

SECTION I

Executive Summary

EMBL – A Unique Track Record of Scientific Excellence

Since its founding in 1974, the European Molecular Biology Laboratory (EMBL) has sought to serve its member states by making fundamental discoveries in molecular biology that fuel a rich economy of knowledge, training researchers and research leaders at the leading-edge for academic and industrial research, as well as by developing new technologies that meet societal needs in the areas of medicine, agriculture, and ecology.

EMBL is Europe's only intergovernmental organisation for life science research. Established to advance the study of molecular biology across Europe, to nurture young talent, new ideas, and technologies, EMBL is constantly evolving and innovating. EMBL undertakes pioneering research and provides cutting-edge biological services and infrastructures that are essential for European science. EMBL's foundational research has, during the past four decades, enabled a better understanding of the molecular basis of life. Since its creation during the era of classical molecular biology, EMBL has played a leading role in the rapid evolution of modern biology, participating in major breakthroughs in the realms of structural, cellular and developmental biology, in the birth of genomics and molecular genetics, and – more recently – in integrative approaches and systems biology. EMBL scientists have discovered many of the fundamental principles by which organisms are built, including how their building blocks are generated, modified, and brought together in time and space. EMBL research has generated cutting-edge facilities to serve member states and beyond.

EMBL occupies a unique position in European science, possessing incredibly strong foundations across all areas of molecular biology. It attracts talented individuals from many disciplines and trains the next generation of scientists who then move on to become global leaders in Europe and around the world. EMBL has played a major role in laying the groundwork of the current scientific revolution, and has spurred the development of many of the tools that scientists use today, including cryo-electron microscopy (cryo-EM), genomics, and advanced imaging. The vital nature of these technologies is aptly illustrated by their global deployment in the current fight against COVID-19 and other diseases.

Molecules to Ecosystems, the New EMBL Programme 2022–2026

EMBL strives to be at the forefront of modern biology and to build the foundations for future success. The curiosity-driven scientific discoveries made at EMBL serve as the basis for the next generation of applications and discoveries, driving new technology developments and service provision.

Through its next five-year scientific Programme, EMBL intends to propel Europe into a new era of biological understanding, from the molecular building blocks of life through to the complexity of ecosystems – the context within which all life forms exist. EMBL's ambition in the new Programme, titled Molecules to Ecosystems, is to establish the molecular basis of **life in context**, to gain new knowledge that is relevant to understanding life on Earth, and to provide translational potential to support advances in human and planetary health.

Most molecular biology research has focused on studying organisms that are isolated in a laboratory setting, where environmental variation can be minimised. However, life does not happen in isolation, but in the context of communities where organisms interact with each other and respond to constantly changing physicochemical conditions. Due to incredible advances in technology and quantitative data generation, molecular biologists now have the capacity to follow the dynamics of living matter in real time and at multiple scales. Gaining molecular and mechanistic insights to understand how organisms respond to changing environments, and how they influence their environment, are at last feasible. This will be of fundamental importance in achieving a true understanding of the basis of life. It will also be relevant to the important issue of scientific reproducibility, as differences in biological samples due to subtle environmental parameters can influence the interpretation of experimental data.

The challenges facing life on Earth today are huge: the spread of infectious diseases, the loss of biodiversity, environmental degradation, and climate change. To take on these challenges, drastic improvements are needed in understanding the processes of life in their natural context. A better understanding of life means a greater ability to preserve it. EMBL will seize this unique opportunity at a critical time for society, and will undertake bold and potentially transformative discovery science that will also be a force for good for humankind and the planetary ecosystem services that sustain us.

Through the new Programme, EMBL will build on its existing and globally recognised expertise in molecular biology to expand into new areas including planetary biology, human ecosystems, infection biology, and microbial ecosystems. Central to the research strategy will be the development of advanced data sciences and theoretical approaches. Through close collaborations with scientists from different domains and within all of EMBL's member states, this Programme will enable EMBL to build new bridges with disciplines, including ecology and epidemiology, while keeping its firm foundations in molecular biology. Within this Programme, EMBL will also look to train a new generation of interdisciplinary scientists who will address real-life scientific questions and prepare for future challenges.

