from 2010–2014

Future plans 2017–2021

We plan to further improve the Visitor and Scholar Programme by developing it into a top-quality international exchange programme. Building on the EMBL Partnerships and other existing links, we intend to team up with universities and research institutes in molecular biology as partners for this active exchange. The Programme will therefore also provide greater benefit to the EMBL member states.

In collaboration with EMBL Resource Development (Section F.2.2), we intend to also team up with foundations to provide financial support to our visitors. This will allow scientists from less well-funded backgrounds to benefit from the top-quality training at one of our EMBL campuses. EMBL Resource Development is already in contact with foundations that actively support such opportunities.

Member state scientists will especially benefit from a new type of fellowship that will be awarded through our Corporate Partnership Programme to fund scientific visits to the EMBL Core Facilities. The fellowships will be awarded in memory of the late Christian Boulin, EMBL Director of Core Facilities and Services until his premature and unexpected death in April 2014. Ten such fellowships will initially be given per year and they are reserved for scientists from our member states.

To support the EMBL faculty, the Programme team will provide: i) a standardised application procedure that will help to define the expectations from both sides – the applicant and the evaluator; ii) a pre-filtering of incoming applications, so that the task of evaluators is simplified; and iii) a standardised process to collect feedback from our visitors, so that EMBL is informed about the general success of a visit, remains connected to all visiting scientists, and collects testimonials from the visitors and scholars for reporting purposes. The visitors will be encouraged to raise awareness about the Laboratory and the possibilities it offers to their colleagues and peers.

E. Technology Development, Technology Transfer & Interaction with Industry

EMBL Mission: Driving research, innovation and progress through technology development, interaction with industry and technology transfer

EMBL's main research goal is to pursue fundamental knowledge of the principles and processes governing complex biological systems. Basic science, however, is intertwined with innovation, and advances in knowledge produced by curiosity-driven research may lead to breakthrough methods and technologies that have potential applications.

EMBL is a hotbed of innovation, particularly in the field of technology and instrumentation development, in which scientists constantly challenge existing approaches to investigate increasingly complex biological problems. This chapter will showcase EMBL's activity in this area and provide an outlook on trends and future developments in technology development.

In the second section of this chapter, we present case studies of research-driven impact, which showcase selected applications of EMBL technologies and methods, their progress towards commercialisation and the value they create for life science research beyond EMBL and for society at large.

To facilitate the translation of basic research discoveries into practical applications and make new technologies and instruments developed at EMBL available to the broader scientific community, EMBL engages in technology transfer through its commercial subsidiary EMBL Enterprise Management Technology Transfer GmbH (EMBLEM). These activities will be described in the third section of this chapter.

Finally, the fourth section gives an account of EMBL's growing interactions with industry.

1. Enabling Technologies

EMBL has a long tradition and track record in technology development, which is intertwined with and complementary to EMBL's research (Section C) and service (Section D) missions. In pursuing ambitious research, our scientists frequently push the limits of technology and extend the reach of their toolboxes. At EMBL, engineers and technology developers are embedded in the Research Units and Core Facilities to ease exchange with biologists and customise new instrumentation to researchers' needs. This leads to a mutual synergy through which research drives the technologists and technology developments enable novel approaches to research.

Many of the instruments and tools developed at EMBL are of benefit to the scientific community and are made available to the users of our services in bioinformatics and structural biology or through EMBL's Core Facilities. This arrangement is mutually beneficial as it allows service users to profit from new technologies early, often before they become commercially available, while the users provide valuable feedback to EMBL technology developers. In selected cases, after being evaluated by EMBLEM for their potential, EMBL also makes its inventions commercially available through technology transfer.

The Laboratory's pioneering role in developing new tools, and their value to the entire scientific community, is illustrated by a report published in the journal *Nature* in 2014 on the 100 most highly cited papers of all time, each of which has received more than 12,000 citations[1]. Among these were three methods papers describing work carried out at EMBL and one even made it into the top 10. This shows the extent of the uptake of the tools and methods developed at EMBL, and that its activity in this area is of great value to the scientific community and its member states.

EMBL engages in technology development across a broad range of biological disciplines. However, the most prominent activities cluster around three main areas: structural biology instrumentation, imaging technology and computational technology. This section features a selection of EMBL projects in these three areas that have been instigated or significantly advanced over the past few years and outlines future plans in technology development for the next EMBL Programme. A more comprehensive list of the highlights in technology development over the past five years can be found in Appendix 3.

1.1 Experimental Technologies

1.1.1 Structural biology technologies

1.1.1.1 Synchrotron/FEL instrumentation

As described in more detail in the Structural Biology Services section (Section C.2) of this Programme, EMBL scientists are working hand in hand with beamline engineers at the high-brilliance synchrotron radiation sources at DESY in Hamburg and the ESRF in Grenoble to design and optimise the beamline endstations for structural biology applications. Both sites are in transition towards

next-generation X-ray light sources: the establishment of the world's most powerful X-ray free electron laser (European XFEL) in Hamburg and the phase II upgrade of the ESRF. Future technology development at EMBL Hamburg and Grenoble will be targeted towards exploiting these novel opportunities for macromolecular X-ray crystallography (MX), small angle X-ray scattering (SAXS) and future XFEL-based structural biology approaches. This includes development of technologies and instrumentation in the area of optics and data acquisition and processing. A few selected examples are listed below, whereas others can be found in Section C.2.

Serial crystallography with synchrotron radiation

A major limiting factor in crystallography is the radiation damage caused by the X-ray beam, which normally precludes the use of small crystals. However, a recent ground-breaking experiment on the PETRA III beamline P14 in Hamburg used a combination of highly brilliant beams, versatile diffractometry and high-speed detectors to assemble a complete X-ray diffraction dataset from many small rather than one large crystal. The aim now is to refine this approach by developing, testing, and implementing innovative ways to present many small crystals to the X-ray beam with minimal background scatter. There are also many synergies with the methods developments currently underway in FEL-based serial crystallography that will be exploited, notably in sample presentation and data processing. Conversely, synchrotron-based sample characterisation could become a valuable tool for optimising the use of FELs. Methods developed at PETRA III in collaboration with EMBL Grenoble will be implemented at the ESRF after its phase II upgrade. For example, the CrystalDirect (Section E.1.1.1.2) and MD3 micro-diffractometer technologies developed at EMBL allow *in situ* low background data collection from crystals grown in CrystalDirect plates and serial data collection by fast-rastering over multiple micro-crystals harvested in batches.

An integrated tool set for low-resolution X-ray crystallography

With increasing size and complexity of functional multi-protein complexes comes increasing crystal imperfections. These crystals are highly sensitive to radiation and have inferior diffraction properties that limit data collection and model construction to much less than atomic resolution. In the past, such crystals were often discarded due to a lack of suitable X-ray acquisition infrastructure and protocols for validated structure determination and refinement. With the available spectral beam properties at the PETRA III storage ring in Hamburg, it should now be possible to optimise the beam optics, beam size, brilliance and photon flux to significantly improve the data quality obtained from such crystals and, thus, expand the use of X-ray crystallography to structures that were previously not amenable. We are planning to develop an integrated tool set, which also includes advanced phasing approaches that will use novel heavy atom clusters, new automated methods to deliver heavy atoms into crystals (e.g. via CrystalDirect) and/or model phases of shapes and structures that have been obtained by complementary structural biology approaches such as SAXS or electron microscopy (EM).

1.1.1.2 Sample preparation, characterisation, crystallisation, and sample transfer

Full exploitation of the advanced beamline infrastructures in Hamburg and Grenoble requires complementary developments in sample preparation, characterisation, crystallisation, and crystal transfer technologies. EMBL has a track record in developing technologies that function upstream or downstream of the actual beamlines and is planning ambitious projects in this area during the period of the next Indicative Scheme.

CrystalDirect

Producing crystals of biological molecules, such as proteins, is an essential procedure in crystallography. CrystalDirect is a new automated harvesting and crystal manipulation technology developed at EMBL Grenoble, which is currently being developed for commercialisation in collaboration with industry (Section E.3). CrystalDirect enables the automatic harvesting of crystals that are then either frozen or directly transferred to a beamline without their individual removal from crystallisation trays. It also greatly facilitates soaking with additives (e.g. ligands or heavy atoms). The technology will be improved over the coming years to make it more robust and to increase its automation for use at high-throughput. The tight coupling of crystallisation, ligand screening and X-ray data collection will be a cornerstone in offering fully remote integrated crystallographic services. In addition, CrystalDirect will be further developed to facilitate new modes of *in situ* and serial crystallography data collection approaches, including those requiring heavy-atom derivatisation for experimental phase determination. These advances will enable the full exploitation of the powerful micro-focus beamlines already available at PETRA III and which are being constructed during the new ESRF upgrade.

NewPin

In the past decade, the adoption of the ‘Spine’ sample holder standard for frozen crystallography, developed at EMBL Grenoble, has been essential for establishing compatible beamline automation in Europe. Future MX beamlines at next-generation light sources will require only a few tens of milliseconds of exposure to extract the information contained in a crystal, once again making new developments in rapid, large-scale sample handling central to beamline efficiency. Another project, called ‘NewPin’ was launched in 2009 at EMBL Grenoble to develop a compact and precise sample holder to both increase sample storage density and reduce crystal alignment time at the beamline. The first implementation of the standard is currently proposed for evaluation at pilot sites in the context of BioStruct-X, a project funded by the European Commission to establish a state-of-the-art coordinated and multi-site infrastructure to support access for key methods in structural biology. NewPin will continue to be developed during the next Indicative Scheme, with international academic and industrial partners. The plan is to make this new standard available to the user community in this period.

Crystal imaging at sub-µm scale

The development of serial crystallography using micro-focus synchrotron and XFEL beams allows the sampling of very small crystals. This has raised new challenges in sub-µm-scale crystallisation, especially in monitoring crystal growth

under high background conditions. By adapting instruments for dynamic light scattering and detecting crystal chirality, we are planning to extend our analyses of crystal nucleation and aggregation using a technique known as second-order nonlinear optical imaging of chiral crystals (SONICC), which has recently been developed for biological X-ray laser applications. Protocols for optimised and automated crystal assessment will be devised and made available to the external user community.

Sample delivery through ion traps

In vacuo sample delivery techniques allow for low-background data collection on small amounts of sample and are well known in the field of mass spectrometry. However, an online native mass spectrometer consisting of an electrospray, an ion trap and a time of flight detector also has great potential as a sample delivery device for serial crystallography and single-molecule diffraction experiments. The electrospray is a very efficient delivery device with low sample consumption; the ion trap captures macromolecules or crystals and enriches certain charge states, and it can be used as a purification device; lastly, the time of flight detector can also be used for a 'post-mortem' analysis of the particles that were exposed to the X-ray beam to analyse the consequences of beam damage. We intend to investigate the potential of these devices for online sample delivery and analysis during the next Indicative Scheme.

1.1.1.3 Electron microscopy

In contrast to X-ray crystallography, in which EMBL Hamburg and Grenoble are both actively involved in synchrotron beamline instrumentation development and service provision, EMBL does not engage in microscope hardware development for structural cryo-EM, which is largely carried out within companies. Cryo-EM at EMBL is currently set up as a cluster of research groups with a strong focus on research. In this context, the cryo-EM groups see their current and future role mainly as early adopters of new technologies and in working at the limits of current capabilities (e.g. EMBL Heidelberg is a beta test site for the FEI image phase plate for weakly contrasting unstained EM samples). Within these limitations, the following technical developments in the area of cryo-EM are planned:

- single-particle EM for smaller molecules (~200 kDa) exploiting direct detector and phase plate technology
- continuous software development on structures with helical symmetry with a focus on further improving high-resolution structure determination and facilitating analysis of heterogeneous samples
- data collection schemes/optimisation/image processing for subtomogram averaging to resolve high-resolution structural information *in situ*
- development of algorithms and processing schemes that facilitate the seamless integration of EM data with other structure-determination techniques, in particular X-ray crystallography and spatial restraints provided by nuclear magnetic resonance and mass spectrometry.

1.1.2 Imaging technologies

Imaging technologies are experiencing extraordinary new developments, as recently highlighted by the award of the 2014 Nobel Prize in Chemistry for super-resolution microscopy. Light-, EM-, and X-ray-based imaging are constantly improving our ability to study biological specimens. The integration of these usually separate technologies places EMBL in a good position to start to visualise the dynamic function of life's molecular machines down to the level of individual proteins. Imaging technologies have thus become a major driver in the discovery and elucidation of molecular mechanisms and, along with DNA sequencing, generate the largest amount of research data at EMBL. EMBL is unique in that it develops and applies imaging technologies from the atomic to the organismal scale and we plan to strengthen our leading role in this area through new technology developments by our research groups. In addition, if funding allows in the next Indicative Scheme, we will create a new platform for imaging technology development to facilitate the rapid development and dissemination of new instrumentation to researchers in our member states (Box E.1).

New technology developments will focus on three major areas, driven by biological applications:

i) *correlation* of different imaging modalities to seamlessly bridge scales in resolution from atoms to organisms

ii) *increasing* spatial and temporal *resolution* of light-based imaging at low illumination to push single-molecule imaging from cells into living organisms

iii) *automation* of image data acquisition and analysis of all modalities to enable studies of entire molecular machines, pathways and networks.

With these new technology developments, we will move towards generating atomic resolution movies of life's core processes, which will yield unprecedented mechanistic insight and opportunities for understanding health and disease.

1.1.2.1 Fluorescent dyes and reporters for next-generation microscopy

The major limitation for most fluorescence-based imaging technologies is the lack of optimal fluorescent dyes and probes. EMBL's strategic emphasis on chemical biology (Section B.1.1) has enabled major advances in this area, including small molecule-based reporters for monitoring enzyme activities, which have been applied in cells from patients (Section B.2.3). Second, a high-throughput platform for the rapid development of genetically encoded fluorescent sensors for applications in structural, cellular, or developmental biology has been initiated at EMBL. A new method for labelling the amino-acid side chains of specific proteins in intact cells via click chemistry on unnatural amino acids, which has the potential to replace genetically encoded or affinity-based fluorescent labelling methods for many future applications, has also recently been pioneered in a collaboration between EMBL groups.

In future, chemical biology-derived tools need to be developed further with a view to achieving improved photostability, photon yield, blinking behaviour, and better delivery to living cells to allow optimal use of new light-based imaging technologies in tissue and organ settings as well as in cells. New super-resolution microscopy techniques, for instance, use dyes carrying a high negative charge,

which will require the development of chemical methods to deliver them into cells. In addition, the move towards the simultaneous analysis of many proteins in systems and networks requires the ability to combine many dyes and probes to monitor a large number of molecular species and events in parallel. Taking advantage of the high-throughput microscopy technologies developed at EMBL, we have set ourselves the goal of developing methods that allow the observation of up to 50 different intracellular events in real-time in the same experimental set-up.

1.1.2.2 Correlative imaging: enabling molecular movies and cell atlases

Correlative light and electron microscopy (CLEM) allows high-resolution structure–function imaging of highly dynamic or infrequent molecular processes by collecting time-resolved light microscopy images and subjecting them to EM. A strong emphasis on developing correlative technologies in several EMBL groups and Core Facilities has led to major advances and the establishment of specific workflows linking fluorescence imaging and transmission EM for almost any biological model commonly used at EMBL. The unique capability to integrate kinetic with ultra-structural information on the same object has already led to several breakthroughs (Section B.1.1.1.3). Despite its power, this approach is currently very labour intensive, making it difficult to obtain the large quantitative datasets that are necessary to create the type of dynamic molecular movies and molecular cell atlases outlined in EMBL's future research plans (Section B.2.4.2 & B.2.4.2). EMBL scientists will address this bottleneck by driving automation. Future efforts in technology development will focus on integrating next-generation light microscopy techniques developed at EMBL (Section E.1.1.2.3 & E.1.1.2.4) with new EM imaging modalities (Section E.1.1.1.3) to enable high precision as well as high-throughput correlative workflows. These will enable projects such as those illustrated in Boxes B.2.6 & B.2.7.

1.1.2.3 Super-resolution microscopy: a key tool for structural cell biology

Super-resolution microscopy (SRM), which overcomes the resolution limit of optical microscopy by one order of magnitude, has been one of the most exciting new technologies of the past decade. By reaching nanometer resolution with the molecular specificity of fluorescence microscopy inside intact or even living cells, these methods will rapidly narrow the methodological gap between cell and structural biology. Recognising this, EMBL has already recruited research groups that will make use of these new technologies. Currently, difficulties in obtaining 3D information at nanometer resolution from biological specimens and exploiting the power of single-molecule photophysics for absolute measurements of protein copy numbers are the major limitations of SRM. Our future technology developments will therefore focus on establishing nanometer 3D resolution and quantitative counting of protein numbers. We will also use new fluorescence labelling strategies (Section E.1.1.2.1) and implement high-throughput as well as CLEM (Section E.1.1.2.2) and light sheet imaging (Section E.1.1.2.4) to overcome current limitations and establish SRM as a standard tool for structural cell biology. To infer structural information from SRM data, new data analysis tools and standards will be developed. In parallel, we will integrate SRM data with complementary data (e.g. structures, interactions, electron-density maps) using molecular modelling approaches to provide high-resolution 2- and 3D images of molecular complexes within cells.

1.1.2.4 Light sheet microscopy: towards molecular resolution in organisms

Light sheet or single plane illumination microscopy (SPIM) was first developed at EMBL and has proven to be a very powerful imaging technique owing to its unsurpassed acquisition speed and low light dose optical sectioning, which enables many *in vivo* imaging applications that were previously impossible because of light-induced sample damage. The more recent multiview (MuVi-SPIM) technology, developed in Heidelberg during the current Indicative Scheme, now illuminates and images biological samples from multiple directions, leading to even better resolution, speed, and ease of 3D data processing. A commercial prototype of the MuVi-SPIM has been developed and, in small numbers, is being directly marketed to academic laboratories via EMBLEM (Section E.3). Nevertheless, SPIM still has to overcome two limitations in order to realise its full potential: light scattering inside tissues and the dimensions of the light sheets used for imaging generated with traditional optics. Our future technology developments will address scattering by combining MuVi-SPIM with electronically sliding confocal slit detection on the camera detector, improving image quality and doubling the acquisition speed in multiview setups. This will greatly ease image post-processing, data handling, and storage. We will also implement confocal detection for multi-beam scanning for faster imaging speeds, lower light dose for fast organismal imaging, and develop new real-time 3D data fusion algorithms that will dramatically reduce the requirements for raw data storage and processing. A further aim is to create ultra-thin light sheets to improve axial and lateral resolution in biological samples. This will result in subdiffraction-resolution SPIM in all three dimensions while remaining compatible with video-rate imaging of fast subcellular processes over extended periods due to the low light dose employed.

1.1.2.5 High-throughput microscopy: towards live cell proteomics

High-throughput fluorescence microscopy, originally developed roughly 10 years ago at EMBL, has now reached a level of maturity that has allowed the completion of several genome-scale small interfering RNA (siRNA) screens addressing various biological questions in human cells, such as the molecular regulation of mitosis, organelle biogenesis and maintenance, DNA repair, cell signalling, and disease mechanisms (Section B.1). These and future applications form the basis on which comprehensive molecular and functional cell atlases will be developed.

In the future, we will focus on developing high-throughput imaging technologies that allow visualisation of protein networks carrying out essential functions in human cells. These new imaging-based live-cell proteomic techniques will rely on the ability to fluorescently label all copies of an endogenous gene by genome editing. Intelligent imaging by real-time image analysis will allow us to automatically reconstruct cells at the highest possible optical resolution in 3D over time while the function of interest is being carried out. Single-molecule (fluorescence correlation spectroscopy) calibrated imaging will enable determination of absolute abundances, subcellular distributions, and fluxes of proteins in 4D. The integration of data from hundreds of proteins using spatiotemporal landmarks, image analysis, machine learning, and vision will allow us to reconstruct a 4D view of dynamic protein networks at work and predict the formation, composition, function and disassembly of protein complexes. These function and interaction predictions will in turn be validated systematically in live

cells by sampling pairwise interactions with high-throughput fluorescence cross-correlation spectroscopy. This will provide us, for the first time, with a comprehensive view and predictive model of specific molecular processes through an entire cell.