EMBL is well placed to fulfil its bold ambition of gaining a molecular understanding of life in context. In doing so, EMBL will continue to uphold its special responsibility to lead and coordinate European life sciences in its role as Europe's flagship life sciences research organisation. The new Programme will be a truly pan-European initiative, providing scientific services and innovation, and sharing expertise with all of EMBL's member states, while harnessing the strong and dynamic collaborations and networks of partnerships that EMBL has built over many years.

As an international organisation, made stronger by the breadth of its member states and by its physical presence at six sites across five host countries, EMBL is uniquely positioned to deliver this ambitious and timely programme. EMBL's success is due to its dynamic turnover, interdisciplinarity, and a distinctive scientific culture that blends ambition, excellence, cooperation, and openness across borders and societies. EMBL's Programme sets out plans to enhance research coordination, to promote joint standards and open science, and to inform and impact international research and policymaking. The knowledge and technological advances unlocked as a result of this new Programme will directly help EMBL's member states to better understand and address the planetary challenges of climate change, pollution, food security, and emerging pathogens.

1. Introduction

The Foundations of Molecular Biology

Molecular biology can be considered as the collective understanding of how the rich diversity of life on Earth works at the level of molecules, such as DNA, RNA, proteins, lipids, and others. This field of science was created more than half a century ago, with the aim of deciphering how living organisms function, and has led to some of the greatest discoveries of the 21st century. Curiosity-driven research in molecular biology has provided humanity with a wealth of knowledge, including the capacity to read the blueprint of life, to observe and understand how organisms are formed, to identify the events that lead to disease, and to develop new treatments.

Modern biology is in a period of spectacular progress, accompanied by an unprecedented level of excitement. Transformative research technologies such as single-cell sequencing and cryo-EM, coupled with the ability to generate, perturb, and analyse biological systems at scale, have revolutionised our understanding of the core molecular processes that define life. Pioneering developments in molecular biology have applications in new areas of medicine, agriculture, and biotechnology, with genomics being used in patient diagnoses, cellular imaging for efficient food production, and structural biology for the design of drugs and the creation of innovative consumer goods. The discoveries of fundamental molecular biology research include many of the genetic drivers of development and disease. However, despite the capacity to capture and read so much biological information, in reality only a modest fraction of the fundamental workings of living matter is actually known and what has been learnt in the past 50 years is just the beginning.

The Future of Molecular Biology

The Next Frontier

Over recent decades, remarkable molecular insights have been made in model organisms ranging from bacteria to animals, under defined lab conditions. However, living things do not exist in isolation. From plankton in oceans to bacteria in the human gut, every organism in nature is part of a complex and dynamic ecosystem, living in community with other organisms, in physical and chemical environments. An ecosystem, or biome, is defined as a single environment comprising every living organism (biotic) and non-living factor (abiotic) contained within it. From unicellular to complex multicellular organisms, all living systems must respond and adapt to the environment in which they live in order to survive.

While the impact of the environment on phenotypes (the observable characteristics or traits of an organism) is well described at the organism and population levels, the underlying molecular processes and mechanisms remain relatively unstudied. The principles underlying phenotypic variation and the responsiveness of organisms to changing environments have hardly been tackled at the molecular level. **Looking to the future, molecular biology can help to reveal these processes and mechanisms in order to understand how organisms function in their natural contexts.**

The Societal and Economic Value of Ecosystems

Ecosystems, both in terms of biodiversity and balanced interrelationships among organisms, are fundamental to life on our planet and to human well-being. However, human action is destroying ecosystems on a massive scale. Accelerating pollution, deforestation, and climate change, coupled with environmental policy failure, have created major environmental problems such as biodiversity loss, threats to public health including pandemics, and ecosystem collapse.

One major reason for this is that the value of ecosystems to human welfare is severely underestimated. A study carried out by environmental scientists and public policymakers (Living Planet Report 2018; Costanza, R. *et al.* Changes in the global value of ecosystem services. (2014) *Global Environmental Change* 26: 152-158) estimated the notional economic value of **ecosystem products and services** (such as the provision of food, water, fuel, and other raw materials, the pollination of crops, and the prevention of floods or soil erosion) to be US\$125 trillion per year. This value, and conversely the cost to society of losing these products and services if they are not protected, is around two-thirds higher than global GDP. These services also include benefits for human health: the destruction of the planet's ecosystems means that, for the first time in modern history, humanity faces the prospect of losing many of the benefits that medicine has brought, possibly jeopardising advances made over the past two centuries.

Leading the Future

The world is now facing an urgent need to find solutions to challenges such as climate change, the loss of biodiversity, environmental degradation, antibiotic resistance, environmentally driven epidemics, and human diseases such as diabetes, cancer, and mental illness. It is crucial that the life sciences play a leading role in developing new knowledge and innovations for mitigating the impact of human action on ecosystems. A new era of molecular biology, encompassing ecosystems, is needed to help us understand and revolutionise planetary and human health.

Knowledge of ecosystems at the molecular level will be pivotal for the next wave of scientific discoveries, such as an understanding of the emergence of infectious diseases, vaccine development for evolving pathogens, modelling the human brain, and providing ecological therapies for a burdened planet. These are some of the crises of modern society that can be transformed by molecular biology in the next decade.

Molecules to Ecosystems: EMBL's vision is to advance our understanding of ecosystems at the molecular level, applying expertise in molecular biology to study life in its natural context. In so doing, EMBL aims to use fundamental science to tackle societal challenges.

Molecules to Ecosystems

Scientists now have many of the molecular tools needed to address fundamental questions about **life in context**. New technologies are now being developed to collect measurements of ecosystem components at unprecedented volumes, from molecules to cells, organisms, populations, and communities, alongside chemical and physical environmental parameters (Figure IN1). Advances in computational power and artificial intelligence (AI) have also enabled the rigorous analysis and creative integration of these data. This tremendous technological progress in the life sciences can now be coupled with the capacity to gather and analyse data of greater scope, resolution, and quality than ever before. This means that measurements of environmental context can be collected in systematic ways, allowing for the integration of this new level of complexity into the study of biology. Building on its established expertise in molecular biology, EMBL can now take on the study of life in its natural context. **In this new scientific era, researchers at EMBL will strive to understand ecosystems at the molecular level.**




Figure IN1 | Molecules to Ecosystems.

EMBL's ambitious new Molecules to Ecosystems Programme will leverage EMBL's strengths in molecular biology, interdisciplinary research, and its collaborative spirit to work towards a fundamental understanding of life, including human life, in the context of populations and environments. This cutting-edge, interdisciplinary, and societally relevant Programme will incorporate novel areas of research and new technologies. It will also expand EMBL's horizons, creating a new era for the life sciences, while maintaining EMBL's core values of scientific excellence and fundamental research.

The new EMBL Programme will explore **life in context** by studying **both classical and novel model organisms** in the context of their real-world environment. Longitudinal studies will be performed, collecting comprehensive data in specific areas of Europe, in close collaboration with institutes from EMBL's member states. **Fieldwork**, bringing together molecular biologists and ecologists (theoretical and experimental), epidemiologists, and environmental biologists, will be a critical component in better understanding environmental effects on molecular mechanisms and organism composition, symbiotic states, and host–pathogen interactions. Most importantly, to study life **across interconnected scales** (cells, tissues, organisms, populations) in different genetic and environmental contexts, **lab experiments** will be needed to induce **controlled perturbations of genetic or environmental factors** and to measure their impact to gain mechanistic understanding.

In collaboration with EMBL member states, EMBL will initiate specific projects within **mobile labs**, containing specialised equipment and staffed by dedicated personnel with technical and research expertise. EMBL will work in **partnership** with scientists from various fields, including ecologists, zoologists, epidemiologists, population geneticists, climatologists, and toxicologists, to realise these ambitions. EMBL aims to develop **advanced sampling approaches** for gathering data from observational studies. Appropriate computational tools, including AI and theoretical approaches, will be required to extract new knowledge and biological principles from these large and complex datasets.