Box E.1: EMBL Imaging Technologies Development Platform - A European gateway to cutting-edge biological imaging technologies

Over the past years we have witnessed an explosive increase in novel imaging technologies in the life sciences. There is a big demand for these technologies among European scientists. However, the availability of commercial counterparts of microscopes built in research laboratories typically involves a lag of 5-10 years. Furthermore, as most new imaging technologies are very data-intensive, deep integration with bioimage informatics is a prerequisite to exploit their potential. This limits the use of new instruments to a small circle of developer labs and close collaborators, because only once technologies have been commercialised and tested are they made available through the Core Facilities at EMBL or in national research institutions.

If funds become available in the next Indicative Scheme, EMBL plans to address this gap through the creation of a new Imaging Technologies Development Platform that provides access to imaging technologies to a wider community prior to commercialisation. To support user access, the platform will bundle expertise in the use and adaptation of these cutting-edge instruments, custom sample preparation and mounting as well as data handling, processing and visualisation to allow users to achieve optimal results. This expertise will be shared with the community through a web-based interface, which will be openly accessible. The platform will focus on imaging technologies invented and/or developed at EMBL, ranging from light sheet and super-resolution microscopy to correlative light and electron microscopy. This new activity will build on our track record in the development of imaging technologies, our widely recognised imaging facilities for commercial microscopes, close collaborations with microscopy companies and efficient technology transfer mechanisms.

1.2 Computational Technologies

Computational technologies are pervasive in all areas of life science research. Software and other computational tools are required to operate cutting-edge instrumentation and to process, analyse, store and share the data it produces. EMBL scientists develop and adapt computational technologies in many of its research areas. In this section, we feature two prominent example areas to illustrate the kind of computational technology development in which EMBL will engage during the period of the next Indicative Scheme. Examples of software development for structural biology – the third main area of computational technology development at EMBL – can be found in the Structural Biology Services section (Section C.2) and past achievements in Annex 3.

1.2.1 Enabling bioinformatics standards, tools, resources and IT infrastructure

While much of our biocomputational efforts go directly into biology-driven projects, as described in Section B.2.1, their success heavily depends on an enabling infrastructure. Developing the infrastructure itself presents multiple challenges. Firstly, data collection and archiving in a form that is re-usable is far from trivial in the age of Big Data. Integration of each new data type into a growing worldwide body of resources requires considerable effort. Their maintenance alone requires constant adaptation of technologies and the development of new methods. EMBL will continue to actively engage in leading such developments. For example, archiving and updating Big Data is constrained by storage and compute capacity, so compression algorithms and cloud-computing methods will be actively pursued. The integration and interpretation of the archived heterogeneous data also bring new challenges. Although many different methods for data-mining exist, the increased complexity of data types will require new statistical frameworks to be developed or tools that simplify pipeline development allowing customised analysis of a particular combination of data types, as described in Section 1.2.1.1 for multi-omics data. In addition, new visualisation concepts for Big Data need to be developed to allow intuitive reviewing of results. With these advances, the comparability of submitted data has to be ensured, for example by the development of standards and frameworks for the inter-operability of methods.

1.2.1.1 Statistical methods for multi-omics

The advent of routine molecular data gathering through '-omics' (genomics, transcriptomics, proteomics and metabolomics) technologies has transformed molecular biology over the past decade (Section B.2.2). Increasingly, these high-throughput techniques are being applied systematically to a set of samples to provide many molecular readouts from a biological process. This produces rich, multi-dimensional datasets that show great promise in illuminating biology but come with a host of computational challenges, which EMBL is well placed to play a leading role in solving.

The first challenge involves sensible data management. EMBL has extensive experience in a variety of large-scale systematic projects (including the 1000 Genomes Project, the ENCODE project, the BLUEPRINT project, the Pan-Cancer Analysis of Whole Genomes (Box B.2.4), the International Human Microbiome Project (Box B.2.12) and the Human Induced Pluripotent Stem Cells Initiative (HipSci, Box B.2.5)). Careful and constant development of data flow, quality assurance and management is required for these projects to be successful. Much of this work takes place at EMBL-EBI, where the lessons learned from the various projects have been further institutionalised and shared worldwide (e.g. in the effective use of sample ontologies) by the Bioinformatics Service teams (Section C.1).

The second challenge is the effective development and deployment of 'first line' analysis, which often processes the raw data into useable information per sample – examples include RNA quantification, Chip-seq peak calling, metabolite identification and genotype calling, including structural variations. EMBL is a leader in some of these areas (e.g. structural variation, differential RNA-seq assessment, metabolite identification) and, regardless of whether the

development is in-house or external, always aims to deploy world-leading algorithms in robust pipelines. We expect to continue to have a mixture of developing methods to match the innovation in measurement techniques and, whatever the provenance of the methods, to deploy the best method in the pipeline towards the production and storage of useful data. We expect to see a shift in data production methods over the period of the next Indicative Scheme towards long-read-based RNA-seq and other nucleic-acid readouts, more sophisticated proteomics methods – in particular integrating with genomic variation – and more sophisticated metabolomics technology.

The third challenge is the statistical integration of heterogeneous information, for which there is no single framework: the analysis methods need to be tailored both to the question at hand and to the datasets. However there are common recurring themes. A key issue in many analyses is to quantify, characterise and separate out different sources of variance in the measurements, such as experimental confounders or recorded covariates (such as age or weight) in cohort studies. The classification and quantification of these variances is key for downstream success, and EMBL scientists are pioneering the development of sophisticated ways to capture this, for example the PEER and PANAMA software that infer hidden determinants and their effects from gene-expression profiles using factor analysis methods.

Single-cell genomics demands particular attention to sources of variance: first, there is a far higher level of technical noise than in data collected in bulk from many cells; and second, there are a number of known cellular processes, in particular the cell cycle, that influence the cellular data obtained. EMBL scientists have successfully developed methods to quantify and model technical noise and cell-cycle effects (Section B.1.2.3). The close interaction of statistical and experimental scientists, fostered by the EMBL structure, allows for creative exploration and understanding of additional sources of variance and will be key for our further progress into single-cell biology in the next Indicative Scheme (Box B.2.2).

These different techniques require multidisciplinary teams that span experimental techniques through engineering to statistical methods. Over the past decade, EMBL has been very successful in the (bottom-up) creation of these integrated teams, with strength and depth across the organisation leading to spontaneous collaboration, often involving multiple groups. Over the coming Indicative Scheme, EMBL will further develop this area, with both individual innovative methods in first-pass analysis and statistical integration, and example projects that pull together this information for specific biological readouts and illustrate their potential. As well as demonstrating the feasibility of such methods, EMBL shares its expertise, for example via the alignment of EMBL-EBI Service teams (Section C.1) with the needs of these integrated methods, or the participation in European research infrastructures such as ELIXIR (Section F.1.3.1) and Euro-BioImaging (Section F.1.3.2) to enable the transfer of our know-how to the broader community in an efficient way.

1.2.2 Computational technologies for imaging

1.2.2.1 Software for imaging technologies

EMBL has a strong record in developing new imaging instrumentation (Section E.1.1.2). As an integral part of the optical developments, new software has been developed to control the instruments and do the on-board analysis of the raw detector data to allow quick visualisation and real-time interactive or automatic control based on the image data. There are currently three areas with major achievements in instrumentation software – light sheet microscopy, high-throughput imaging and super-resolution microscopy– with additional development efforts in other technology areas, such as correlative light and electron microscopy. In light sheet microscopy, multiview and higher-resolution capabilities have led to real-time 3D data fusion algorithms allowing real-time visualisation of large 3D volumes. In high-throughput microscopy, EMBL-developed software (e.g. Micropilot, Micronaut) allows fully automatic large-scale imaging campaigns with high-end imaging modalities obtained by collaboration with industry partners such as confocal and fluorescence correlation spectroscopy. In super-resolution microscopy, quantitative determination of the resolution and precision of the acquired data has been achieved and new developments focus on quantitative stoichiometry measurements based on the photophysics of the fluorophores under observation. In summary, computer science expertise at the interface of imaging instrumentation and image analysis has been and will continue to be vital to further strengthen EMBL's leading role in imaging technology development.

1.2.2.2 Image analysis for light microscopy

Recent imaging technology captures extremely rich quantitative, multidimensional molecular and structural data in 3D space and over time across various scales of biological organisation, from single cells to whole organisms. To interpret these data, deep computational analysis is required to allow the extraction of biophysical and biochemical parameters and thus enable *in situ* structural biology and proteomics experiments. This is the aim of image analysis, or bioimage informatics, a rapidly expanding emerging field of computer science, with many opportunities for innovation and new technology development.

Image analysis requires various types of hardware, including high-performance computational power, large-capacity data storage, and high-speed connectivity to transfer the large datasets. It also needs new algorithms for image processing, software to rapidly link algorithms into workflows and, finally, software for image data organisation into databases and public libraries. EMBL has significant expertise in bioimage informatics in both the SCB and CBB Units, and this is penetrating into other Units across EMBL. This expertise is broad, with foci driven by the biological questions under study and encompasses methods of analysis of: electron microscopy data, such as helical and subtomogram averaging for structural interpretation; super-resolution microscopy data for structural and proteomic interpretation such as Fourier Ring Correlation; and light microscopy data such as anomalous diffusion modelling, cell architecture and organism lineage reconstruction, for biophysical and cellular and organismal structure interpretation. Experts at EMBL will continue to provide image analysis courses for experimentalists to rapidly disseminate the new and powerful tools they develop. The future recruitment of a new group or team focusing on new

algorithm development in bioimaging informatics is highly desirable and a priority for the above-mentioned Units.

1.2.2.3 Mathematical modelling

Computational tools to simulate biological matter at the cellular and multicellular level are an emerging technology that will play an important integrative role in future molecular biology. They will be an essential pillar of the quantitative revolution in biology driven by imaging technologies. In this computational framework, it is necessary to merge biological knowledge with the laws of physics and chemistry relevant at the scale of interest. In the future, researchers will be able to compare their latest findings with a public 'community' model that represents the state-of-the art quantitative molecular knowledge, in the form of a simulation running on a public server. Ultimately, every pertinent and reproducible piece of experimental information should feed the community model, which would be the ultimate information integration and dissemination tool. Using computer simulations in this way will also hugely facilitate the way researchers work together. Simulation software can be organised modularly, through which each module is developed and reviewed by the respective experts in the field to guarantee highest content quality. At the coordination point, modules are combined, unifying all the expertise into the production of a functional model with a power that printed communication cannot achieve. Continuing to push the development of computational tools for molecular biology and providing the technology for their coordination and integration will be important to EMBL's future mission.

2. Case Studies of Research-Driven Impact

In this section, we showcase selected applications of technologies and methods initially developed at EMBL, and mark the progress towards their commercialisation, which ultimately benefits the community at large. The translation of research-derived innovations into marketable products is a notoriously lengthy process, often requiring many years. For this reason, although the application developments presented here took place in the 2012–2014 period, some of the underlying tools and technologies were invented significantly earlier.

2.1 SP3 proteomics sample preparation

Scientists in the Genome Biology Unit have developed a paramagnetic bead-based method for proteomics sample preparation. This method overcomes the shortcomings of current procedures, which show significant limitations with regard to reagent compatibility, sensitivity, and throughput. The so-called Single-pot Solid-phase-enhanced Sample preparation method (SP3) allows for time-

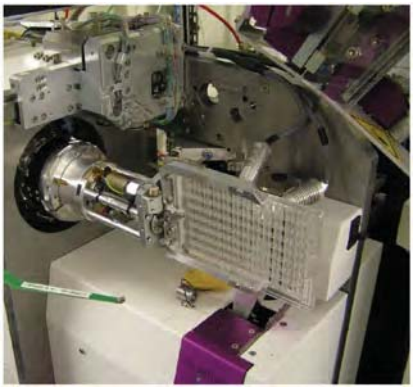
efficient and cost-effective enrichment of proteins or purification of peptides for downstream analysis in high-throughput automated set-ups. The new method addresses a major bottleneck and facilitates next-generation proteomics research, and will thus have a major impact on a large fraction of the molecular life science community. EMBLEM filed a European patent application covering key aspects of this technology in early 2014, and negotiations with potential licensees have begun.

2.2 New SPIM developments

A good example of technology development at EMBL is single plane illumination microscopy (SPIM, Section E.1.1.2.4). The technology was invented and developed at EMBL and led to a first patent application being filed in 2002. An exclusive license was granted to Zeiss and resulted in the launching of the first commercial SPIM microscope, the ZEISS Lightsheet Z.1. SPIM has revolutionised the field of light microscopy and live imaging. It allows scientists to study large, living specimens from many different angles, under physiological conditions and with minimal harm to the specimen. This is well-suited to three-dimensional imaging of transparent tissues or whole and living organisms as specimens are exposed only to a thin plane of light, and photobleaching and phototoxicity are thereby minimised.

A more recent development is the multiview MuVi-SPIM microscope that images biological samples from multiple directions, a commercial prototype of which has been developed in-house and, in small numbers, is being directly marketed to academic laboratories via EMBLEM. In 2013, EMBL researchers also developed a SPIM microscope that allows the fluorescent imaging of live mouse embryos for the first time. These two MuVi-SPIM systems are envisioned to form the basis of a new EMBL spin-out company, which will further develop and advance these technologies, making them broadly available to end users.

2.3 CrystalDirect: fully automated crystal harvesting and analysis



CrystalDirect, developed at EMBL Grenoble, enables full automation of the crystal harvesting and presentation to the beamline process without the need to transfer the crystal out of the crystallisation plates (Section E.1.1.1.2). EMBL aims to establish CrystalDirect© for routine use in macromolecular crystallography. A series of pilot projects are being performed to develop new applications, including automated high-throughput crystal harvesting for ligand screening based on co-crystal structures. The first prototype is now in operation at EMBL Grenoble. A patent application was filed by EMBLEM in December 2012 and the intellectual property (IP) was successfully licensed to the French company Maatel in May 2014. The company plans to launch the first commercial system in 2015, which is a remarkably short time for a development pipeline of this type.

2.4 Infinity MAB: selection and stable propagation of antibody-secreting hybridoma cell lines

Scientists at EMBL Monterotondo have initiated a project, financed by the EMBL Technology Development Fund (TDF), that aims to develop new mice that allow for antibiotic-mediated selection of monoclonal antibody-producing hybridoma cells and maintenance of industrial hybridoma stability. The project addresses a significant market as transgenic mice remain the most important source of monoclonal antibodies, particularly in the diagnostics and life science research sectors. When combined with existing transgenic technology platforms, this approach is expected to significantly contribute to the future development of diagnostic monoclonal antibodies. A patent application protecting the overall approach and the use of such cell lines for the production of monoclonal antibodies was filed in November 2014. The IP will subsequently be licensed to interested partners in the biotech and diagnostics industries.

2.5 Acquifer AG: an EMBL start-up building robust screening microscopes and developing solutions for large data handling

The long-term stability of both the optical and mechanical parts of a microscope is important for screening applications. EMBL scientists working with EMBL alumni at the Karlsruhe Institute of Technology (KIT) have designed a compact screening microscope comprising key features indispensable for robust long-term imaging. The microscope includes stable mechanical parts such as positioning stages and focusing units, as well as a moving optical system yielding reproducible results over days and months of operation.

Based on this innovation, the company Acquifer AG was founded jointly by KIT and EMBL in 2012, with the aim of further developing the system and exploiting the EMBL screening microscope technology in the field of life science research. A joint patent application was filed by EMBLEM and KIT and licensed exclusively to Acquifer AG. The company, located on the KIT campus in Karlsruhe, aims to commercialise the screening microscope in combination with platforms for large image dataset handling. Specific research applications include its use in zebrafish, fruit fly, and yeast models. The company has recently completed a significant investment round to finance worldwide expansion of its activities.

3. Technology Transfer

The technology transfer activities of EMBL are carried out exclusively by EMBLEM Technology Transfer GmbH (EMBLEM; http://www.embl-em.de), the wholly owned limited liability company of the Laboratory. Established in 1999, EMBLEM identifies, protects and commercialises intellectual property developed at EMBL, by EMBL alumni and by third parties including the life science faculties of the University of Heidelberg. The proactive technology transfer approach ensures the rapid commercial development of promising innovations while concomitantly securing the free dissemination of knowledge for basic research purposes. Together with the associated venture capital fund, managed by EMBL-Ventures, EMBLEM also helps in the structuring, creation and financing of start-up companies in the life sciences in any of the EMBL member states.

The success of the technology transfer activities is reflected both in the broad engagement of scientific staff – more than 500 EMBL scientists are on record as inventors with nearly 700 invention records – and in the 300-plus satisfied commercial licensees of EMBL technologies, more than half of which are recurring customers interested in establishing a long-term relationship with EMBL and EMBLEM. EMBLEM manages a portfolio of over 250 individual granted patents and patent applications, 130 copyrights and trademarks and 16 spin-out companies (Box F.3).

3.1 Backward look and highlights 2012–2014

At just over half-way through the current Indicative Scheme, all milestones towards the goals set out in the EMBL Programme 2012–2016 have been achieved. During the 2012–2014 period (Box E.2), 130 invention disclosures were received, 51 priority patent applications were filed and 11 software copyrights were secured. More than 801 license and collaboration agreements were executed and annual turnover has increased by 47% (2014 compared with 2011). Cumulative turnover for the period 2012–2014 is in excess of EUR 16 million. Two proof-of-concept projects, ‘Infinity MABs’ (Section E.2.3) and ‘Phosphatase of Regenerating Liver-3’ (PRL-3), both developed in the Genome Biology Unit, were funded in this period. Three EMBL spin-outs (Box E.3) were created in this time, Paratopes Ltd. (UK), GBiotech Sárl (CH), and Acquifer AG (D) (Section E.2.4).

The raising of the second EMBL Technology Fund (ETF-II), a venture capital fund managed by EMBL-Ventures, was successfully completed in 2012. Of the EUR 40 million raised, EUR 38.5 million came from external investors and business angels, thereby substantially leveraging the EUR 1.5 million commitment of EMBLEM to ETF-II.

In 2012, the EMBL spin-off company Cellzome was acquired by GlaxoSmithKline (GSK), and the Vienna-based Savira Pharmaceuticals spin-off signed a collaboration agreement with Roche to develop small polymerase inhibitors for the treatment of influenza virus infections.

Box E.2: EMBLEM technology transfer in numbers (2010–2014)

Year	2010	2011	2012	2013	2014	Total
Income (k€)	3,563	4,033	4,482	5,563	5,944	23,585
License & Collaboration Contracts Concluded	141	186	203	329	269	1,128
Invention Disclosures	41	39	47	48	35	210
Priority Patent Applications Filed	11	9	15	21	15	71
Copyrights	8	11	5	2	4	30
Material Transfer Agreements (MTAs)	242	262	374	347	390	1,615
Patents Granted	13	9	13	13	20	68
Patent Families Reassigned	8	9	7	4	4	32
Gross Patent Costs (k€)	391	519	535	366	479	2,290
Patent Cost Recovered (k€)	259	280	307	211	267	1,324
Net Patent Cost(s) (k€)	132	238	228	155	212	965
Start-Ups Created	1	0	3	0	0	4

3.2 Future plans 2017–2021

EMBLEM has built a strong track record of successes and achievements and has followed a commercial strategy that balances short-term income against mid-term financial return to ensure the long-term sustainability of technology transfer activities. Scientific collaboration agreements between EMBL scientists and industry play a central role in this model and additional efforts will be made in the next five-year period to trigger an increasing number of collaborations and to build larger intellectual property portfolios around key areas of EMBL expertise.