EMBL researchers working in areas relevant to many of the new themes have held or planned topic-specific workshops to bring together interdisciplinary experts, discuss strategic synergies, and solicit guidance for EMBL's future plans, providing a rapid horizon scan of the scope and challenges in these research areas. EMBL also plans to establish focused advisory boards for some of the new scientific themes, whose support and direction will be critical in pursuing research questions that are of the most scientific value. To explore the feasibility of expanding EMBL research in these new directions, EMBL initiated a set of **pilot projects**. Some of these were newly conceived proof-of-principle projects, while others were extensions of existing projects relevant to the Molecules to Ecosystems research themes. The projects are described in the following chapters, preceded by the pilot project icon .

Why EMBL?

EMBL is positioned at the heart of the current revolution in the life sciences. With its dynamic turnover model and its vibrant community of young scientists, EMBL is Europe's flagship laboratory for fundamental research in molecular biology. EMBL enables its 27 member states and two associate member states to join forces and be at the forefront of the life sciences on the global stage. Professor Edith Heard was appointed as Director General of EMBL in 2019 by EMBL's governing body, the EMBL Council, composed of representatives of all member and associate member states. Her appointment corresponded to a wish by Council to usher in a new era at EMBL.

EMBL is Europe's only intergovernmental organisation for life science research. Since its creation in 1974, EMBL's goal has always been to foster the development of excellent scientists, to enable discoveries in molecular biology that fuel a rich economy in knowledge, and to develop technologies that meet societal needs in the areas of medicine, agriculture, and ecology. Today, almost 1,800 personnel from more than 90 countries advance EMBL's activities across its six sites: Heidelberg (headquarters) and Hamburg in Germany, EMBL's European Bioinformatics Institute (EMBL-EBI) in Cambridge in the United Kingdom, Grenoble in France, Rome in Italy, and Barcelona in Spain. Each site offers a unique and highly complementary mix of research and services (Chapter 14: People, Processes, and Places).

The principle of regular staff turnover (nine-year rule) and EMBL's five-year programmatic funding enable the organisation to be agile and to evolve rapidly via the recruitment of talented scientists studying exciting questions in biology using cutting-edge approaches (Figure IN2). The regular turnover at EMBL also provides Europe with a regular supply of highly trained scientific personnel. EMBL's operational model has been employed across Europe and has served as a framework for many national centres of excellence.