Specifically, in the 2017–2021 period EMBLEM expects to receive in excess of 200 invention disclosures and to execute more than 1,000 new licensing, consultancy, and collaboration agreements. Cumulated turnover is expected to increase to above EUR 25 million. The goals of EMBLEM activities in 2017–2021 are: to maintain the high level of technology transfer services to EMBL and the member states; to ensure that promising inventions and technologies are developed to commercial maturity to benefit scientific endeavour and society at large; to attract new commercial partners and further develop the EMBL-EBI Innovation & Translation Initiative; and finally, to broaden the industrial target group of the “EMBL Science Days”.

Box E.3: EMBL spin-off companies

Name	Field	Founding Year	Status	Location
Sygnis (ex Lion Bioscience)	Bioinformatics	1997	Post-IPO	D
Cenix Bioscience	RNAi	1999	Alive	D
Cellzome	Chem. Proteomics	2000	M&A Sale	D
Anadys Inc.	Anti-viral	2000	Post-IPO	USA
Gene Bridges	Genetic Eng.	2000	Alive	D
EVP Inc.	Neuronal Disorders	2001	Alive	USA
SLS	Software	2002	Closed	D
Hybricore	HT mAb Prod	2002	Closed	D/I
Triskel	Oncology	2006	Alive	IRE
Elara Pharma	Oncology	2006	Closed	D
BioBytes	Bioinformatics	2008	Alive	D
Savira Pharma	Influenza	2009	Partnered	A
PEPperPRINT	Peptide Microarrays	2010	Alive	D
Acquifer AG	Microscopy/Screening	2012	Alive	D
GBiotech	Protein Expression	2012	Alive	CH
Paratopes Ltd.	HT Monoclonals	2012	Alive	UK

Status of the spin-out companies (December 2014). 3 out 16 spin-offs (SLS, Hybricore & Elara) are no longer operational, hence survival rate of spin-offs between 1997 and 2014 is 81.25%. Lion Bioscience and Anadys Inc. are post-initial public offering (IPO) and one company exited through Merger and Acquisition (Cellzome was bought by GSK). Savira has been partnered to Roche.

4. Industry Relations

EMBL has a history of collaborating with industry and small businesses in a variety of ways, ranging from strategic institutional programmes to less formal project-based collaborations. These interactions provide industry partners in our member states with access to the expertise of our scientists, our infrastructure and our training events. They also ensure that EMBL's knowledge and technologies are being broadly disseminated and thereby help to translate our

fundamental research into tangible benefits for society. EMBL together with EMBLEM constantly search for new avenues for collaboration.

4.1 Strategic interactions with industry through institutional programmes

4.1.1 EMBL-EBI Industry Programme

The EMBL-EBI data resources, as well as the tools to interrogate them, are used extensively not only by academic institutions but also by industry R&D departments. The pharmaceutical industry has traditionally been the biggest non-academic consumer of the bioinformatics services.

The subscription-based EMBL-EBI Industry Programme is the main way through which Service teams interact with commercial users and industry partners. Members of the Industry Programme were initially all pharmaceutical companies but now also represent the agri-food, nutrition, healthcare and consumer goods sectors. The programme provides quarterly strategy meetings, expert-level workshops on topics prioritised by its members and other forms of communication including webinars and face-to-face meetings. In addition, EMBL-EBI also supports small to medium-sized enterprises (SMEs) through subsidised outreach and training activities either at the new EMBL-EBI South Building or locations within EMBL member or associate member states.

Box E.4: Members of the EMBL-EBI Industry Programme

- Astellas
- AstraZeneca
- Bayer
- Biogen Idec
- Boehringer Ingelheim
- Bristol-Myers Squibb
- GlaxoSmithKline
- Eli Lilly and Company
- F. Hoffmann-La Roche
- Janssen
- Merck Serono
- Nestlé Institute of Health Sciences
- Novartis
- Novo Nordisk
- Pfizer Inc.
- Syngenta
- Sanofi-Aventis
- UCB
- Unilever

Backward look and highlights 2012–2014

Since 2009, six new companies have joined the Industry Programme, among them the first member from Japan, two new European and a number of US-based companies. Forty-seven knowledge exchange workshops on topics including the 1000 Genomes and ENCODE projects, oncogenomics, quantitative systems pharmacology and causal inference in interpretation of '-omics' data took place. In addition, there were 24 quarterly strategy meetings, which allowed EMBL-EBI to present forthcoming new initiatives and receive input from the partners about their needs and priorities.

With financial support from one or more of the companies including AstraZeneca, GlaxoSmithKline, Roche and Pfizer, several small development projects have been undertaken on a pre-competitive basis, the early stages of the development of a commercial product during which competitors who belong to the Industry

Programme collaborate towards achieving a common goal. The results of such projects are freely available.

Future plans 2017–2021

Over the course of the next Indicative Scheme the EMBL-EBI Industry Programme will continue to build upon the successes of the existing model of industry interaction. At the same time, we foresee interactions with industry partners becoming stronger as we work together to address challenges and opportunities created by Big Data, such as the management of large amounts of data, data integration and confidentiality of proprietary, licensed and personal information.

4.1.2 The EMBL ATC Corporate Partnership Programme

The Corporate Partnership Programme (CPP) acts as a facilitating platform for mutually beneficial interactions between EMBL and industry. In close collaboration with the relevant parts of EMBL, the CPP provides corporate partners with opportunities to draw on EMBL's expertise and track record of cooperation with industry for the co-development of training, products and services, as well as in the broader sector of technology transfer.

Since its foundation, funding obtained through the CPP has supported talented young scientists and their participation in and contribution to the world-class scientific gatherings held at the Advanced Training Centre (ATC). Thanks to the generous support of the EMBL Corporate Partners, more than 770 scientists (up until late 2014) have received Corporate Partnership Programme fellowships, making the EMBL ATC one of the key meeting places for young biologists in Europe and beyond. Through the Corporate Partnership Programme, EMBL and its corporate partners also support the development of innovative scientific events at the ATC. At the beginning of 2015, EMBL launched a new type of fellowship supported by the CPP, the Christian Boulin Fellowships, commemorating EMBL’s former Director of Core Facilities and Services who passed away in 2014. The Christian Boulin Fellowships provide support to scientists who wish to visit EMBL to make use of its Core Facilities (Section C.3) but who do not have access to funding to make this possible.

Backward look and highlights 2012–2014

In the years 2012 and 2013, the initial corporate partnership agreements completed their first term. We were happy that almost all were successfully renewed or re-negotiated, reflecting the satisfaction of our partners and strengthening the relationships with the 12 current members of the Programme. Corporate partners are provided with the opportunity to discuss current and future trends of molecular biology with the scientific leadership of EMBL at the annual CPP event. In addition, the CPP has facilitated a variety of meetings with corporate partners interested in exploring collaborations with the Laboratory at different levels, thus connecting corporate scientists and senior management with group leaders, members of the Core Facilities and EMBLEM (Sections C.3 & E.3). In 2014, the Founders and Corporate members of the CPP took part in the newly launched EICAT Consultation Panel meeting, a forum to discuss and broadly explore possible training synergies with industry.

Future plans 2017–2021

Looking forward, the expanding number of interactions with industry facilitated by the CPP is an indication that industry values the CPP as a platform for engagement with the world-class research, services, and training carried out at EMBL. As an attractive entry point for industry relations, the CPP plans to steadily enhance its facilitation services for our industry partners in pre-commercial and pre-competitive areas that are mutually beneficial and meaningful to both scientists at EMBL and industry training as well as R&D departments. In the years 2017-2021 the CPP is also expected to grow membership towards its currently anticipated maximum of 20 companies.

4.1.3 New strategic partnerships with industry

The pharmaceutical and agri-food industries are experiencing a period of major change. There has been increasing pressure to deliver more cost-effective solutions to patients, healthcare providers and other customers as well as to ensure ethical and environmental responsibility. These pressures are driving companies to collaborate in ‘pre-competitive’ business areas and to form strategic alliances so they may access cutting-edge tools and expertise they cannot afford to develop internally.

EMBL plans to capitalise on this development and, in the coming years, aims to forge new, mutually beneficial, strategic partnerships with industry. The idea is to use EMBL’s expertise in various areas, such as data management and analysis or technology development, to complement the needs of biotechnology, pharmaceutical and agri-food companies in pre-competitive research areas. This new type of strategic partnership complements the existing EMBL Industry Programme described above. In this new initiative, EMBL develops larger and more sustained collaborations with commercial partners in order to ensure the translation of our skills and resources into innovative products that benefit society.

EMBL will benefit from these proposed larger-scale collaborations through increased capacity to perform research and develop new tools and products and by enhancing our reputation for cutting-edge technology development. We have already developed one such programme, the Centre for Therapeutic Target Validation (CTTV) (Box E.5) in collaboration with GSK and the Wellcome Trust Sanger Institute. During the next EMBL Programme period, we anticipate development of similar initiatives with other industry partners. The successful launch of the CTTV has stimulated much interest and we are actively responding to enquiries from other firms.

Box E.5: The Centre for Therapeutic Target Validation

The Centre for Therapeutic Target Validation (CTTV) is a pioneering pre-competitive public-private partnership supported by significant funds and in-kind contributions from EMBL-EBI, the Wellcome Trust Sanger Institute and GlaxoSmithKline (GSK). The aim of this initiative is to harness the power of Big Data and genome sequencing to improve the success rate of drug discovery. The CTTV's mission is to use genome-scale experiments and analysis to become a world-leading centre for human biological target validation, to provide evidence for the biological validity of therapeutic targets and an initial assessment of the likely effectiveness of pharmacological intervention directed towards these targets. It aims to address a wide range of human diseases and will share its data openly in the interests of accelerating the development of novel therapies. The CTTV also fosters collaborations between consortium members that extend beyond the human target validation arena and aims to attract further academic and industrial partners in the future.

The CTTV is established within the Innovation & Translation Suite in the EMBL-EBI South building and brings together staff with complementary expertise from each of the three institutions.

4.2 Project-based cooperation with industry

In addition to the above-described institutional, formalised ways of interacting with industry, EMBL also engages in a diverse set of less formal, project-based collaborations.

Where common interests exist, EMBL scientists enter into collaborations with industrial R&D groups. In 2014, 27 scientists collaborated with 43 industry partners on a broad variety of projects related to EMBL's research and service goals. Many of these fall in the areas of method, tool or technology development (Section E.1). Others, particularly those involving Core Facility staff, centre around the testing of new products. The Core Facilities are attractive as sites for beta-testing of new instrumentation. They offer an efficient way of exposing new products to a broad group of high-end users who test them on biological questions and provide feedback and expert advice that companies use to further refine their products.

EMBL's X-ray based structural biology services attract many users from the European pharmaceutical industry. They rely on crystallography to provide structural insight into biological macromolecules and enable structure-aided drug design. They frequently do so in collaboration with EMBL research groups. Examples from the past three years include investigations of anti-influenza drugs, anti-infectives targeting leucyl-tRNA synthetase or the TBK1 kinase (with EMBL Grenoble) or testing and pioneering new concepts in mycobacterial drug discovery and tuberculosis vaccines in collaboration with EMBL Hamburg.

In addition to the Industry Programme discussed above, EMBL-EBI also engages in bilateral projects with industry partners. Over the past decade, these have

ranged from obtaining seed money for specific activities (e.g. the development of the ArrayExpress database), bilateral projects leading to the incorporation of proprietary datasets into public resources like ChEMBL (the EMBL-EBI database of bioactive drug-like small molecules) to projects with a small number of industry partners. All of these smaller-scale projects were financed by industry for a period of time and all of the results of these projects were immediately released to the public. In addition, EMBL-EBI is also involved in publicly funded projects that have strong industry participation, for example the FP7 and H2020 Innovative Medicine Initiatives (IMI 1 and IMI 2) and other EU-Framework Programme projects. During IMI 1, EMBL-EBI participated in ten projects and in the recently started IMI 2, it is already a partner in a number of proposals.

In addition to project-based collaborations, EMBL scientists also make their expertise available through consultancies with companies. In the period 2012–2014, EMBLEM registered around 50 consultancy agreements between EMBL researchers and industry.

F. International Integration and External Relations

EMBL Mission: Playing a leading role in the integration of life science research in Europe

1. International Integration

1.1 Relations with EMBL Member States, Prospect Member States and Associate Member States

EMBL is owned and governed by its members, who provide financial support and decide on its strategic direction. Therefore maintaining close relations and dialogue with the scientific communities and government representatives in member states and associate member states is of key importance to EMBL in pursuing its missions to the benefit of the member states.

EMBL was initially founded by 10 member states (Austria, Denmark, France, Germany, the Netherlands, Israel, Italy, Sweden, Switzerland, and the UK), in the 1980s four more countries joined (Finland, Greece, Norway, Spain), in the 1990s two countries joined (Belgium, Portugal) and in the 2000s four countries: Ireland, Iceland, Croatia, Luxembourg. The global financial crisis that started in 2007/2008 was a major reason for a seven-year period in which no new EMBL member states joined. However, membership in EMBL remained stable, thanks to continuous support within the member states, particularly those that were most seriously affected, and considerable efforts on our behalf to remain responsive to member state needs and provide added value to their national research systems.

In the last few years, EMBL membership expansion to Central and Eastern Europe has been a main strategic focus. The aim is to help develop and integrate Europe's scientific landscape in the life sciences, and to contribute to harnessing the scientific talent and potential of all European states. This is a strategy that is also pursued at the national level by EMBL's member states. To this end a new Prospect Membership policy has been developed and approved by EMBL Council. Prospect Membership is available to countries that are members of the Council of Europe. During the first year of implementation, three countries (Slovakia, Hungary, Poland) have already decided to join the new scheme. We will continue to approach all Central and Eastern European countries in the coming years to continue to broaden EMBL's membership and thereby foster integration of all European countries into the Life Science research community.

Sir John Kendrew, EMBL's first Director General, considered science the most developed international activity in building world peace and international cooperation is embedded in Article II (4) of the founding Agreement of EMBL. For

EMBL, international cooperation, which refers to relations with countries outside of Europe, offers an opportunity to improve the quality of research through competitively striving for excellence, assembling critical mass and by facilitating the flow of ideas and perspectives to enable complementarity and synergy. Although EMBL is a leading research institute in molecular biology in Europe, it cannot pursue all areas of research in the life sciences and therefore must engage with partners that have complementary activities. International cooperation is not an end to itself. Its value is in sharing the benefits of scientific activities among the global scientific community. EMBL's strategic mechanism for promoting international cooperation in Life Sciences is the EMBL Associate Membership Scheme.

Associate membership has been possible since 2003. Australia became the first EMBL Associate member state in 2008. This was viewed by EMBL Council as a test case and evaluated after four years by EMBL's Scientific Advisory Committee (SAC). Following SAC's strategic advice, Council unanimously approved the revision of the EMBL's Associate Membership Scheme to improve and simplify its terms. In November 2013, Argentina joined EMBL as its second Associate Member State under the new terms, and Australia applied for renewal of its membership.

The selection of international partner countries builds on several key principles that were discussed with EMBL Council in Summer 2014: First, and most importantly, the principal of reciprocal responsibility and benefit. Thus, Associate Membership is concluded under the condition that the cooperation will be of mutual benefit. Second, in line with EMBL's missions, associate member candidates should demonstrate strong commitment to advancing life science research nationally as well as on a regional and global scale, including supporting and empowering young scientists. Third, to ensure that pollination of ideas, projects and infrastructure is possible, EMBL considers whether potential future partners have scientific complementarity and thus bring synergy.

Building on a very successful cooperation to operate the ESRF beamline BM14 in Grenoble together with ESRF and India since 2010, a Statement of Intent was signed with India in late 2014 building on long-standing interactions, particularly in structural biology and bioinformatics. A similar agreement is planned with South Africa in 2015. These reflect an interest to join EMBL as an associate member state. A visit to Chile is planned for the second half of 2015 in order to initiate discussion with representatives of the scientific community and the government. Such discussions are the product of approaches to EMBL and are likely to continue throughout the next Programme period.

The EMBL Partnership Programme, which is exclusively available to the Member and Associate Member States, facilitates the dissemination of EMBL's successful operational model and contributes to the build-up of highly skilled critical mass of researchers nationally. It was evaluated by EMBL SAC in 2011, who saw the existing partnerships as very beneficial and recommended not to modify the policy. The number of partnerships has grown steadily, with currently nine in place. The programme represents one of the best options for EMBL member states and associate member states to form closer national links to EMBL at the institutional level.

Figure F.1 EMBL's 21 member states and 2 associate member states (dark petrol) and 4 prospect member states (light petrol).

1.1.1 Member States

Backward look and highlights 2012-2014

In the past three years, EMBL was able to intensify relations with its member states. EMBL's Director General and Director International Relations visited various member states and met with Ministers responsible for science and research or Heads of Research Councils, for example in Greece, Ireland, Iceland, Israel, Portugal, Spain, and Luxembourg. These meetings were very positive as Ministers reconfirmed their support for EMBL and acknowledged the important contribution of EMBL to furthering European life sciences. EMBL also continued working closely with its host countries Germany, the UK, France, and Italy.

In several countries, links between EMBL and national research communities were intensified via formal collaboration agreements. Agreements encouraging scientific collaborations were signed with the Ministry for Higher Education and Research of Luxembourg, Systems Biology Ireland, Karolinska Institutet (Sweden), the University Clinic Hamburg-Eppendorf (UKE) and the National Center for Tumour Diseases (NCT), Heidelberg (both Germany). The cooperation with UKE was strengthened by an agreement to jointly award PhD degrees. EMBL is also a founding member of the Hamburg-based Centre for Structural Systems Biology (CSSB), a joint initiative of nine research partners from Northern Germany, including three universities and six research institutes. In 2014 the Head of the EMBL Hamburg outstation was appointed as the first Scientific Director of the CSSB. The construction of a new research facility on the DESY campus is expected to be completed in 2016. Research activities will focus on infection biology and will aim to unravel the mechanisms of pathogenic processes with the ultimate goal to discover more effective treatment options for bacterial and viral infections.

Another way of linking EMBL and national research communities is the organisation of dedicated scientific workshops designed to attract in particular a large number of young researchers. Workshops took place for example in Greece, Ireland, and in Portugal.

In 2014, the Czech Republic became EMBL's 21st member state. In the years leading to membership, EMBL had already established close collaboration and signed Memoranda of Understanding with two of its life science centers, CEITEC and BIOCEV. Following meetings with representatives from the scientific community and the Minister in 2013, in 2014 EMBC and EMBL Council also endorsed membership of Malta, which is in the process of finalising its national ratification procedures and thereby becoming the 22nd EMBL member state.

In the past two years, relations with Italy were focused on possible relocation of the Monterotondo outstation and revision of the Host Site Agreement between EMBL and Italy. The relocation is no longer necessary because the National Research Council (CNR), the host organisation of the Italian outstation, has decided to invest into the Monterotondo campus and to relocate a significant number of CNR researchers to the site. Renovation and expansion of the EMBL facilities in Monterotondo is now being discussed with CNR and the ministry.

The EMBL-CRG Partnership Unit for Systems Biology was established in 2006 and has been the foundation for establishing the fifth EMBL outstation in Barcelona, Spain (Section B.2.5). A scientific plan for the new Outstation for Tissue Biology and Disease Modeling was developed and evaluated positively by EMBL SAC in May 2012. A Host Site Agreement has been negotiated between EMBL and the Spanish government and a financial plan for the period 2015-2021 drafted. EMBL Council approved the Host Site Agreement in November 2014 and gave in principle approval to establish the Spanish EMBL Outstation in Barcelona.

Future plans 2017-2021

In the upcoming 2017 – 2021 Programme period EMBL remains committed to strengthening relations with its member states. In particular, EMBL will aim to intensify links to those countries which the EMBL leadership was not able to visit during the past programme period. Hosting visitors at EMBL, entering into formalised collaboration with excellent national research institutes, bringing communities together via scientific workshops and participating in important conferences are some of the activities that will be continued as they have proven very successful in the past.

Dialogue with our host countries will remain a priority for EMBL. In particular, in view of Spain hosting a new EMBL outstation as of 2017, EMBL will work closely with the Spanish government in order to ensure its smooth launch and operation. The first Head of Outstation will be recruited once the Host Site Agreement has been signed by the Spanish King and the EMBL Director General in 2015. The outstation is expected to become fully operational in 2017.

With Italy we will continue to negotiate the changes in the Host Site Agreement that are required to establish employment conditions comparable to the other EMBL sites.

A stronger focus will be put on linking to the communities of new EMBL member states, Czech Republic and Malta, but also other countries which may accede to EMBL in the meantime, including the Prospect Member States.

1.1.2 Prospect Member States

Backward look and highlights 2012-2014

In 2012 EMBL Council established a working group tasked with suggesting a new scheme to encourage non-member European countries, in particular from Central and Eastern European, to engage in EMBL. EMBL's aim, as already stated above, is to help develop and integrate Europe's scientific landscape in the life sciences, and to contribute to harnessing the scientific talent and potential of all European states. The working group designed a Prospect Membership Policy, which was approved by the Council in 2013. The policy foresees that European countries, by expressing an intent to accede to EMBL, are given the opportunity to participate in EMBL, its training and obtain access to services under the same conditions as member states, as well as an observer status in the EMBL Council, for a period of three years. The policy has since proven to be of considerable interest to European countries. After visits to several countries and discussion with their scientific communities and political leadership, Slovakia, Hungary, and Poland became prospect member states in the course of 2014 and Lithuania in 2015.

European countries retain the possibility to accede to EMBL without an intermediate prospect membership status and discussions on this possibility have taken place with several countries. In this respect, EMBL established a very fruitful collaboration with the Russian Foundation for Basic Research, through which two rounds of successful joint research projects have been launched to date. Relations also exist with the Kurchatov Institute, to which EMBL donated equipment from the Hamburg outstation when the DORIS beamlines were decommissioned, and which takes part in the project to establish an XFEL-based Biology Infrastructure (XBI) at the European XFEL in Hamburg (Section C.2.3).

Future plans 2017-2021

EMBL will continue to promote its Prospect Member Policy in European countries that do not yet participate in the Laboratory. In addition, EMBL and its prospect member states will engage in joint activities in line with the policy. In 2017, the first prospect memberships will expire (Slovakia, Hungary and Poland) and it is currently expected that these countries will accede to EMBL during 2017-2018. Future prospect member states may therefore also accede to EMBL during the Programme period.

1.1.3 Associate Member States

Backward look and highlights 2012-2014

As agreed between the EMBL Council and the Australian Government, a mid-term review of Australia's associate membership took place in early 2013 to ensure that the cooperation is mutually beneficial and is working well despite the

geographical distance. On the basis of the mid-term review, EMBL SAC delivered a strong recommendation to EMBL Council to continue its engagement with Australia, but also to further boost its interactions with the global life science community via Associate Memberships.

EMBL took the Australian mid-term review as an opportunity to evaluate the 2003 Associate Membership Scheme. In 2013, EMBL Council unanimously approved the proposed revision of the Associate Membership Scheme, as supported by EMBL SAC. The new Scheme aligned the associate membership with full membership more closely by offering access to all EMBL programs, services and training activities at a reduced membership and entry fee, and observer status in EMBL Council. The new Scheme attracted the interest of several non-European countries. Following discussions with Argentina's political leadership and scientific community, Argentina, that has a long history of European collaboration in the biomedical sciences, became the first country to benefit from the new scheme in 2014. In late 2013 Australia also applied for renewal of its membership under the conditions of the new scheme.

EMBL not only expanded its associate membership but also deepened the existing cooperation with its associates by organising exchanges of scientific visitors, joint scientific and industry workshops, conferences and training activities to build stronger bridges between EMBL and the national communities of researchers.

Adding to the new Associate Membership Scheme, in 2014 EMBL introduced a policy on Strategic International Cooperation in Life Sciences Research, which was welcomed by EMBL Council. The policy defines the criteria which EMBL follows when initiating cooperation with countries interested in associate membership. The main requirement is that the cooperation is of significant mutual benefit for both EMBL and the international partner.

Building on this new framework for international cooperation, EMBL initiated discussions on potential associate membership with several non-European counties, including South Africa and India. EMBL and India, represented by the Department of Biotechnology, were already involved in a successful collaboration in the joint operation of the ESRF beamline BM14 at EMBL Grenoble, which was renewed in 2014.

EMBL also formalised two long-standing research alliances – one in the US and one in Japan– by signing a statement of collaborative intent with Stanford University and the National Institutes for Natural Sciences, respectively.

Future plans 2017-2021

In keeping with its missions and with the policies guiding international cooperation, EMBL will continue to engage with associate member candidates that demonstrate strong commitment to advancing life science research nationally and globally. As at present, EMBL will seek to expand its associate membership by integrating the scientific communities of non-European countries with a well-developed national molecular life science program to allow for cross-pollination of ideas, projects, and exchange of researchers and infrastructure access.

EMBL will also continue to work actively with the relevant political actors and the scientific communities of the existing Associate Member States Australia and Argentina to ensure that they make maximal use of the benefits and opportunities provided by their EMBL associate membership.

1.1.4 EMBL Partnerships

EMBL institutional partnerships are close cooperative affiliations between EMBL and external institutions of comparable standard, vision and international orientation. They are working relationships at the institutional level that are based on shared institutional goals and scientific synergy or complementarity. The partnership programme is beneficial for EMBL's member and associate member states because it promotes distribution of the EMBL model by creating a network of centres of excellence in Europe and beyond. Partnerships are established under the principal of mutual benefit and add value to EMBL by exploring topics that are not part of EMBL's core research activities, particularly translational research (Section B.2.3).

Figure F2 EMBL has nine institutional partnerships with institutions in eight countries.

At present, EMBL has nine institutional partnerships that are categorised according to their scientific scope but also proximity to an EMBL site.

Local partnerships involve institutions on or near EMBL campuses, and have emerged from the benefits of sharing infrastructure and equipment. These are:

- Partnership for Structural Biology (Grenoble, France)
- Unit of Virus Host Cell Interactions (Grenoble, France)
- Molecular Medicine Partnership Unit (Heidelberg, Germany)
- Deutsches Elektronen-Synchrotron (Hamburg, Germany)
- Wellcome Trust Sanger Institute (Hinxton, UK)

Remote partnerships were inspired by the desire of Member States to make sure that EMBL's research strategy and successful operational model are exported and implemented on the national level. These are:

- Nordic EMBL Partnership for Molecular Medicine (Helsinki, Finland; Umea, Sweden; Oslo, Norway; Aarhus, Denmark)
- Sars International Centre for Molecular Marine Biology (Bergen, Norway)
- EMBL Australia Partnership Laboratory (Melbourne, Adelaide, Sydney, Perth, Australia)
- EMBL-CRG Partnership Unit for Systems Biology (Barcelona, Spain)

Backward look and highlights 2012-2014

In May 2011 the Partnership Programme was reviewed for the first time by EMBL SAC. The report prepared for the review aimed to provide insight into the progress, structure and strategic value of the Programme to EMBL and the partner institutes. The positive outcome of the review encouraged EMBL to continue exporting its successful operational model and high scientific standards to the member and associate member states through the mechanism of institutional partnerships.

The first three years of the current indicative scheme saw several important developments for the EMBL Partnership Programme as outlined below.

EMBL and the Wellcome Trust Sanger Institute

In 2012 the long-standing collaboration between the EMBL and the Wellcome Trust Sanger Institute was transformed into a full-fledged institutional partnership, facilitating deeper scientific collaboration and cooperation in areas such as services, research, training, technology, public engagement and campus activities.

Nordic EMBL Partnership for Molecular Medicine

In 2013 the Nordic EMBL Partnership for Molecular Medicine celebrated two important milestones: the renewal of the partnership agreement for an extended period of 10 years, and the expansion of the Nordic EMBL network with the official opening of the Danish Research Institute of Translational Neuroscience (DANDRITE) at Aarhus University, which became its Danish node.

The Nordic EMBL Partnership has the unique feature of being geographically distributed across four Member States – Norway, Sweden, Finland and Denmark. The four Nodes have complementary expertise in areas of molecular medicine and work tightly together to create a Nordic research network of excellence in life science.

Partnership in Systems Biology between EMBL and the Center for Genomic Regulation (CRG), Spain

The success of the partnership between EMBL and CRG in Spain ultimately led EMBL Council to support the establishment of the fifth EMBL outstation for Tissue Biology and Disease Modeling based on the existing Partnership Unit (Section B.2.5 outlines the scientific plans). The outstation is planned to become operational in 2017 and will continue to be the focus for collaboration with the CRG.

Unit of Virus Host Cell Interactions, France

In Grenoble, France, the Unit of Virus Host Cell Interactions was established as an Unité Mixte Internationale between EMBL, CNRS and the Université Joseph-Fourier, Grenoble. The agreement expires in 2015 and discussions have been initiated to establish a new cooperation between EMBL Grenoble and the Institut de Biologie Structurale, part of the Université Joseph-Fourier, which has in the meantime moved onto the European Photon and Neutron Science Campus where EMBL Grenoble is also located. As possible legal framework a Féderation de Recherche is being considered.

Partner Laboratory Network in Australia

In Australia, the Partner Laboratory Network grew from an initial two groups to twelve in the course of 2013-2015. The groups are distributed across Australia, but linked through the EMBL partnership model and aim to develop a network of complementary expertise. EMBL has been providing continuous support and know-how for the recruitment and review of scientific faculty in Australia. EMBL, particularly through the EMBL-EBI, supported the establishment of Bioinformatics Resource Australia - BRAEMBL by the provision of strategic guidance. BRAEMBL will provide help and user support covering the bioinformatics service needs of Australian researchers. In addition to enabling optimal exploitation of the bioinformatics tools and data by Australian scientists and contributing to global bio-molecular information infrastructure showcasing Australian science, BRAEMBL will offer extensive user training. The link with EMBL-EBI has been essential for the success of BRAEMBL and will remain a high priority in the future.

Molecular Medicine Partnership Unit and DESY, Germany

Two of EMBL's oldest partnerships – DESY and MMPU – were renewed in 2014 and 2015, respectively, for an extended period of 10 years. Both partnerships have been highly successful and have enabled the exchange of complementary expertise in research and service provision for the benefit not only of the participating institutes, but also of the local research community.

Bridging partnership expertise

In order to leverage the successful partnership model and link its partner institutes working in the field of molecular medicine, EMBL, together with the CRG, organised a joint conference on the topic of 'Perspectives in Translational Medicine' in 2012 in Barcelona. The scope of the event was to offer a forum for bridging the expertise of the different partners and to facilitate the creation of a larger European network. The wide spectrum of topics reflected the diversity of research being pursued by the partnerships, including human genetics, infectious disease, regulatory networks, stem cells and cancer among others.

Future plans 2017-2021

With the expansion of the EMBL membership, EMBL's network of institutional partnerships is likely to grow as interest from new member and associate member states is high. Moreover, many of the existing member states have expressed interest to enhance collaboration with EMBL through the Partnership Programme.

EMBL will continue supporting its existing partnerships by providing scientific and administrative expertise. This will be achieved by participating in review and recruitment panels, by providing know-how for the establishment of core facilities, of training programmes and by advising on the set-up of various science administration aspects in the partnerships.

EMBL will also seek to further integrate all of its existing partnerships by enabling events and initiatives that will bridge the different institutional partners and will bring them together to exchange ideas, establish collaborations and ultimately create a well-functioning network of excellence across Europe and in the associate member states.

The European Commission has introduced a number of new instruments in the research framework programme, Horizon 2020, that aim at widening participation by linking institutions in more developed parts of the EU with institutions in a lower performing country that aims to increase its research capacity. One of these instruments is 'Teaming' and EMBL has successfully obtained EC funding for the first set-up phase of the 'Hungarian Centre of Excellence for Molecular Medicine', a collaboration between three Hungarian universities (Szeged University, Debrecen University and Semmelweis University), the Biological Research centre of the Hungarian Academy of Sciences and EMBL.

1.2 EU Relations

Backward look and highlights 2012-2014

EMBL is Europe's only intergovernmental research organisation in the life sciences and, by following its missions in research, service provision, training and technology development and transfer as well as integration of life science research, makes important contributions to the European Research Area. This has been recognised and the European Commission (EC) and EMBL have established a strong partnership over the past 20 years since their first

agreement on cooperation signed in 1995. Today, the collaboration is based on a Memorandum of Understanding signed in 2011 and implemented through bi-annual Work Plans. The current Work Plan outlines cooperation in areas like research programming, e-infrastructures, mobility of researchers, technology transfer and international cooperation. The EC is also an Observer at EMBL Council and thus is informed about recent developments and upcoming trends at the Laboratory and can contribute to the discussions. The EU funding programmes for research remain the biggest external funding source of the Laboratory and EMBL therefore regularly contributes to EC research policy initiatives and provided advice and input during Horizon 2020 preparation. With regard to the latter, important links were made also to key Members of the European Parliament to convey EMBL's view on the future funding programme during its negotiations amongst the EU institutions.

Future plans 2017-2021

EMBL will continue to maintain a close relationship with the EC in the future. Following the appointment of the new Commissioner for Research, Science and Innovation in late 2014, EMBL will engage in dialogue with him and his services, aiming at advising on new policies. In 2017-2019 the emphasis will be on providing advice and input to the process of preparation of the next EU research funding programme. EMBL will again establish dialogue with key Members of the European Parliament and thereby convey its view on the future of EU research funding.

1.3 European Strategy Forum for Research Infrastructures

The European Strategic Forum on Research Infrastructures (ESFRI) was established by the EU member states more than 10 years ago to provide strategic advice and coordinate large research infrastructure projects in all scientific fields. EMBL as an intergovernmental organisation, has the right to propose its own projects to ESFRI but has never made use of this option. As a long-standing provider of European-scale life science research infrastructure, EMBL however could provide valuable advice and help to the communities who were planning ESFRI projects, and did so in a number of cases. EMBL coordinated two ESFRI preparatory phase projects funded by the EC FP7: ELIXIR and Euro-BioImaging, and participated in a number of other ESFRI preparatory phase projects from the first and second ESFRI roadmaps published in 2006 and 2008 (Instruct, Infrafrontier, BBMRI, EU-Openscreen, EMBRC). During the ELIXIR and Instruct preparatory phase projects, an EMBL Council Working Groups explored how these new research infrastructures could be established and what EMBL's role could be. This included the possibility of setting up new research infrastructures as EMBL special projects, using EMBL as a host. In 2013 the ERIC (European Research Infrastructure Consortium) became available as a new legal instrument for European Research Infrastructures. By that time the ELIXIR Interim Board had decided to use the EMBL special project model. Instruct was established as a UK limited liability company that was created as a spin-off by Oxford University. Infrafrontier was established as a German limited liability company and BBMRI decided to

establish an ERIC. All projects from the 2008 roadmap will be established as ERICs, including Euro-Bioimaging.

The biomedical ESFRI projects will all be established as distributed RIs consisting of a hub and nodes which are not part of the same legal entity but are linked through collaboration agreements. EMBL can participate in ERICs but no financial contributions can be made unless approved by EMBL Council. Since EMBL is not a country, and since none of the ESFRI projects have member states that overlap 100% with the EMBL member states, challenges to our participation remain. In addition, taking part in an ERIC creates a high administrative burden and can only be justified for EMBL if there is a strong scientific and strategic interest in a new RI. This is the case for ELIXIR and Euro-Bioimaging, however EMBL is no longer formally affiliated with any of the other biomedical ESFRI projects.

1.3.1 ELIXIR – European Life-science Infrastructure for Biological Information

ELIXIR is a European research infrastructure that brings together national bioinformatics capacities with those of the European Bioinformatics Institute (EMBL-EBI) to scale up the collective capability to archive, integrate, analyse and exploit the large and heterogeneous datasets produced in modern life science research. ELIXIR ensures that users can easily access data resources that are sustainable, built on strong community standards and safeguarded for the long term.

In 2006 the European Strategy Forum on Research Infrastructures (ESFRI) identified ELIXIR as one of 35 projects with the potential to become a large-scale pan-European infrastructure. In 2007, ELIXIR entered a preparatory phase funded through a competitive grant from the European Commission (EC), coordinated by Janet Thornton, the Director of EMBL-EBI. Since then EMBL, together with the ELIXIR consortium partners, has been driving ELIXIR by developing the strategic concept, securing EC funding and ensuring successful completion of the preparatory and interim phases of the project. ELIXIR's preparatory phase ended in 2012, moving into an interim phase based on a Memorandum of Understanding (MoU). 17 countries and EMBL signed the MoU, which was endorsed by EMBL Council in 2011 as a step towards setting up ELIXIR as an International Consortium. Part of the construction process was the establishment of an interim decision making body - the Interim ELIXIR Board - which assembled representatives of all MoU signatories. The MoU countries agreed on bridging support to fund employment of ELIXIR staff. In May 2013, Niklas Blomberg took up his function as ELIXIR's first Director.

Throughout this process EMBL led the development of ELIXIR's legal and governance structure. After the evaluation of different legal framework options for ELIXIR, which involved an EMBL Council Working Group, an International Consortium Agreement between ELIXIR member countries and EMBL was negotiated, and later approved by EMBL Council. The agreement specifies EMBL as the hosting organisation for ELIXIR and as the provider of so-called Core Resources, which include acting as a legal personality for ELIXIR, employing the ELIXIR staff and making facilities and infrastructure available to the ELIXIR Hub.

As the big data challenge in the life sciences cannot be handled single-handedly by one institution (Sections B.2.1 & C.1), ELIXIR has adopted a distributed hub-and-nodes model, with the Hub co-located with EMBL-EBI on the Wellcome Trust Genome Campus in Hinxton and a growing number of Nodes, , located at centres of excellence in the ELIXIR member states. In addition to hosting the ELIXIR hub, EMBL-EBI also functions as an ELIXIR Node. All the Nodes provide resources and services that are part of ELIXIR. These include: data resources; bio-compute centres; services for the integration of data, software, tools and resources; training; and standards expertise. The integration of these Nodes across Europe is coordinated by the ELIXIR Hub.

The construction of the ELIXIR Hub in Hinxton, was supported by the UK Biotechnology and Biological Research Council (BBSRC) as part of a 75 million GBP (approximately €90 million) award from the UK's Large Facilities Capital Fund in February 2011. In 2009 the BBSRC had already contributed GBP 10 million (approximately €11.5 million) to support a new Data Centre for EMBL-EBI in the context of ELIXIR preparation.

On 18 December 2013 ELIXIR was officially launched in Brussels. By the end of 2014, 11 countries had signed the ELIXIR Consortium Agreement and joined ELIXIR as members. They fund the Hub jointly. In addition Member States locally fund their national Nodes. Additional funding is obtained from the EC framework programme, Horizon 2020, and other sources.

In 2014 was selected as one of Europe's three priority research infrastructures by the European Council and ESFRI, which allows it to apply to dedicated funding in Horizon 2020 to finance its operations.

ELIXIR's first five year Scientific Programme and Financial Plan were approved by the ELIXIR Board in 2014. As ELIXIR is now organised separately from EMBL, with its own Scientific Programme and funding, we will not present detailed future plans for ELIXIR in the context of this EMBL Programme. However EMBL-EBI is a critical part of ELIXIR, providing many of the core data services for Europe and the world. As part of ELIXIR, we will continue to contribute to ELIXIR's role in coordinating bioinformatics service provision in Europe.

1.3.2 Euro-BioImaging

The ESFRI research infrastructure Euro-BioImaging will be an ERIC and provide open user access to a complete range of state-of-the-art imaging technologies in biological, molecular and medical imaging for life scientists in Europe and beyond. Euro-BioImaging will offer image data support and training for infrastructure users and providers, and continuously evaluate and include new imaging technologies to ensure cutting-edge services in a sustainable manner. The infrastructure will consist of a set of complementary, strongly interlinked and geographically distributed Nodes (specialised and expert imaging facilities) to reach European scientists in all member states. The physical user access will take place at these Nodes. The pan-European infrastructure will be empowered by a strong supporting and coordinating entity, the Euro-BioImaging Hub. The Hub will provide the single entry point from which users will be directed to their desired imaging technology as served by the respective Euro-BioImaging Nodes,

and it will coordinate dedicated data management and training activities tailored to the needs of Euro-BioImaging users.