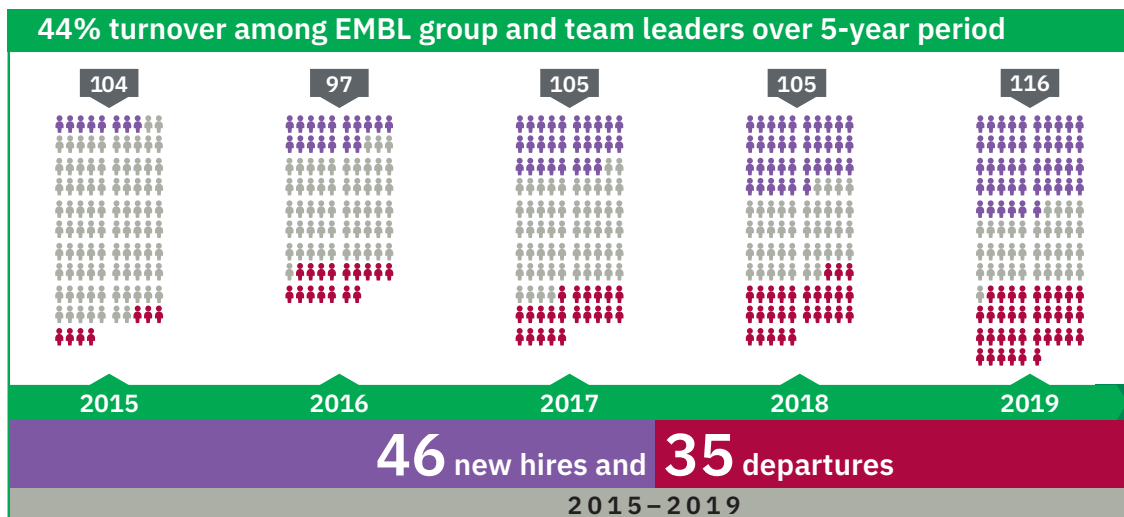


Figure IN2 | EMBL's dynamic turnover among group and team leaders.

In the five-year period up to and including 2019, nearly half of EMBL's group and team leaders joined EMBL (purple) or departed (red). This model, based on a nine-year rule, ensures that EMBL can quickly and seamlessly adapt and develop its scientific directions via the continuous intake of new group and team leaders.

EMBL's Missions and Achievements

Mission 1: To perform excellent fundamental research in molecular biology

EMBL's **overarching goal** is to understand the molecular basis of life. Research at EMBL emphasises experimental and computational analyses of biological organisation, from the molecule to the organism. Research areas cover a wide spectrum of biology, including structural biology, genome biology, cell biology, developmental biology, tissue and organ biology, neurobiology, microbiology, bioinformatics and computational biology, and molecular medicine. EMBL prides itself on the collaborative and interdisciplinary nature of its work, reflected in the high percentage of papers published and grants secured in collaboration with academic and industry research groups in EMBL member states and worldwide (Figure IN3).

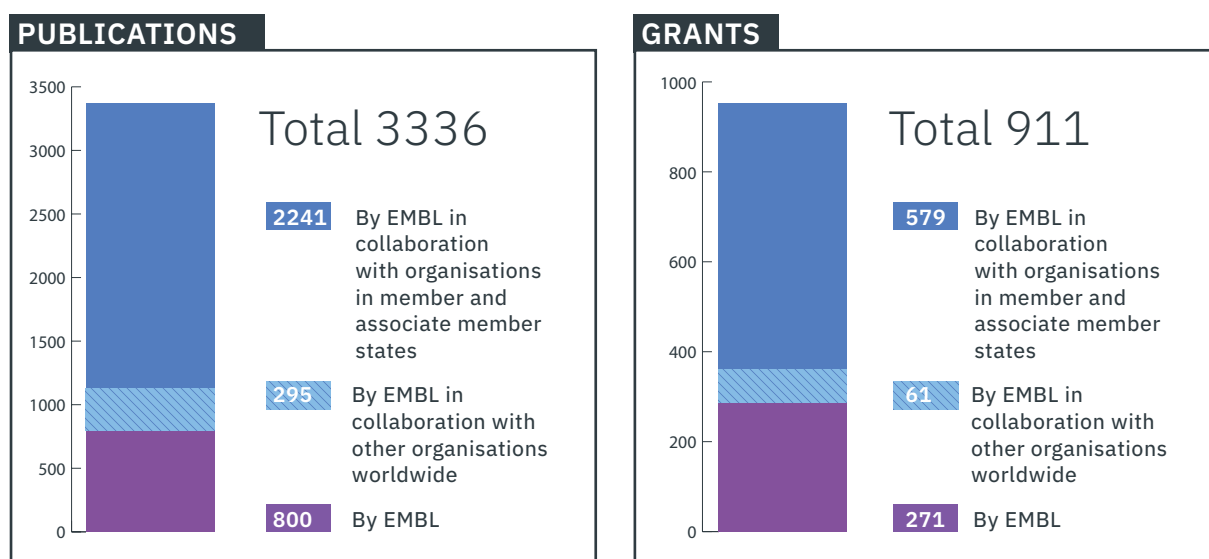


Figure IN3 | EMBL collaborations: publications and grants (2015–2019).

Of the 3,336 scientific publications by EMBL from 2015–2019, more than three-quarters (76%) were published as part of collaborations with other organisations worldwide, many of which are in member states. Among the 911 grants secured by EMBL during the same period, including 45 ERC grants, a similar proportion (70%) involved collaborations with organisations worldwide, with the majority of those organisations also in EMBL member states.