Backward look and highlights 2012-2014

During the Preparatory Phase (2010-2014) that was coordinated by EMBL, Euro-BioImaging conducted extensive consultation with the European scientific communities of imaging providers and users, gained support by 3000 stakeholders, analysed supply and demand of the imaging technologies in Europe, successfully demonstrated technical feasibility of its operational model in a six months proof-of-concept phase, identified and evaluated possible Nodes in European States and published its recommendations for the infrastructure model, governance structure and finance plan in the *Euro-BioImaging Business Plan*. Since March 2014, the Euro-BioImaging Memorandum of Understanding (MoU) has been signed by 13 countries, (Belgium, Czech Republic, Finland, France, Israel, Italy, Norway, Poland, Portugal, Slovakia, Spain, The Netherlands, United Kingdom) and EMBL, which together aim to implement the Euro-BioImaging infrastructure. More countries have indicated that they wish to follow soon including Austria, Hungary, Portugal, Sweden and the German Research Foundation DFG (Germany), which are already participating as Observers. The MoU signatories constitute the Euro-BioImaging Interim Board that governs the Interim Phase prior to operation. Its key tasks now are to prepare the submission of the Euro-BioImaging ERIC statutes to the European Commission (EC), and to decide on the final governance structure, finance plan and user access policy. In parallel, the Interim Board has invited its members to make proposals to host the Euro-BioImaging Hub. Once the ERIC statutes are approved by the EC and signed by the member states, Euro-BioImaging will be established as a pan-European research infrastructure supported and owned by its member states and international organisations and ultimately start its operation.

During the Interim Phase, the EuBI MoU signatory countries commonly contribute to the annual budget of €300.000 for funding the Interim Phase Secretariat, which supports their work towards implementation. In parallel, the EuBI Interim Board has applied for several EC Horizon 2020 grants (total budget applied for by Euro-BioImaging: € 3.6 million) with EMBL as coordinator or major partner in all proposals. On behalf of the Interim Board and together with the other Interim Board members interested in hosting the Euro-BioImaging Hub, EMBL will submit a Preparatory Phase II proposal for the respective H2020 call in Spring 2015.

Since 2010, EMBL administratively and scientifically coordinated the preparatory phase of Euro-BioImaging and supported the organisation and communication among 24 national biological imaging communities across Europe. In addition, it fostered the international cooperation with the Australian and Indian imaging infrastructure communities to prepare the ground for international cooperation and future international user access.

During the Interim Phase, EMBL hosts three of the Interim Phase Secretariat staff, funded by the MoU signatory countries. In addition to this support for the Interim Board, EMBL's Legal Services coordinate the group of legal advisors across several Interim Board member states, and EMBL chairs the Working Group WG4, Funding Acquisition.

Future plans

EMBL has unique expertise and track record in the management of microscopy facilities, the organisation of European and international open user access and support, cutting-edge imaging technology development and the implementation of European-level research infrastructures. We will continue to make this expertise available to Euro-BioImaging by continuing to play a central role in the new infrastructure in the future. EMBL's involvement in Euro-BioImaging will contribute to our missions of technology development and transfer, service provision to our member states, advanced training and European integration. Moreover, Euro-BioImaging further strengthens EMBL's role as a leader in imaging technologies. As a European Intergovernmental Organisation, EMBL is in an ideal position to coordinate European access to biological imaging technologies in the context of Euro-BioImaging.

In 2015, the Interim Board will invitie the MoU signatories to submit their proposal for the Euro-BioImaging Hub, to host the statutory seat for the Euro-BioImaging ERIC and/or the coordination and support activities (user access, training and image data) in biological imaging and/or medical imaging. EMBL has started discussion with the Interim Board member states to develop a joint proposal for hosting the Euro-BioImaging Hub. This can only happen as a joint effort with one or more Euro-BioImaging member states, as only countries can host the statutory seat of the ERIC. If invited by the Interim Board members, EMBL will host the Hub services for coordination of biological imaging (but not for medical imaging). This includes management of open user access to biological imaging technologies at EuBI Nodes; training of facility staff and users of EuBI biological Nodes and provision of common image data services (e.g. access to image analysis tools or image repositories).

1.4 EIROforum

EIROforum is an organisation consisting of eight European intergovernmental scientific research organisations: CERN, EFDA-JET, EMBL, ESA, ESO, ESRF, ILL and XFEL. The mission of EIROforum is to support European science in reaching its full potential through both their individual efforts and by launching common initiatives. In addition EIROforum serves as a forum for its members to share their expertise and exchange best practices in the areas of basic research and the management of large, international infrastructures, facilities and research programmes.

EIROforum remains the only arena where the largest European Intergovernmental Research Infrastructures meet and discuss points of common interest, including European science policy, international relations, human resources, instrumentation, outreach and education and information technology. EMBL, as the only EIROforum partner performing research in the life sciences, is an important partner in the association and actively contributes to all its activities.

Backward look and highlights 2012-2014

EIROforum has been particularly active in the area of building relations with the EC and the EU Council Research Working Party, where it provided input on various EC initiatives, among others implementation of the European Research Area, the draft Charter on Access to Research Infrastructures and its view on

sustainability of research infrastructures. Cooperation between EIROforum member organisations focused on the topics of instrumentation, technology transfer and information technology. EMBL's additional contributions to EIROforum were chairing the Thematic Working Group for International Affairs in the period July 2013-June 2015, hosting the EIROforum website and the editorial office Science in School, the interdisciplinary magazine for science teachers that is published by EIROforum.

Future plans 2017-2021

EMBL will continue to actively contribute to EIROforum activities, its strategy and policies. Its most notable contribution will be the chairmanship of EIROforum in 2017-2018. The priorities for EMBL chairmanship will be prepared in for that period.

2. External Relations

2.1 Alumni Relations

Since its foundation in 1974, EMBL has produced a body of alumni that is fast reaching 7,000 scientists, science communicators and administrators. After leaving EMBL over 80% of them have chosen to work in Europe and remain in science research. 30% are in senior positions.

This distributed network of alumni is one of EMBL's biggest assets and a major benefit for our member states. As a centre of excellence for research in training, EMBL aims to attract talented young scientists from across the world to Europe and provide them with advanced training and ideal conditions to pursue research in molecular biology. Our turnover system, based on time-limited appointments, ensures that the skills and expertise that scientists acquire during their stay at EMBL will afterwards feed into the national research systems of our member states. Helping to train scientists and to create networks and collaborations is a unique service that EMBL offers to its member states and European science.

Our alumni are EMBL ambassadors; they share their experiences of working in a unique international and interdisciplinary environment and often go on to successfully implement aspects of the EMBL model in member state institutions. They also support the Laboratory in various ways beyond their stay at EMBL, for example by raising awareness, sending new recruits or offering positions to EMBL leavers. Beyond the individual level, alumni facilitate institutional partnerships, countries joining EMBL as new (associate/prospect) members, or offer access to their personal networks that are highly valuable for example for our resource development activities. Senior Alumni also play a critical role in mentoring newly arrived alumni, helping them for example to navigate the best way through the national funding system.

EMBL Alumni Relations actively engages our growing body of alumni to the benefit of the institute, of the alumni themselves, and of the (member) states in which they work. For EMBL, Alumni Relations serves numerous purposes including support for current staff and alumni, recruitment of new staff to EMBL, facilitation of international relations, the EMBL Courses and Conferences Programme, and resource development.

Backward look and highlights 2012-2014

Alumni Relations has reached a very exciting stage of development. Staff, alumni and sponsors are now approaching us proactively with suggestions for new events, initiatives, and collaborations. Such alumni-initiated actions typically attract high participation and receive positive feedback from the community. The receptiveness of our alumni community rests firstly on the successful identification and tracking of a large percentage of EMBL's alumni, and secondly on the close and trusted working relationship between EMBL, the EMBL Alumni Association members and its elected board, and thirdly our commitment to engage and listen to current staff and alumni to develop meaningful and mutually beneficial services.

A survey conducted by Alumni Relations enjoyed a response rate of 33% of all alumni, a value that reflects their high level of active engagement. EMBL responded by introducing tools that instantly made alumni visible and contactable online. Also based on the survey results, significant changes were made to EMBL's news content, news channels and publication formats, including more alumni-related news. The EMBL anniversary events that took place in numerous member states in 2014 were planned with the benefit of this survey feedback. More than 1500 staff and alumni joined celebrations in Hinxton, Monterotondo, Heidelberg and Hamburg (events for Grenoble are yet to take place), and more than 100 alumni contributed as organisers, speakers and sponsors.

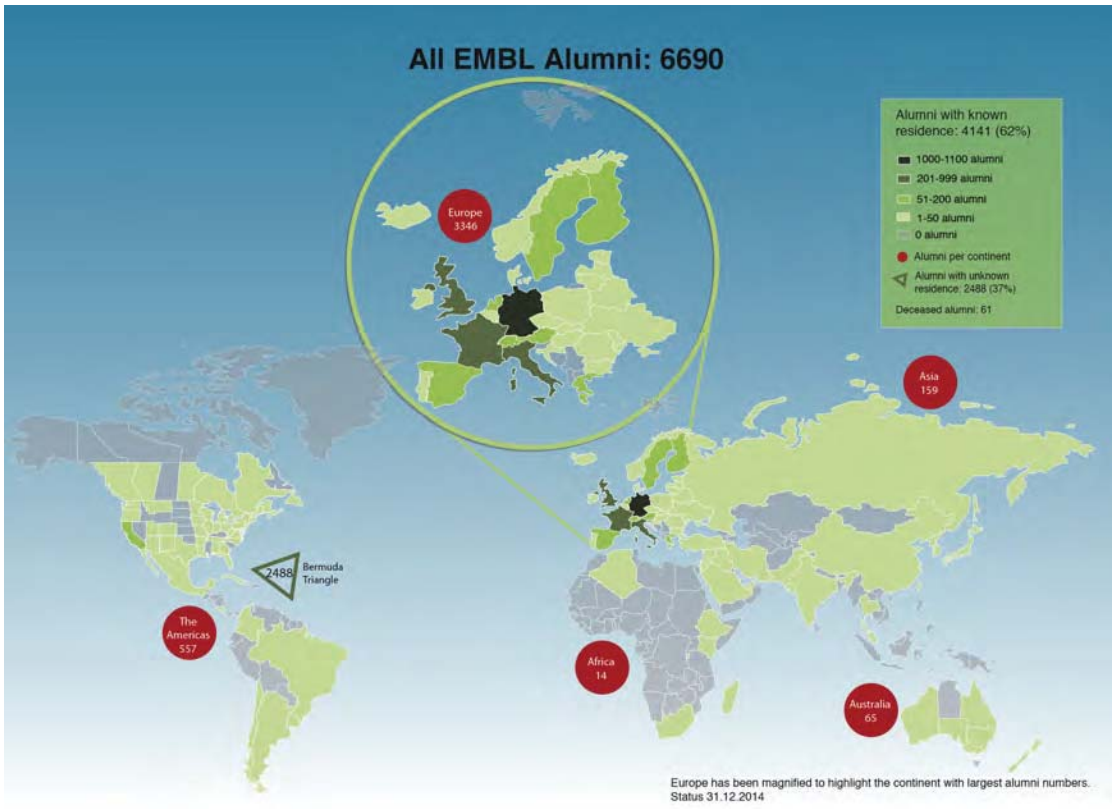
The format of national alumni association meetings, known as 'local chapters', has also evolved based on dialogue with alumni. We now invite non-EMBL scientists from the host countries to these events in an effort to inform them about EMBL and its opportunities. EMBL staff, alumni and their lab members can expand their networks at these national meetings, and explore career options both within alumni labs in the member states and at EMBL. Overall, the number of local chapter meetings in member states has doubled in the last five years, from four to eight. Events following the new format were introduced in Greece in 2012, Portugal in 2013, and Ireland and Belgium in 2014. The positive feedback and the catalytic effect that these meetings have had on EMBL's interactions with scientists from the hosting member states motivates us to intensify these efforts in the future.

Future plans 2017-2021

Alumni Relations will focus on two priorities for the period 2017-21. First, we will design and carry out a survey to capture the value of alumni for EMBL. From regular informal feedback we know that it is significant, but we would like to be able to understand and measure this value in a more detailed and structured way, that could then be communicated in tangible terms back to our supportive alumni community as well as be used for strategic planning purposes.

Second, Alumni Relations together with the Alumni Association and EMBL alumni will build an EMBL Archive (Section F.2.3) that communicates and makes visible EMBL's institutional history, its role in the history of molecular biology and its scientific impact worldwide. We have done much groundwork towards this aim in the last four years and an Archivist has been recruited. Already, the project to build our archive served as a community-building project that meaningfully connects alumni and current staff, and that has received generous support in an internal fundraising campaign conducted together with EMBL Resource Development.

Figure F.3 Overview of EMBL Alumni and their whereabouts.



2.2 Resource Development

As EMBL enters its 5th decade, we are building on its scientific reputation, loyal alumni and positive recognition in the general public and in philanthropic circles by creating a growing group of advocates. With them, we will work towards the identification of private financial support for aspects of EMBL's activity that are not covered by existing member state contributions or third party funding.

Resource Development is still a comparatively new activity at EMBL and our activities are largely strategic and geared towards building a growing group of advocates and sustainable financial supporters for EMBL's programmes and research projects. Resource Development also keeps its eyes open for arising opportunities when they are in line with the overall institutional strategy. The generous support (~15 million €) for the ATC building by the Klaus Tschira Foundation is a prime example of this.

Backward look and highlights 2012-2014

Within the current Indicative Scheme, we focused on putting in place much of the groundwork for the systematic development of private support, including but going beyond contributions that we receive from industry through the Corporate Partnership Programme (CPP), the EMBL-EBI Industry Programme and other more specific initiatives described elsewhere.

EMBL's first strategic Resource Development effort, the CPP, has raised a total of over 1.7 million Euro in its first five years. This income has supported the development of new directions in our internationally acclaimed Courses and

Conferences program at the ATC, and enabled over 700 young scientists (graduate students, postdocs, young PIs) from underfunded laboratories to attend these events. In addition, the CPP has become an important additional vehicle for interactions with industry partners (Section E.4.1.2).

Setting up a fully functional Resource Development operation at EMBL required considerable groundwork. Fundraising guidelines were developed and made available to all staff. Our processes for gift solicitation and stewardship were strengthened, and key messages to develop EMBL's institutional case for support identified. Importantly, Resource Development has begun to develop relationships with several individuals featured on international lists of "the wealthy" who have visited EMBL, shown an interest in what EMBL does, and appear to recognise the special role that it plays.

At EMBL's 40th Anniversary *Festakt* (gala evening) in September 2014 we launched "Friends of EMBL". Designed as an international group of EMBL supporters from the general public, the "Friends of EMBL" offers the possibility to build an active network of influential and sometimes philanthropic people who will advise and support the institute.

Future plans 2017-2021

Resource Development in the years 2017-2021 will continue to pursue targeted donor cultivation activities and will build on the foundations that have been laid. It will also benefit from our growing network of motivated, engaged alumni who offer generous access to their personal networks in support of our Resource Development efforts. Resource Development activities will be promoted at all EMBL sites, and aim to raise support for both local projects and for institutional priorities from within and beyond the member states. Particular emphasis will be placed on developing a solid pipeline in the major donor segment, including corporate and private foundations. Until the end of the next Indicative Scheme, we hope to move significantly toward establishing the principal of an endowment that yields meaningful interest income in support of important aspects of the EMBL mission but that for any reason cannot be covered by the member state contributions. Should appropriate European legislation be passed, this endowment could be established as a European Foundation that offers tax incentives for the majority of European nationals.

2.3 EMBL Archive

The EMBL Archive was initiated to capture the institutional history of EMBL and the stories of its past, current and future staff. Being Europe's flagship laboratory for molecular biology, EMBL's history is strongly correlated with that of European molecular biology as a whole. As such, the purpose of the EMBL Archive is necessarily twofold:

- By safeguarding the institution's memory for future generations, the EMBL Archive will record EMBL's history and accomplishments.
- As a record of molecular biology in Europe, the EMBL Archive will be a valuable resource to communicate the importance of molecular biology first-hand and especially communicate EMBL's role within the life sciences.

2014 marked the 40th anniversary of EMBL. The celebrations with alumni and staff were a timely occasion to look back on the Laboratory's accomplishments, but also brought to the fore the need to capture EMBL's history, 50 years after the inception of EMBO. The Laboratory's entrenched policy of staff turnover and the lack of a facility (physical and digital) in which to deposit items has created a situation where unique material is often discarded as staff become alumni: it is all too easy for drafts of papers, sketches of instruments and photos of social events to be deleted or discarded. The EMBL Archive will become a resource for staff to enable them to deposit their records and materials in perpetuity and for the wider research community to be able to access in due course.

The EMBL Archive will work with staff and alumni from EMBL's five sites to actively gather material (analogue and digital) to constitute the EMBL Archive. Complementary resources as well as collaborations within and beyond EMBL will further enhance the user-experience of the resource, and help meet the objectives stated above.

In the long term, the EMBL Archive will be accessible to future generations of scientists and social scientists as a witness of the contributions of EMBL to molecular biology, collaborative science and the better understanding the natural world.

2.4 Communications, Education and Outreach

As a publicly funded organisation, EMBL has both a responsibility and an interest to communicate its innovative research and technology widely, and to welcome the questions, concerns and responses of the public. It does this through a number of channels, each of which aims to make sophisticated science understandable and convey the importance and fascination of our research to people from many different backgrounds.

EMBL maintains links with journalists, public figures, industry representatives, funders, policymakers, students, teachers, online social communities, the wider scientific community and the general public at large through its Office of Information and Public Affairs (OIPA) and the European Learning Laboratory for the Life Sciences (ELLS) in the main laboratory in Heidelberg, and its External Relations team at EMBL-EBI. The primary responsibilities of EMBL's communications and outreach professionals are to communicate within the organisation and externally using diverse channels, and to empower EMBL scientists and other staff to present EMBL effectively to external audiences.

EMBL communications professionals publicise EMBL's excellence in science, its relevance to society and the opportunities it offers for scientific collaboration, training and career development. They also promote EMBL's crucial position as a provider of life-science infrastructure, and the valuable services it provides for researchers in non-profit and commercial sectors.

EMBL's teams make use of multiple communication channels to reach diverse target audiences, for example press and media engagement, websites and social media, printed publications and promotional materials, slide presentations, exhibitions and presentations for professional conferences, public engagement, visitor programmes, high-profile events and teacher training courses.

EMBL's communication activities evolve according to the needs of stakeholders which include scientists, policy makers and funders, industry, science educators and the general public and in response to the rapidly changing landscape of digital communication.

2.4.1 Communication

Backward look and highlights 2012-2014

Media work

The Office of Information and Public Affairs (OIPA) and EMBL-EBI External Relations coordinate all media relations between EMBL scientists and journalists, communicating important developments in research, technology development, training and other EMBL activities to more than 1,000 media contacts worldwide.

In the period between January 2012 and December 2014, EMBL issued 76 press releases, which resulted in more than 6,000 articles in news outlets throughout the world. Traffic to EMBL's news websites well exceeded a quarter of a million hits in the second half of 2014 alone.

Publications

During the past Programme, EMBL repositioned some of the publications in its diverse portfolio – which includes the EMBL and EMBL-EBI Annual Reports, the *EMBLetc.* magazine, *Research at a Glance*, brochures, and other information materials – and enhanced them to further complement each other. The EMBL Annual Report underwent a significant revision both in terms of content and design, to better satisfy the needs of its target audience, mainly member states' representatives. *EMBLetc.* has evolved into a high-quality website and magazine with a print distribution of 6,500 and targeted accompanying email newsletters, to staff, alumni, close collaborators and interested external parties.

EMBL online

EMBL operates websites and intranets for each of its five sites in Europe. Depending on the profile of the EMBL site, their websites fulfil different purposes and serve different audiences. While the Grenoble, Hamburg, Heidelberg and Monterotondo websites aim to inform the scientific community and the public at large about EMBL's science, training, facilities and services, EMBL-EBI's website is particularly geared towards users seeking access to a vast offering of public data resources.

Driven by the need to provide researchers with a more consistent experience when navigating between data resources, the EMBL-EBI website was redesigned, rebuilt and re-launched in 2013 – a major endeavour with contributions from staff throughout the institute. To celebrate the Laboratory's 40th anniversary in 2014, the EMBL websites, too, received a design makeover and were relaunched concurrently with a successful news website. The modern look

and feel reduces redundancy, is easier to navigate and features EMBL's research and scientific stories more prominently, including on mobile devices.

EMBL and EMBL-EBI launched and quickly expanded their social media presence, for example reaching 10,000 followers on EMBL-EBI's Twitter in 2014. This has proved a successful way to promote news, events, jobs, awards, training opportunities and other information to large audiences on Twitter, Facebook, LinkedIn, Google+ and YouTube. Videos posted on the EMBL Media YouTube channel have been particularly successful, with the most popular attracting between 5,000 and 25,000 views. EMBL's scientific images are often featured in newspapers, news websites and popular science magazines. The high level of engagement within these networks – likes, retweets, shares and comments – reveals the considerable impact of social media in amplifying our messages and sparking online conversations about EMBL's activities and services.

Future plans: 2017-2021

Unified, updated EMBL brand

As a distributed organisation with a broad set of activities, EMBL needs a strong, uniform brand that makes it recognisable as one organisation delivering excellent science, services and training. A unified brand will optimise visibility of EMBL services and the many opportunities on offer for the scientific communities in our member states. The communications and outreach teams at EMBL will work together to establish a clear visual identity and brand for the organisation, focusing on consistent visual presentation and tone of voice. This body of work involves three main stages and is expected to take two years from initiation to launch. The first stage is consultancy with internal and external stakeholders to benchmark and establish how the organisation is actually perceived. The second involves hands-on workshops with individuals throughout the organisation to establish positioning statements and the necessary tools to implement a refreshed brand. The final stage is implementation.

New website and digital content strategy

Over the next five years, EMBL communications will develop and implement a digital content strategy that streamlines working practices and empowers more people throughout the organisation to create and distribute high-quality, discoverable content on multiple channels. A user-led approach to content generation and distribution will allow all teams to fully exploit materials – written, graphical and audiovisual – created for the press, funders, policymakers, alumni and educators. This will be realised mainly by replacing the current web content management system (CMS) with a more user-friendly and flexible one that integrates seamlessly with other systems and tools, such as existing content databases, EMBL's research information system and a new media asset management system. The latter will bridge the gap between print and digital. To this aim, the EMBL websites' design will continue to evolve based on users' needs, also bringing the EMBL and EMBL-EBI websites closer together. Finally, a coordinated digital content strategy will allow EMBL communications teams to use consistent metrics to monitor the success of each project, and to adjust their approaches based on these data.

2.4.2 Education

The European Learning Laboratory for the Life Sciences (ELLS) has a long-standing tradition of linking secondary school science teachers, students and EMBL scientists through its multifaceted training activities. ELLS has been launched by EMBL in 2003 to address the demand for continuing professional development of European secondary school science teachers. In addition, ELLS has been driving a variety of projects to bring pupils into the research lab and to co-develop teaching resources with teachers and students.
Over the last years ELLS has continuously expanded its portfolio and now offers a variety of formats ranging from face-to-face training activities to the development of e-learning resources, the provision of interactive online seminars and outreach activities.

Backward look and highlights 2012-2014

ELLS LearningLABs

Engaging teachers with contemporary science and bringing them in contact with EMBL researchers is a hallmark of ELLS' actions. In ELLS LearningLABs, high school science teachers develop hands-on expertise and refresh their knowledge. These multi-day workshops for international groups of teachers bring the participants in contact with the institute's vibrant scientific environment through a blend of practical experiments, hands-on modules, presentations by EMBL scientists, and visits to world-class research facilities. In addition to wet-lab courses ELLS also offers bioinformatics training courses and is developing bioinformatics teaching resources to be used in the classroom.

In order to further disseminate the LearningLAB model, ELLS co-organises pilot courses at research institutions in EMBL member states who are looking to implement similar programmes for teachers.

ELLS Webinars, ELLS Science Chats and EMBL Insight Lectures

In addition to face-to-face training courses, ELLS offers online training, including engaging online seminars by EMBL scientists, called ELLS Webinars, and an interactive online ELLS Science Chat for students. In EMBL Insight Lectures EMBL scientists inform young people about current trends in life science research and spark discussions about how research influences our lives.

EMBL School Ambassadors

In 2013, ELLS launched the EMBL School Ambassadors programme. The EMBL School Ambassadors are EMBL scientists who have chosen careers in the interdisciplinary life sciences. They visit schools in Europe and share their experiences of working as a scientist.

EMBLog – the ELLS web portal

All of ELLS activities are underpinned by the new ELLS portal EMBLog, which ensures lasting contacts to ELLS target audiences. It is an exchange platform for teachers, supports the dissemination of ELLS resources in multiple European languages, represents a central hub for ELLS projects and is instrumental in maintaining the growing ELLS network.

Future plans: 2017-2021

In the EMBL Programme 2017-2021 ELLS will focus on:

- Providing a blend of innovative face-to-face and online training formats based on modern science education methodologies;
- Disseminating strategies to provide EMBL's member states with educational resources and training activities;
- Integrating additional education and outreach activities into the existing portfolio to optimally serve various target audiences and to further develop scientific citizenship;
- Increasing involvement of young EMBL scientists via the EMBL School Ambassadors Programme to actively participate in outreach activities and inform the next generation of scientists about scientific career options.

2.4.3 Outreach

In its over 40-year history, EMBL has gained widespread recognition in the scientific community. As EMBL enters into its 5th decade, we aim to build on its scientific reputation to foster public engagement and support. EMBL engages with the public and local communities in its host countries in diverse ways. Raising awareness of contemporary life sciences research, its impact on society and the opportunities it provides will be an essential factor for the future work of EMBL. It will help to sustain the integration of EMBL's research outcomes into societal contexts but also to transfer the strong brand image that EMBL enjoys within the global scientific community to the general public.

EMBL's outreach activities are handled in a close collaboration between OIPA, ELLS, Alumni Relations, Office of Resource Development and EMBL-EBI External relations teams.

Backward look and highlights 2012-2014

EMBL's 40th anniversary in 2014 offered an excellent opportunity to reinforce EMBL's outreach activities and to build lasting relationships with local communities. A pilot campaign to raise public awareness and engagement, focusing initially on Germany as host country, was launched in 2014. It was centered on a specially developed core theme that communicates EMBL's mission to the general public: "Vom Leben lernen" ('Learning from life').

The campaign included a web area (the "Discovery pages", www.embl.de/leben) as the hub for all activities, a short cinema ad shown regionally, as well as a variety of public outreach events. In regular scientific Sunday matinees PhD students gave presentations about their research work to large public audiences. 'Creativity' and 'the quest to understand life' as common denominators between the arts and science were highlighted in several events, including the photo exhibit "DNA | Portraits by Horst Hamann", or a concert hosted at EMBL's Advanced Training Centre within the Heidelberg festival 'EnjoyJazz'. The 'Forschercamp' (Research Camp) offered an opportunity for members of the general public to gain hands-on experience in laboratory experiments over a

weekend. The radio announcement of the event alone reached almost one million listeners every morning for a week.

All campaign elements generated significant media coverage. Strategically, we combined the above activities for larger audiences with more selective, individual engagement events (Section F.2.2).

Apart from the 2014 anniversary campaign, EMBL regularly hosts visits from secondary-school and university students (among many others). EMBL Heidelberg saw almost one such public visit per week 2012-2014. In addition, EMBL takes part in local public science outreach initiatives such as the Girls' Day initiative, the 'Long Night of Science' festivals and the Science Days in Germany, the Cambridge Science Festival in the UK and *Cafés Scientifiques* in France.

EMBL also exhibits regularly at high-profile conferences to promote careers at the Laboratory and to raise awareness of our services and training opportunities. For example, EMBL has a strong presence at the EuroScience Open Forum, Europe's largest general science meeting.

Future plans 2017-2021

During the 40th anniversary campaign we have increased our outreach activities and EMBL has in general become more visible at its sites. In the future we plan to use the core structures established through the campaign to raise awareness, communicate in simple language what EMBL does and why it is important for society, and engage public support. We will continue to build on formats that have proven successful. For example, innovative educational resources developed by ELLS will be used to engage with visitors and to highlight EMBL research topics and scientific achievements. In addition, we will also develop new ways to engage new target audiences. We plan for example, to develop a set of devices to be used in temporary and permanent exhibitions at specific EMBL locations in order to meet the increasing number of requests from visitors.

2.5 Science and Society

The last quarter of the twentieth century saw an important change in the public interest in and perception of molecular biology. With the advent of new biotechnologies and their application to food production, pharmaceuticals, and biomedicine, common knowledge of molecular biology – ranging from plain supermarket wisdom to sophisticated interdisciplinary expertise – spread far beyond academic boundaries. In this process, the socio-economic, regulatory and moral significance of biology also became increasingly apparent.

As a flagship scientific organisation and training institution in Europe, EMBL recognises its obligation to expose its scientists to the evolving social concerns and ethical debates relating to applications growing out of the life sciences. EMBL also aims to promote a better understanding of science and its societal implications by stimulating an active dialogue between its scientists and members of the public. To this end, EMBL launched a Science and Society Programme in 1998 and has since been at the forefront of developing means to actively incorporate societal issues within its institutional frameworks.

Backward look and highlights 2012–2014

A variety of Science and Society activities and events, organised across all EMBL sites, bring together members of the life science community, scholars of other disciplines, as well as members of the public for discussion and communication extending beyond professional boundaries[1]. Some of the activities are targeted primarily to the EMBL research community (e.g. the EMBL Forum seminar series[2]), whereas others are directed towards a broader audience. The EMBL Science and Society Programme has initiated successful collaborations with EMBO and other leading life science and cultural institutes in the area. Together, they organise thematic symposia and yearly interdisciplinary conferences, most notably the annual EMBL/EMBO Science and Society Conference that regularly attracts several hundred participants[3]. EMBL, the German Cancer Reseach Centre (DKFZ), and the University of Heidelberg also jointly launched a distinguished lecture series, 'Heidelberg Forum – Biosciences and Society', aimed at informing and engaging local audiences[4]. Together with participating local EMBL staff members, the Science and Society Programme manager has regularly (co-) organised events at all EMBL sites.

Across all of these different formats, the Science and Society Programme has organised around 50 events in the first three years of the current Indicative Scheme (2012-2014). The discussions have covered topics as diverse as the implications of personal genomes, biodiversity, the impact of food on body and mind, play and creativity and astrobiology, to name only a few examples.

Future plans 2017–2021

For the EMBL Programme 2017-2021 we anticipate that we will continue to make use of, and further develop, our various channels and means of Science and Society communication (conferences, symposia, lectures series, on-line videos, informal discussion meetings, etc.). For future topics, we intend to focus on the societal and ethical aspects linked to the research plans outlined in this document: personal genomes and personalised medicine, data access and data privacy, evolution, research conduct and bioethics among others. At the same time, the EMBL Science and Society Programme's mission goes beyond issues directly relevant to EMBL's research and aims to stimulate its scientists and the public with insights into new trends in many disciplines. The life sciences have enormous potential for further development and practical application. However, a popular consensus needs to be developed as to how to assess and deal with the diverse repercussions of such developments. More than ever, in the years ahead there will be a need for interdisciplinary dialogue to inspire synthetic insights and a common worldview. The new ways in which science is now being applied for the production of knowledge and economic wealth must be carefully adjusted to public interests and the value systems across Europe. It is the common responsibility of all – scientists as well as non-scientists – to engage in an ongoing process of carving out a shared understanding of science. The EMBL Science and Society Programme will continue to work towards that important goal.

[1] http://www.embl.de/aboutus/science_society/index.html
[2] http://www.embl.de/aboutus/science_society/forum/index.html
[3] http://www.embl.org/aboutus/sciencesociety/conferences.html
[4] http://www.embl.de/aboutus/science_society/hd_forum/index.html

G. Administration

The EMBL Administration supports EMBL's missions by handling administrative aspects so that EMBL staff can focus their activities in the areas of research, technology development, service provision, training or international integration. EMBL operates across five sites in four countries and has more than 1,800 staff from over 80 nations. Considering these facts, EMBL works with a comparatively lean Administration. Across the Laboratory as a whole there are currently 104 people employed in the traditional administrative activities (Finance, Purchase, Human Resources, and Legal Services), which is around 6% of EMBL's total staff. An additional 148 staff improve the working environment for scientists by providing them with support services for their laboratories and social services to them and their families. This category includes, for example, facility management, health and safety, security, the canteen and cafeteria, the kindergarten and gardening in Heidelberg.

Most of EMBL's administrative tasks are handled centrally at the headquarters in Heidelberg. In addition, the outstations employ a small number of local administrative staff with the exception of EMBL-EBI, the size of which now requires substantial local administrative support. EMBL's Administration adopts a unified approach across all of its sites, and activities in the outstations largely mirror the activities in Heidelberg. However, the local administrations also engage in site-specific issues. These include, for example, tasks centred around the new EBI South Building, maintenance of the new data centre and hosting of the ELIXIR hub at EMBL-EBI, support for local collaborations with DESY, the European XFEL and the CSSB in Hamburg as well as ESRF, ILL and national organisations in Grenoble and challenges around the accommodation of servers in Monterotondo. Most of these tasks are long-term issues, which will continue to occupy outstation administrators (as well as others in Heidelberg) well into the new Indicative Scheme.

The delivery of an efficient administrative service is an increasingly multidisciplinary enterprise with administrative staff often cooperating to form *ad hoc* teams, usually with the participation of scientists, to address difficulties or improve procedures. This is partly a pragmatic result of the requirement for electronic systems that interface smoothly with each other, but it also reflects a need for administrative systems to be a closer fit with the EMBL culture of cooperation and flexibility. A common thread running through these teams is the need for legal advice, whether on matters of employment law, international agreements or contracts and a small legal service has been formed to meet this need centrally.

Below we describe major developments that have taken place during the first three years of the current Indicative Scheme, and the future plans and projects of the EMBL Administration in more detail.

1. People

During the current Indicative Scheme, EMBL's progress in implementing the European Charter for Researchers and the Code of Conduct for Recruitment of Researchers was recognised by the granting of the HR Excellence in Research Award by the European Commission.

Working conditions for EMBL fellows have been improved by including them in the EMBL pension scheme as well as entitling them to allowances and benefits formerly restricted to staff members.

The Council Working Group on terms and conditions of employment completed its five-yearly review in 2013, based on a survey of 14 national and international organisations, and concluded that EMBL's terms and conditions are broadly appropriate to attract and retain highly qualified staff. Nevertheless, it suggested some small adjustments to align some terms with those in other organisations and to enhance EMBL's competitiveness. At the same time, EMBL carried out a review of the Health Insurance Scheme to halt a rapid increase in the scheme's funds. This resulted in a readjustment of some benefits and a reduction in the rates of employer and employee contributions. Council currently has a working group looking into the EMBL pension scheme with a view to ensuring its sustainability and fitness to meet future needs.

A joint team made up of staff from Human Resources and other areas, including scientific areas, has embarked on a major review of the Staff Rules and Regulations to provide a necessary update of the regulatory framework governing the terms and conditions of employment. This exercise will continue well into the next Indicative Scheme.

Over the next few years, Human Resources aims to streamline its standard processes. Work to improve and automate time and leave recording systems has already started. Processes for contract management and systems relating to the appointment and departure of staff will follow next. In the area of recruitment, future focus areas include enhancing EMBL's website and monitoring gender balance issues. In particular, women are under-represented in senior research positions at EMBL and so an internal working group on gender balance was established in 2014 with the task of identifying gaps and developing recommendations as to how to address this. Several relevant changes in researcher recruitment practice have recently been implemented.

Changes in health and safety requirements across Europe will continue to require significant effort to ensure that our practices and procedures are at least as good as national standards and we will be undertaking additional work to develop our procedures up to and beyond required standards.

2. Systems and Processes

Over the first three years of the current Indicative Scheme, EMBL's Administration dedicated significant effort to the optimisation, streamlining and modernisation of its administrative systems. This initiative began with the implementation of a financial budgeting and reporting system during the previous Indicative Scheme. Expanding this effort to encompass other administrative processes, existing systems have been upgraded to improve the availability and accuracy of non-financial data as well as the user-friendliness of their interfaces. The implementation of new software solutions, in particular ones that are inter-linked around SAP's business warehouse system and compatible specialised software, has not only significantly improved EMBL's ability to monitor and report on its financial and staffing situation but also allows the integration of data from other databases, such as Human Resources, grants, publications, visitors and Finance. In addition, EMBL has implemented and is engaged in customising a Research Information System, which enhances the tracking of our publication output and other indicators. Taken together, all these system upgrades have greatly improved

EMBL's ability to quickly and reliably produce meaningful and informative reports for its different stakeholders.

EMBL's Administration has been working to reduce paper-based processes, improve their efficiency and speed, widen the online availability of information and enable staff to work remotely whenever and from wherever their working practices dictate. Projects that have been completed in the first three years of the current Indicative Scheme include the transition to electronic sales invoice processing, timesheets, stores management and grants administration. In the immediate future, we expect to complete similar projects touching on absence administration and electronic capture and processing of purchase invoices, with electronic purchasing catalogues and stores following before the end of the current Indicative Scheme. The implementation of all these changes will continue to require significant training and an internal communication effort to help staff adapt to the new systems and processes. It also involves the purchase, customisation and integration of multiple software packages.

For the period of the next Indicative Scheme, the EMBL Administration plans to further streamline and upgrade its processes and systems. In particular, we aim to make access to administrative systems available via portable devices, further integrate stand-alone systems, streamline the production and reporting of statistics and develop interactive systems for modelling and scenario planning.

Developments in ethics regulation covering conflicts of interest and the use of human and animal material have increased the amount of regulation and recording required of EMBL by various funding bodies. To comply with these new requirements, EMBL will need to support current paper-based procedures with automated systems that can cope with issues of confidentiality and risk management.

3. Buildings

EMBL is a highly dynamic organisation as a result of both its staff turnover system and the rapid advances being made in the life sciences. Our building requirements often alter with the departure and recruitment of key scientific staff, changes in research directions and advances in technology. Facility management works to adapt our buildings and facilities to continually satisfy the needs of EMBL's scientists.

EMBL Administration moved to the Advanced Training Centre (ATC) on completion of the new building in 2010. The free space created by this move allowed a comprehensive rationalisation and a still-ongoing modernisation of laboratory space in Heidelberg to bring the accommodation of all Research Units up to current standards. A second major refurbishment project in Heidelberg is the replacement of the façade of the main building, which has deteriorated over time. This work started in 2014 and will continue until the end of the current Indicative Scheme. Care is being taken to minimise the disruption to the scientific work of the Laboratory. Apart from these major projects, a continuous programme of maintenance and improvement is being carried out to increase energy efficiency and to ensure that a build-up of delayed work does not cause financial problems in future.

Major building works at the other EMBL sites in the first three years of the current Indicative Scheme include the repair of buildings and the improvement of safety at Monterotondo and provision of extensive and continuous expert input into the CSSB and European XFEL building projects in Hamburg to ensure that the needs of EMBL life

scientists were taken into account when facilities and laboratories were designed and built. Administration at EMBL-EBI in Hinxton played a central role in managing the construction of the EMBL-EBI South Building from 2009–2013, which was funded by a grant from the UK Research Councils to EMBL and was completed on time and within budget.