EMBL has a long and distinguished history of groundbreaking scientific achievements (Figure IN4). Among the most notable contributions is that made by Nobel laureates Christiane Nüsslein-Volhard and Eric Wieschaus, who unravelled the genetic and molecular mechanisms by which multicellular organisms develop. Such research has contributed to the subsequent understanding of the molecular pathways that are disrupted in the development of cancer. EMBL has also made significant contributions to the study of the fundamental units of life, such as the cell and its molecular components. For example, EMBL research has led to the characterisation of the cellular transport machinery, analysis of cytoskeleton organisation, and an understanding of the function and regulation of RNA metabolism.

In parallel, groundbreaking technologies are also developed at EMBL, driven by fundamental biological questions. For example, cryo-EM was developed by Nobel laureate Jacques Dubochet during a quest to decipher protein structures. The further development of cryo-EM technology at EMBL has enabled scientists to study biological structures *in situ*, leading to new medicines and vaccines. Other notable inventions include the first functional light-sheet microscope by Ernst Stelzer, used to track live cell movements in embryos, and mass spectrometry-based protein analyses by Matthias Mann. All of these technologies are widely used today in academia and industry.

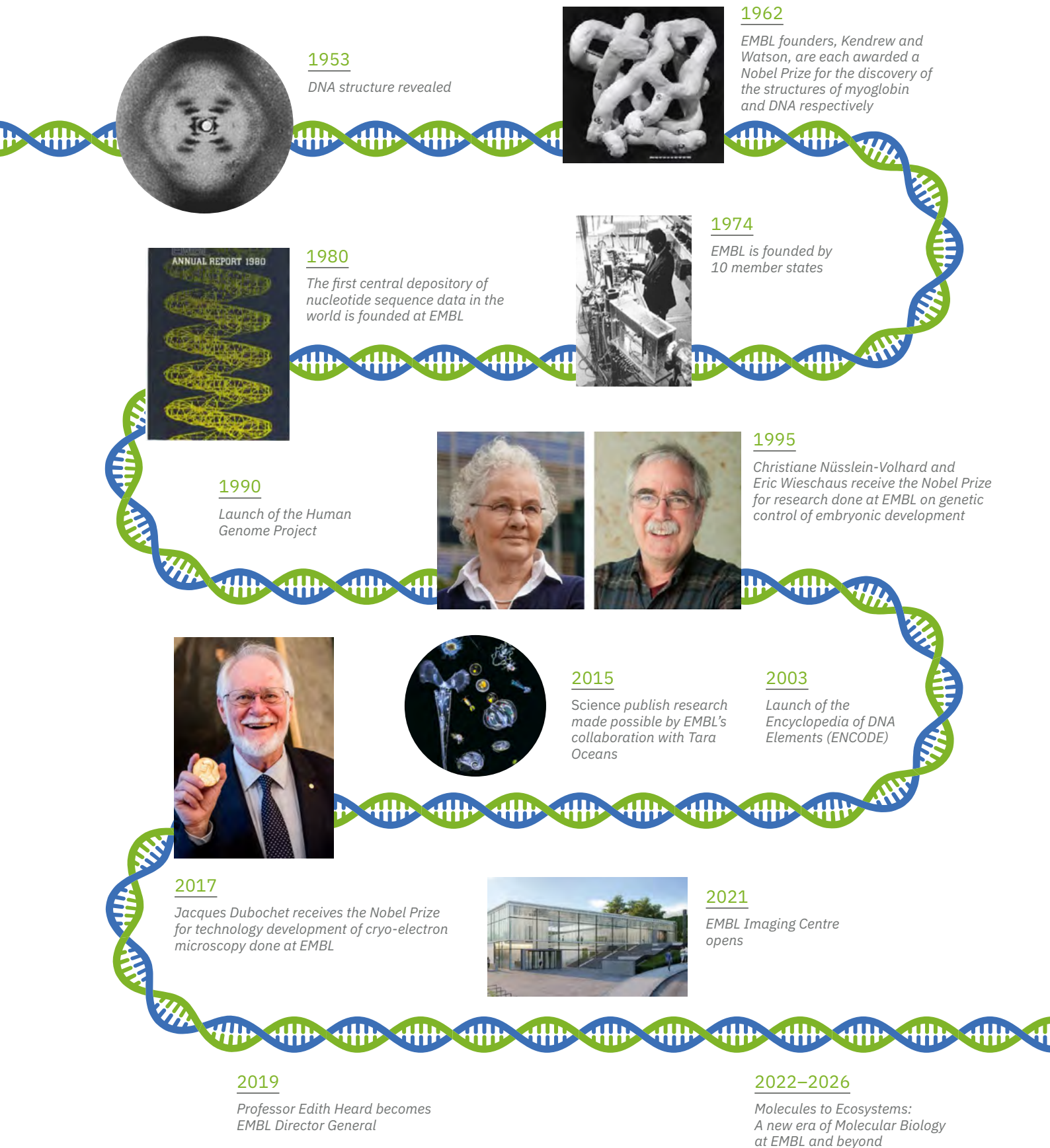


Figure IN4 | The foundations of molecular biology and the birth of EMBL.

EMBL's foundations and successes have been built upon the discoveries and visions of its scientific predecessors. EMBL's future will rely on its expertise, experience, and collaboration with the scientific community to drive forward a new era of life science research.

Mission 2: To offer vital services to scientists in the member states and around the world

The scientific services provided by EMBL include data, structural biology, and imaging services, as well as state-of-the-art core facilities.

Core biomolecular databases and bioinformatics tools

EMBL pioneered open access data resources in the 1980s and currently hosts the most comprehensive integrated set of open biomolecular data in the world. Over forty data resources are developed and made openly available to the worldwide scientific community by EMBL-EBI. Databases include information on hundreds of millions of genome and RNA sequences, protein structures, protein