Foreseeable future projects in the area of facility management for the next Indicative Scheme include, most pressingly, a need to continue the renovations in Monterotondo. We are currently working with our Italian host organisation, the CNR, to plan and prioritise this work. In preparation for future developments on the Heidelberg campus, EMBL Administration is already liaising with local authorities in Heidelberg and work has begun on expanding the capacity of the sewer system. One potential, but not yet definite, space requirement in the next period would be to enable EMBL to play an active role in hosting part of Euro-BioImaging (Section F.1.3.2) and house other equipment for imaging initiatives.

Appendix 1

Research highlights from the external scientific community enabled by EMBL Structural Biology Services in 2012-2014

A. EMBL Hamburg

Monecke T. et al. (2012) Structural basis for cooperativity of CRM1 export complex formation. Proc Natl Acad Sci U S A 110:960-965. doi: 10.1073/pnas.1215214110

Santos K.F. et al. (2012) Structural basis for functional cooperation between tandem helicase cassettes in Brr2-mediated remodeling of the spliceosome. Proc Natl Acad Sci U S A 109:17418-23. doi: 10.1073/pnas.1208098109

Tidow H. et al. (2012) A bimodular mechanism of calcium control in eukaryotes. Nature 491:468-472. doi: 10.1038/nature11539

Elegheert J. et al. (2012) Allosteric competitive inactivation of hematopoietic CSF-1 signaling by the viral decoy receptor BARF1. Nat Struct Mol Biol 19:938-947. doi: 10.1038/nsmb.2367

von Castelmur E. et al. (2012) Identification of an N-terminal inhibitory extension as the primary mechanosensory regulator of twitchin kinase. Proc Natl Acad Sci U S A 109:13608-13613. doi: 10.1073/pnas.1200697109

Mastny M. et al. (2013) CtpB assembles a gated protease tunnel regulating cell-cell signaling during spore formation in Bacillus subtilis. Cell 155:647-658 doi: 10.1016/j.cell.2013.09.050

Fernandez-Tornero C. et al. (2013) Crystal structure of the 14-subunit RNA polymerase I Nature 502:644-649 doi: 10.1038/nature12636

Juergens M.C. et al. (2013) The hepatitis B virus preS1 domain hijacks host trafficking proteins by motif mimicry. Nat Chem Biol 9:540-547. doi: 10.1038/nchembio.1294.

Civril F. et al. (2013) Structural mechanism of cytosolic DNA sensing by cGAS. Nature 498:332-337 doi:10.1038/nature12305

Paulus J.K. et al. (2013) Greater efficiency of photosynthetic carbon fixation due to single amino-acid substitution. Nat Commun 4:1518 doi: 10.1038/ncomms2504

Malvezzi F. et al. (2013) A structural basis for kinetochore recruitment of the Ndc80 complex via two distinct centromere receptors. EMBO J 32:409-423. doi: 10.1038/emboj.2012.356

Castro-Roa D. et al. (2013) The Fic protein Doc uses an inverted substrate to phosphorylate and inactivate EF-Tu. Nat Chem Biol 9:811-817. doi: 10.1038/nchembio.1364

Motz C. et al. (2013) Paramyxovirus V Proteins Disrupt the Fold of the RNA Sensor MDA5 to Inhibit Antiviral Signaling. Science 339:690-693. doi: 10.1126/science.1230949

Leidig C. et al. (2013) Structural characterization of a eukaryotic chaperone--the ribosome-associated complex. Nat Struct Mol Biol 20:23-28. doi: 10.1038/nsmb.2447

Ribeiro E.D. et al. (2014) The Structure and Regulation of Human Muscle α-actinin. Cell 159:1447-1460. doi: 10.1016/j.cell.2014.10.056

Tamulaitis G. et al. (2014) Programmable RNA Shredding by the Type III-A CRISPR-Cas System of Streptococcus thermophilus. Mol Cell. 56:506-517. doi: 10.1016/j.molcel.2014.09.027

Lamontanara A.J. et al. (2014) The SH2 domain of Abl kinases regulates kinase autophosphorylation by controlling activation loop accessibility. Nat Commun 5:5470. doi: 10.1038/ncomms6470.

Finci L.I. et al. (2014) The crystal structure of netrin-1 in complex with DCC reveals the bifunctionality of netrin-1 as a guidance cue. Neuron 83:839-849. doi: 10.1016/j.neuron.2014.07.010

Soykan T. et al. (2014) A conformational switch in collybistin determines the differentiation of inhibitory postsynapses. EMBO J 33:2113-2133. doi: 10.15252/embj.201488143

De D. et al. (2014) Inhibition of master transcription factors in pluripotent cells induces early stage differentiation. Proc Natl Acad Sci U S A. 2014 111:1778-1783. doi: 10.1073/pnas.1323386111

Chaves-Sanjuan A. et al. (2014) Structural basis of the regulatory mechanism of the plant CIPK family of protein kinases controlling ion homeostasis and abiotic stress. Proc Natl Acad Sci U S A111:E4532-4541 doi: 10.1073/PNAS.1407610111

Kuhle B. and Ficner R (2014) A monovalent cation acts as structural and catalytic cofactor in translational GTPases. EMBO J 33:2547-2563. doi: 10.15252/EMBJ.201488517

Finci L.I. et al. (2014) The Crystal Structure of Netrin-1 in Complex with DCC Reveals the Bifunctionality of Netrin-1 As a Guidance Cue. Neuron 83:839-849 doi: 10.1016/J.NEURON.2014.07.010

Kuhle B. and Ficner R. (2014) eIF5B employs a novel domain release mechanism to catalyze ribosomal subunit joining. EMBO J. 33:1177-1191 doi: 10.1002/embj.201387344

Verstraete K. et al. (2014) Structural basis of the proinflammatory signaling complex mediated by TSLP Nat Struct Mol Biol 21:375-382. doi: 10.1038/nsmb.2794

B. EMBL Grenoble/ESRF

Gayathri P. et al. (2012) A bipolar spindle of antiparallel ParM filaments drives bacterial plasmid segregation. Science 338:1334-1337. doi: 10.1126/science.1229091

Perez C. et al. (2012) Alternating-access mechanism in confromationally asymmetric trimmers of the betaine transporter BetP. Nature 490:126-30. doi: 10.1038/nature11403

Alon A. et al. (2012) The dynamic disulphide relay of quiescin sulphydryl oxidase. Nature 488:414-418. doi: 10.1038/nature11267

Kellosalo J. et al. (2012) The structure and catalytic cycle of a sodium-pumping pyrophosphatase. Science 337:473-476. doi: 10.1126/science.1222505

Neubauer C. et al. (2012) Decoding in the absence of a codon by tmRNA and SmpB in the ribosome. Science 335:1366-1369. doi: 10.1126/science.1217039

Lee C. et al. (2013) A two-domain elevator mechanism for sodium/proton antiport. Nature 501, 573-577. doi: 10.1038/nature12484

Maeda K. et al. (2013) Interactome map uncovers phosphatidylserine transport by oxysterol-binding proteins. Nature 501:257-261. doi: 10.1038/nature12430

Hondele M. et al. (2013) Structural basis of shitone H2A-H2B recognition by the essential chaperone FACT. Nature 499:111-114. doi: 10.1038/nature12242

Mevissen T.E. et al. (2013) OUT deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. Cell 154:169-184. doi: 10.1016/j.cell.2013.05.046

Bell C.H.et al. (2013) Structure of the repulsive guidance molecule (RGM)-Neogenin signaling hub. Science 341:77-80. doi: 10.1126/science.1232322

Kosinska Eriksson U. et al. (2013) Subangstrom resolution X-ray structure details aquaporin-water interactions. Science 340:1347-1349. doi: 10.1126/science.123430

Kim Y.J. et al. (2013) Crystal structure of pre-activated arrestin p44. Nature 497:142-146. doi: 10.1038/nature12133

Vieira-Pires R.S. et al. (2013) The structure of the KtrAB potassium transporter. Nature 496:323-328. doi: 10.1038/nature12055

Baradaran et al. (2013) Crystal structure of the entire respiratory complex I. Nature 494:443-448. doi: 10.1038/nature11871

Zanier K. et al. (2013) Structural basis for hijacking of cellular LxxLL motifs by papillomavirus E6 oncoproteins. Science 339:694-698. doi: 10.1126/science.1229934

Nardini et al. (2013) Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. Cell 152:132-143. doi: 10.1016/j.cell.2012.11.047

Galej W.P. et al. (2013) Crystal structure of Prp8 reveals active site cavity of the spliceosome. Nature 493:638-643. doi: 10.1038/nature11843

Wright K.E. et al. (2014) Structure of malaria invasion protein RH5 with erythrocyte and blocking antibodies. Nature 515:427-430. doi: 10.1038/nature13715

Hassaine G. et al. (2014) X-ray structure of the mouse serotonin 5-HT_3 receptor. Nature 512:276-281. doi: 10.1038/nature13552

Takala H. et al. (2014) Signal amplification and transduction in phytochrome photosensors. Nature 509:245-248. doi: 10.1038/nature13310

Pérez-Vargas J. et al. (2014) Structural basis of eukaryotic cell-cell fusion. Cell 157:407-419. doi: 10.1016/j.cell.2014.02.020

Burke J.E. et al. (2014) Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. Science 344:1035-1038. doi: 10.1126/science.1253397

Grotwinkel J.T. et al. (2014) SRP RNA remodeling by SRP68 explains its role in protein translocation. Science 344, 101-104. doi: 10.1126/science.1249094

Sayou C. et al. (2014) A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science 343:645-648. doi: 10.1126/science.1248229

Boer D.R. et al. (2014) Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. Cell 156:577-589. doi: 10.1016/j.cell.2013.12.027

Appendix 2

Selected research projects that have been enabled by Core Facilities in the period 2010-2014

1. Genomics Core Facility

- **Potential of fecal microbiota for early-stage detection of colorectal cancer**
  Zeller G. et al. (2014) Potential of fecal microbiota for early stage detection of colorectal cancer. Molecular Systems Biol 10:766. doi: 10.15252/msb.20145645

- **Genome sequencing of cancer tissue**
  Weischenfeldt J. et al.(2013) Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell 23:159-70. doi: 10.1016/j.ccr.2013.01.002

  Jones D.T. et al. (2012) ICGC PedBrain: Dissecting the genomic complexity underlying medulloblastoma. Nature 488:100-105. doi: 10.1038/nature11284

  Rausch T. et al. (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 148:59-71. doi: 10.1016/j.cell.2011.12.013

- **Analysis of coding transcript 3'-ends**
  Gupta I. et al. (2014) Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol Syst Biol 10:719. doi: 10.1002/msb.135068

  Wilkening S. et al. (2013) An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res 41:e65. doi: 10.1093/nar/gks1249

- **Evolution of transcriptional regulation**
  Pougach K. at al. (2014) Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat Commun 5:4868. doi: 10.1038/ncomms5868

2. Protein expression and Purification Core Facility

- **Crystal structure of the 14-subunit RNA polymerase I**
  Fernández-Tornero C. et al. (2013) Crystal structure of the 14-subunit RNA polymerase I. Nature 502:644-649. doi: 10.1038/nature12636

- **Structural basis of histone recognition by the FACT chaperone**

Hondele, M. et al. (2013) Structural basis of histone H2A-H2B recognition by the essential chaperone FACT. Nature 499:111-114. doi: 10.1038/nature12242

3. Proteomics Core Facility

- **Structure and composition of the nuclear pore**
  Bui K.H. et al. (2013) Integrated structural analysis of the human nuclear pore complex scaffold. Cell 155:1233-1243. doi: 10.1016/j.cell.2013.10.055

  Ori A. et al. (2013) Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. Mol Syst Biol 9:648. doi: 10.1038/msb.2013.4

- **Quantitative proteomics of Drosophila oocyte maturation**
  Kronja I. et al. (2014) Quantitative proteomics reveals the dynamics of protein changes during Drosophila oocyte maturation and the oocyte-to-embryo transition. Proc Natl Acad Sci U S A. 111:16023-16028. doi: 10.1073/pnas.1418657111

  Kronja I. et al. (2014) Changes in the Posttranscriptional Landscape at the Drosophila Oocyte-to-Embryo Transition. Cell Reports 7:1495-1508. doi: 10.1016/j.celrep.2014.05.002

- **Mapping of kinase-specific phosphorylation sites**
  Polonio-Vallon T. et al. (2014) Src kinase modulates the apoptotic p53 pathway by altering HIPK2 localization. Cell Cycle 13:115-125. doi: 10.4161/cc.26857

- **Identification and characterization of RNA polymerase subunits**
  Fernández-Tornero C. et al. (2013) Crystal structure of the 14-subunit RNA polymerase I. Nature. 502:644-649. doi: 10.1038/nature12636

  Taylor N.M. (2013) RNA polymerase III-specific general transcription factor IIIC contains a heterodimer resembling TFIIF Rap30/Rap74. Nucleic Acids Res 41:9183-9196. doi: 10.1093/nar/gkt664

- **Identification of spindle assembly factors**
  Yokoyama H. et al. (2014) CHD4 Is a RanGTP-Dependent MAP that Stabilizes Microtubules and Regulates Bipolar Spindle Formation. Curr Biol 23:2443-2451. doi: 10.1016/j.cub.2013.09.062

4. Electron Microscopy Core Facility

- **Detailed ultrastructural analysis of cellular components**
  Foresti O. et al. (2014) Quality control of inner nuclear membrane proteins by the Asi complex. Science 346:751-755. doi: 10.1126/science.1255638

- **Correlative light and electron microscopy**
  Romero-Brey I. et al. (2012) Three-dimensional architecture and biogenesis of membrane structures associated with hepatitis C virus replication. PLoS Pathog 8: e1003056. doi: 10.1371/journal.ppat.1003056

  Ronchi P. et al. (2014) Positive feedback between golgi membranes, microtubules and ER-exit sites directs golgi de novo biogenesis. J Cell Sci 127:4620-4633. doi: 10.1242/jcs.150474

  Kukulski W. et al. (2011) Correlated fluorescence and 3D electron microscopy with high sensitivity and spatial precision. J Cell Biol 192:111-119. doi: 10.1083/jcb.201009037

- **Intravital imaging and electron microscopy in model organisms**
  Durdu S. et al. (2014) Luminal signalling links cell communication to tissue architecture during organogenesis. Nature 515:120-124. doi: 10.1038/nature13852

  Karreman M. et al. (2014) Correlating Intravital Multi-Photon Microscopy to 3D Electron Microscopy of Invading Tumor Cells using Anatomical Reference Points. PLoS One 9:e114448. doi: 10.1371/journal.pone.0114448

5. Flow Cytometry Core Facility

- **Characterization of Drosophila developmental enhancer activity**
  Bonn S. et al. (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44:148-156. doi: 10.1038/ng.1064

- **Correlative fluorescence microscopy and flow cytometry methods**
  Mahen R. et al. (2014). Comparative assessment of fluorescent transgene methods for quantitative imaging in human cells. Mol Biol Cell 25:3610-3618. doi: 10.1091/mbc.E14-06-1091

- **mRNA interactome composition and dynamics**
  Castello A. et al. (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149:1393-1406. doi: 10.1016/j.cell.2012.04.031

- ***In vivo* analysis of protein dynamics**
  Khmelinskii A. et al. (2012) Tandem fluorescent protein timers for in vivo analysis of protein dynamics. Nat Biotech 24:708-714. doi: 10.1038/nbt.2281

6. Advanced Light Microscopy Facility

- **Analysis of protein-lipid interactions**

Saliba A.E. et al. (2014). A quantitative liposome microarray to systematically characterize protein-lipid interactions. Nat Methods 11: 47-50. doi: 10.1038/nmeth.2734

- **Genome-wide RNAi screening**
  Almaça J. et al. (2013) High-content siRNA screen reveals global ENaC regulators and potential cystic fibrosis therapy targets. Cell 154:1390-1400. doi: 10.1016/j.cell.2013.08.045

  Gudjonsson T et al. (2012). TRIP12 and UBR5 suppress spreading of chromatin ubiquitylation at damaged chromosomes. Cell 150:697-709. doi: 10.1016/j.cell.2012.06.039

- **Molecular structure of the nuclear pore complex**
  Szymborska A. et al. (2013) Nuclear Pore Scaffold Structure Analyzed by Super-Resolution Microscopy and Particle Averaging. Science 341:655-658. doi: 10.1126/science.1240672

- **Analysis of receptor protein lifetime and influence on chemokine signalling**
  Donà E. et al. (2013) Directional tissue migration through a self-generated chemokine gradient. *Nature* 503:285-289. doi: 10.1038/nature12635

- **Dynamics of embryonic patterning**
  Lauschke V.M. (2013) Scaling of embryonic patterning based on phase-gradient encoding. *Nature* 493:101-105. doi: 10.1038/nature11804

7. Chemical Biology Core Facility

- **Identification of novel:**
  - **Agonist/antagonists of mitochondrial calcium channels**
  - **phosphatase inhibitors**
  - **inhibitors of NMD**

Bhuvanagiri M. et al. (2014) 5-azacytidine inhibits nonsense-mediated decay in a MYC dependent fashion. EMBO Mol Med 1593-1609. doi: 10.15252/emmm.201404461

Appendix 3

Selected technology development highlights 2012-2014

This appendix features a list of selected technology development highlights in which EMBL scientists were involved over the first three years of the current Indicative Scheme.

1. Structural Biology

- **New beamlines at PETRA III**: three state-of-the-art beamlines for SAXS and MX were designed, constructed and brought into regular user operation between late 2012 and early 2013, forming the core of the world-leading high brilliance X-ray radiation source PETRA III at EMBL Hamburg.

- **New beamlines at the ESRF:** EMBL scientists contributed to the design and construction and are now in charge of the operation of several public structural biology beamlines at the ESRF in Grenoble. The beamlines provide the structural biology community with advanced macromolecular crystallography and bioSAXS facilities.

- **MD3 high-precision diffractometer:** a highly versatile diffractometer with a precision compatible to the micro-focus conditions on the beamline P14 at EMBL Hamburg. The MD3 technology, which allows to routinely process micron-sized crystals, has been licensed to the scientific instrumentation companies Bruker and Arinax/Maatel.

- **Serial Crystallography with synchrotron radiation at cryogenic temperatures:** a strategy for collecting data from many micrometre-sized crystals presented to an X-ray beam in a vitrified suspension. By refining structural models previously obtained using free-electron laser radiation, this method opens new avenues for the complementary use of synchrotron and XFEL technologies.

Gati C. et al. (2014) Serial crystallography on in vivo grown microcrystals using synchrotron radiation. IUCrJ 1:87-94. Doi: 10.1107/S2052252513033939

- **Automated crystal harvesting and automated crystal treatment (CrystalDirect):** a new method for automated crystal harvesting based on laser-induced photoablation of thin films. This method is critical for the advancement of challenging projects that require systematic testing of large numbers of crystals.

Cipriani F. et al. (2012) CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. Acta Crystallogr D Biol Crystallogr 68:1393-1399.

Márquez J.A. and Cipriani F (2014) CrystalDirect™: a novel approach for automated crystal harvesting based on photoablation of thin films. Methods Mol Biol 1091:197-203. Doi: 10.1007/978-1-62703-691-7_14

- **New sample holders for frozen crystallography:** new sample holder standards, which increase the capacity of sample changers used at MX beamlines and reduce the storage and transporting costs of frozen samples.

- **BioSAXS sample changer:** an automated sample changer for solution scattering experiments, which allows processing of several hundred samples stored in microtiter plates in one minute turnover.

Round A. et al. (2015) BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. Acta Crystallogr D Biol Crystallogr 71:67-75. Doi: 10.1107/S1399004714026959

- **Automated miniaturized high-throughput pipeline for eukaryotic structural complexomics:** integration of the ACEMBL, MultiBac and polyprotein multi-expression technologies into the first miniaturized fully robotised pipeline for high-throughput structural biology of eukaryotic proteins and their complexes.

- **Automatic data processing routine for the ESRF beamlines:** a system for the rapid automatic processing of MX diffraction data from single and multiple positions on a single or multiple crystals; it was developed through the incorporation of standard integration and data analysis programmes into the ESRF data collection, storage and computing environment.

Monaco S. et al. (2013) Automatic processing of macromolecular crystallography X-ray diffraction data at the ESRF. J Appl Crystallogr 46:804-810.

2. Imaging

a. Probes and reporters

- **New ratiometric fluorescent reporters:** specific small-molecule ratiometric reporters based on energy transfer that allow for spatially resolved monitoring of protease activity.

Gehrig S. et al. (2012) Spatially resolved monitoring of neutrophil elastase activity with ratiometric fluorescent reporters. Angew Chem Int Ed Engl 51:6258-61. doi: 10.1002/anie.201109226

- **Tandem fluorescent reporters:** fusions of two single-color fluorescent proteins that mature with different kinetics to analyse protein turnover and mobility in living cells.

Khmelinskii A. et al. (2012) Tandem fluorescent protein timers for in vivo analysis of protein dynamics. Nat Biotechnol 30:708-14. doi: 10.1038/nbt.2281

- **Chemical dimerizers:** the first rapidly reversible small molecule based dimerisation system with a switch-off sufficiently rapid to determine kinetics of lipid metabolising enzymes in living cells.

Feng S. et al. (2014) A rapidly reversible chemical dimerizer system to study lipid signaling in living cells. Angew Chem Int Ed Engl. 53:6720-3. doi: 10.1002/anie.201402294

- **Genetically encoded click and Diels-Alder chemistry:** methods for genetically encoded copper-free click and Diels-Alder chemistry for site-specific labelling of proteins with fluorogenic dyes in living cells.

Plass T. et al. (2011) Genetically encoded copper-free click chemistry. Angew Chem Int Ed Engl 50:3878-81. doi: 10.1002/anie.201008178

Plass T. et al. (2012) Amino acids for Diels-Alder reactions in living cells. Angew Chem Int Ed Engl 51:4166-70. doi: 10.1002/anie.201108231

- **Single-molecule FRET microscopy:** a novel microfluidic platform that performs multisecond observation of single molecules with millisecond time resolution while bypassing the need for immobilisation procedures.

Tyagi, S. et al. (2014) Continuous throughput and long-term observation of single-molecule FRET without immobilization. Nat Methods 11:297-300. doi: 10.1038/nmeth.2809

- **Live-cell click labelling for fluorescence light microscopy:** design of new unnatural amino acids with improved biocompatibility and stability for rapid dual-colour labelling of live mammalian cells and visualisation by super-resolution microscopy.

Nikić, I. et al. (2014) Minimal Tags for Rapid Dual-Color Live-Cell Labeling and Super-Resolution Microscopy. Angew Chem Int Ed Engl 53(8):2245-9. doi: 10.1002/anie.201309847

b. Light and electron microscopy

- **Multiview SPIM:** a novel multiview light sheet microscope, comprising two detection and illumination objective lenses, which allows rapid *in toto* fluorescence imaging of biological specimens with subcellular resolution (e.g. high quality 3D reconstruction of Drosophila development and the zebrafish brain).

Krzic U. et al. (2012) Multiview light-sheet microscope for rapid in toto imaging. Nat Methods 9:730-3. doi: 10.1038/nmeth.2064

- **Mitotic spindle arrays:** a method to generate arrays of mitotic spindles *in vitro*, using deep UV photochemistry to attach chromatin-coated beads on a glass surface according to a pattern of interest. The immobilised beads act as artificial chromosomes, and induce the formation of mitotic spindles that can be imaged by confocal microscopy.

Tarnawska K. et al. (2014) Mitotic spindle assembly on chromatin patterns made with deep UV photochemistry. Methods Cell Biol 120:3-17. doi: 10.1016/B978-0-12-417136-7.00001-X

- **Golgi depletion from living cells by laser nanosurgery:** a method that uses growth of cells on micropatterns to displace the Golgi complex from its position, and laser nanosurgery to subsequently deplete it from living cells; Golgi-depleted karyoplasts can be imaged by time-lapse microscopy to follow *de novo* Golgi synthesis.

Tängemo C. et al. (2011) A novel laser nanosurgery approach supports de novo Golgi biogenesis in mammalian cells. J Cell Sci 124:978-987. doi: 10.1242/jcs.079640

Ronchi P. et al. (2013) Golgi depletion from living cells with laser nanosurgery. Meth Cell Biol 118:311-324. doi: 10.1016/B978-0-12-417164-0.00019-7

- **Fluorescence correlation spectroscopy:** a microscope based on light-sheet illumination that allows massively parallel fluorescence correlation spectroscopy (FCS) measurements, thereby allowing for quantitative fluorescence imaging of protein diffusion and interaction in living cells.

Capoulade J et al. (2011) Quantitative fluorescence imaging of protein diffusion and interaction in living cells. Nat Biotechnol. 29:835-839. doi: 10.1038/nbt.1928

- **Correlative light and electron microscopy:** a technique that combines light microscopy acquisition on living cells with transmission electron microscopy to allow the precise localisation of dynamic events observed by fluorescence microscopy, and their visualisation at the ultrastructural level.

Spiegelhalter C et al. (2014) Correlative light and electron microscopy: from live cell dynamic to 3D ultrastructure. Methods Mol Biol 1117:485-501. doi: 10.1007/978-1-62703-776-1_21

- **Correlated fluorescence and 3D electron tomography:** a method for direct and highly precise mapping of signals originating from fluorescent protein molecules to 3D electron tomograms in order to visualize cellular processes at the ultrastructural scale.

Kukulski W et al. (2011) Correlated fluorescence and 3D electron microscopy with high sensitivity and spatial precision. J Cell Biol 192:111-119. doi: 10.1083/jcb.201009037

- **SPRING method for EM helical reconstruction:** an image processing package for single-particle based helical reconstruction from electron cryomicrographs combining Fourier based symmetry analysis and real-space helical processing into a single workflow.

Desfosses A. et al. (2014). SPRING - an image processing package for single-particle based helical reconstruction from electron cryomicrographs. J Struct Biol 185:15-26. doi: 10.1016/j.jsb.2013.11.003

- **Sub-tomogram averaging and chemical crosslinking/MS:** a method relying on optimisation of data collection and defocus determination steps to determine the structure of proteins and macromolecular complexes from cryoelectron tomograms using sub-tomogram averaging with unprecedented resolution.

Schur F.K. et al. (2013) Determination of protein structure at 8.5 Å using cryo-electron tomography and subtomogram averaging. J Struct Biol 184:394-400. doi: 10.1016/j.jsb.2013.10.015

- **Mouse embryo image analysis:** tools for semi-automatic cell lineage tracking, gene expression analysis and cell membrane segmentation of the early mouse embryo, which pave the way for quantitative systems-level analysis of embryogenesis.

3. Software

a. Software for structural biology

- **Small-angle scattering data analysis (ATSAS):** a major new release of the ATSAS program suite for small-angle scattering data analysis from biological macromolecules. ATSAS is now used by over 9500 scientists from over 50 countries.

Petoukhov M.V. et al. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr 45:342-350.

- **Determination of protein crystal structures (ARP/wARP):** novel approaches allowing for a more objective determination of macromolecular structure implemented into the ARP/wARP software and applicable to a wide range of structure determinations, including *de novo* phased XFEL nanocrystal diffraction data.

Wiegels T. and Lamzin V.S. (2012) Use of noncrystallographic symmetry for automated model building at medium to low resolution. Acta Crystallogr D Biol Crystallogr. 68:446-453. doi: 10.1107/S0907444911050712

Langer G.G. et al. (2013) Visual automated macromolecular model building. Acta Crystallogr D Biol Crystallogr. 69:635-641. doi:10.1107/S0907444913000565

Barends T.R. et al. (2014) De novo protein crystal structure determination from X-ray free-electron laser data. Nature 505:244-247. doi: 10.1038/nature12773

- ***In silico* ligand-based drug design (ViCi):** an innovative ViCi-based method that identifies novel predicted protein ligands by rapidly screening a large database of compounds. The ViCi software describes small molecule structures using a combination of mathematical descriptors of molecular size, shape and topology.

Carolan C.G. and Lamzin V.S. (2014) Automated identification of crystallographic ligands using sparse-density representations. Acta Crystallogr D Biol Crystallogr. 70:1844-1853. doi: 10.1107/S1399004714008578

- **Crystal reorientation:** new methods and software routines for the alignment of crystals to obtain optimal diffraction in macromolecular crystallography experiments; the software, which is freely available, is in use at several international synchrotrons.

Brockhauser S. et al. (2013) The use of a mini-κ goniometer head in macromolecular crystallography diffraction experiments. Acta Crystallogr DBiol Cristallogr 69:1241-1251. doi: 10.1107/S0907444913003880

b. Software and statistical algorithms for omics data

- **Microbiome analysis (specl and mOTU profiling):** methods that enable accurate and universal delineation of prokaryotic species and reference-independent species profiling using universal marker genes.

Mende D. et al. (2013) Accurate and universal delineation of prokaryotic species. Nat Methods 10:881-884. doi: 10.1038/nmeth.2575

Sunagawa S. et al. (2013) Reference-independent accurate species profiling of metagenomes using universal marker genes. Nat Methods 10:1196-1199. doi: 10.1038/nmeth.2693

- **RNA sequencing data analysis (easyRNASeq):** a complete pipeline for data analysis of RNA sequencing data, which simplifies data processing by bringing the complex interplay of the required software packages into a single functionality.

Delhomme N. et. al. (2012) easyRNASeq: a bioconductor package for processing RNA-Seq data. Bioinformatics 28:2532-2533.

- **Differential Expression Analysis for sequencing data (DESeq):** a software and statistical model that allows the detection of dynamic changes in expression between different samples from next generation sequencing data. DESeq has been downloaded by over 10,000 users and has >1000 citations.
Anders S. and Huber W. (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. doi: 10.1186/gb-2010-11-10-r106

Love M.I. et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550

- **Singe-cell RNA sequencing data analysis:** a quantitative statistical method to distinguish true biological variability from the high levels of technical noise in single-cell RNA-seq experiments.

Brennecke P. et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 10:1093-1095. doi: 10.1038/nmeth.2645

- **Bioconductor:** an international open-source project for the production of software packages covering a range of bioinformatics and statistical applications for the analysis of high-throughput omics data, co-founded and co-led by EMBL scientists.

Huber W. et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 2015 12:115-21. doi: 10.1038/nmeth.3252

4. Omics and high throughput technologies

- **Sub-microliter microfluidic based microsonicator**: a miniaturized DNA shearing device capable of processing sub-microliter samples based on acoustic shearing within a microfluidic chip. In view of its small scale, this device could represent a first step towards performing ChIP experiments with clinical samples.

Tseng Q. et al. (2012) Fragmentation of DNA in a sub-microliter microfluidic sonication device. Lab Chip 12:4677-82

- **Liposome array for high-throughput screening of protein-lipid interactions:** a liposome microarray based assay that measures protein recruitment to membranes in a quantitative, automated, multiplexed and high-throughput manner.

Saliba, A.-E. et al. (2014) A quantitative liposome microarray to systematically characterize protein-lipid interactions. Nat Methods 11:47-50. doi: 10.1038/nmeth.2734

- **High-throughput cell lysis assay:** a new high-throughput quantitative assay for monitoring envelope perturbations and/or cell lysis in a wide range of bacterial cells. It can be used for screening of novel antibiotics and adjuvants for their ability to penetrate the bacterial envelope.

Paradis-Bleau C. et al. (2014) A genome-wide screen for bacterial envelope biogenesis mutants identifies a novel factor involved in cell wall precursor metabolism. PLoS Genet 10:e1004056. doi: 10.1371/journal.pgen.1004056

- **Functional single-cell hybridoma screening using droplet-based microfluidics:** a microfluidic platform allowing to rapidly screen several hundred thousands hybridoma cell clones for the release of antibodies with a desired functional property (e.g. binding, inhibition of specific drug targets, etc.).

El Debs B. et al. (2012) Functional single-cell hybridoma screening using droplet-based microfluidics. Proc Natl Acad Sci USA. 109:11570-11575. doi: 10.1073/pnas.1204514109

- **Transcript isoform sequencing (TIF-Seq):** a method that allows the genome-wide profiling of full-length transcript isoforms defined by their exact 5' and 3' boundaries.

Pelechano V. et al. (2014) Genome-wide identification of transcript start and end sites by Transcript Isoform Sequencing, TIF-Seq. Nat Protoc 9:1740-1759. 0.1038/nprot.2014.121

- **Batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP):** a universal method enabling cell-type-specific ChIP for analysis of histone modifications, transcription factor binding, or polymerase occupancy within the context of a multicellular organism or tissue (e.g. a developing embryo).

Bonn S. et al. (2012) Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. Nature Protoc 7:978-994. doi: 10.1038/nprot.2012.049

5. Databases and resources

- **TRACER:** a resource that centralizes information from a large on-going functional exploration of the mouse genome with different transposon-associated regulatory sensors. TRACER data can be easily accessed and provides information on the regulatory activities present in a large number of genomic regions, notably in gene-poor intervals that have been associated with human diseases.

Ruf S. et al. (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. Nat Genet 43: 379–386. doi: 10.1038/ng.790

Chen C.K. et al. (2013) TRACER: a resource to study the regulatory architecture of the mouse genome. BMC Genomics 14:215. doi: 10.1186/1471-2164-14-215

- **DEPOD (human DEPhOsphorylation Database):** an online resource with information about active human phosphatases, their substrates, and the pathways in which they function. The database includes links to kinases and chemical modulators of phosphatase activity and contains a sequence similarity search function for identifying related proteins in other species.

Li X. et al. (2013) Elucidating human phosphatase-substrate networks. Sci Signal 6:rs10. doi: 10.1126/scisignal.2003203

- **The European Nucleotide Archive (ENA):** Europe's primary nucleotide sequence resource; it collects, maintains and presents comprehensive nucleic acid sequence and related information, covering the spectrum from raw data to assembled and functionally annotated genomes. ENA has undergone major content developments and service enhancements over the last few years.

Leinonen R. et al. (2011) The European Nucleotide Archive. Nucleic Acids Res 39(database issue):D28-31. doi: 10.1093/nar/gkq967

Amid C. et al. (2012) Major submissions tool developments at the European Nucleotide Archive. Nucleic Acids Res 40(database issue):D43-47. doi: 10.1093/nar/gkr946

Cochrane G. et al. (2013) Facing growth in the European Nucleotide Archive. Nucleic Acids Res 41(database issue):D30-35. doi: 10.1093/nar/gks1175

Pakseresht N. et al. (2014) Assembly information services in the European Nucleotide Archive. Nucleic Acids Res 42(database issue):D38-43. doi: 10.1093/nar/gkt1082

- **The 1000 Genomes Project:** the largest coordinated data production and analysis project yet undertaken in genomics, with data on more than 2500 human genomes from around the world. Our scientists lead the project's data coordination and make the datasets freely available to the scientific community through EMBL-EBI online resources.

Mills R.E. et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59-65. doi: 10.1038/nature09708

Conrad DF, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43:712-714. doi: 10.1038/ng.862

Lappalainen T. et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506-511. doi: 10.1038/nature12531

- **Ensembl Genomes:** an integrative resource that provides tools for annotation, analysis and dissemination of genome-scale data from non-vertebrate species through a consistent set of programmatic and interactive interfaces. Recent developments include the addition of important new genomes and related datasets including crop plants, vectors of human disease and eukaryotic pathogens, along with an upscaling in the representation of bacterial genomes (now over 9000).

Kersey P.J. et al. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res 40(database issue):D91-97. doi: 10.1093/nar/gkr895

Kersey P.J. et al. (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res 42:D546-552. doi: 10.1093/nar/gkt979

- **BioSamples:** a database that stores information about biological samples used in molecular experiments and provides an integration point between technology-specific databases at the EMBL-EBI, projects such as ENCODE and reference collections such as cell lines, thus allowing researchers to cross-reference multiple datasets that pertain to a single sample.

Gostev M. et al. (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. Nucleic Acids Res 40(database issue):D64-70. doi: 10.1093/nar/gkr937

Faulconbridge A. et al. (2014) Updates to BioSamples database at European Bioinformatics Institute. Nucleic Acids Res 42(database issue):D50-52. doi: 10.1093/nar/gkt1081

- **Europe PubMed Central (PMC):** a free online archive offering access to full-text biomedical and life sciences journal literature that benefits researchers throughout the world.

McEntyre J.R. et al. (2011) UKPMC: a full text article resource for the life sciences. Nucleic Acids Res 39(database issue):D58-65. doi: 10.1093/nar/gkq1063

Kafkas Ş. et al. (2013) Database citation in full text biomedical articles. PLoS One 8:e63184. doi: 10.1371/journal.pone.0063184

- **EMBL-EBI Metagenomics**: an integrated metagenomics portal that allows users to easily submit raw nucleotide reads for functional and taxonomic analysis by a state-of-the-art pipeline, and have them automatically stored – together with descriptive, standards-compliant metadata – in the European Nucleotide Archive.

Hunter C.I. et al. (2012) Metagenomic analysis: the challenge of the data bonanza. Brief Bioinform 13:743-746. doi: 10.1093/bib/bbs020

Hunter S. et al. (2014) EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 42(database issue):D600-606. doi: 10.1093/nar/gkt961

- **The Enzyme Portal:** an EMBL-EBI portal that mines and displays comprehensive information about enzymes – including protein sequence, biochemical reactions, biological pathways, small molecule chemistry, disease information, 3D protein structures and relevant scientific literature – from public repositories via a single search.

Alcántara R. et al. (2013) The EBI enzyme portal. Nucleic Acids Res 41(database issue):D773-780. doi: 10.1093/nar/gks1112

- **MetaboLights:** the first general-purpose, open-access curated repository for metabolomics studies, their experimental data and associated metadata, maintained by the EMBL-EBI. The MetaboLights repository is cross-species and cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations, concentrations and raw data from metabolic experiments.

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

Haug K. et al. (2013) MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 41(database issue):D781-786. doi: 10.1093/nar/gks1004

Salek R.M. et al. (2013) Dissemination of metabolomics results: role of MetaboLights and COSMOS. Gigascience 2:8. doi: 10.1186/2047-217X-2-8

Salek R.M. et al. (2013) The MetaboLights repository: curation challenges in metabolomics. Database (Oxford) 2013:bat029. Salek R.M. et al. (2013) The MetaboLights repository: curation challenges in metabolomics. Database (Oxford) 2013, bat029. doi: 10.1093/database/bat029

- **The Resource Description Framework (RDF) platform:** a platform that uses the emerging RDF technology to provide easy links between related but differently structured information, enabling the meaningful and intuitive sharing of molecular data amongst different applications. It provides a new entry point to querying and exploring integrated resources available at EMBL-EBI, such as UniProt, ChEMBL, the Expression Atlas, Reactome, BioSamples and BioModels.

Jupp S. et al. (2014) The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30:1338-1339. doi: 10.1093/bioinformatics/btt765

- **European Variation Archive (EVA):** launched in 2014, it is the first archival resource at EMBL-EBI to provide a single access point for submissions, archiving, and access to high-resolution genetic variation data of all types. The data in the EVA, which includes 1.7 billion submitted variants, is linked with external resources including Ensembl, the European Genome-phenome Archive (EGA) and the European Nucleotide Archive (ENA).

- **The International Mouse Phenotyping Consortium (IMPC):** the first comprehensive, functional catalogue of a mammalian genome, offering free, unrestricted access to a centralised data resource for mutant mice and related gene-to-phenotype associations. Tools for automated statistical analysis, annotation with biomedical ontologies and data integration with other resources make this an invaluable platform for researchers studying the genetic contributions of genes to human diseases.

Mallon A.M. et al. (2012) Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. Mamm Genome 23:641-652. doi: 10.1007/s00335-012-9428-9

Koscielny G. et al. (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res 42(Database issue):D802-809. doi: 10.1093/nar/gkt977