folding domains, cell metabolites, phenotypes, and on the effects of drugs on cells and tissues, as well as biological image data. EMBL-EBI's open sharing of biological data in standardised formats with the life science community has been integral to generating countless research insights worldwide. These data resources have become a fundamental infrastructure for genomic medicine and the analysis of complex microbial ecosystems, and are becoming critical in other areas including agritech and biodiversity tracking.

Beamlines, instrumentation, and high-throughput technology for structural biology

EMBL provides structural biology infrastructure for biologists from all over Europe at the European Synchrotron Radiation Facility (ESRF) in Grenoble and Deutsches Elektronen-Synchrotron (DESY) in Hamburg. At each site, synchrotron beamlines for macromolecular crystallography and small-angle X-ray scattering are complemented by advanced sample preparation facilities offering integrated access to services, expertise, and user training. These are widely used by the European and global scientific community in conjunction with EMBL scientists, and have resulted in landmark discoveries. One recent illustrative example is the discovery of the nature of protein–RNA complexes involved in viral replication, which is expected to inform research on diseases like COVID-19.

Imaging facilities with access to world-class microscopy and technologies

At EMBL Heidelberg, the new EMBL Imaging Centre is scheduled to begin operations in 2021. The centre will offer access to the latest light and electron microscopy technologies, along with data analysis facilities and expert support. In addition, EMBL Barcelona hosts a Mesoscopic Imaging Facility which, in conjunction with the Electron Microscopy Facility and Advanced Light Microscopy Facility in Heidelberg, provides scientists with access to microscopy and modelling technologies designed for studying tissues.

Core facilities that provide cost-effective and efficient access to methods and technologies

EMBL's core facilities offer scientists at EMBL and in its member states access to state-of-the-art equipment and expert support, enabling them to achieve their research goals. The facilities currently offer services in the following areas: advanced light microscopy, chemical biology, electron microscopy, flow cytometry, genomics, metabolomics, protein expression and purification, and proteomics, genetic and viral engineering, histology, gene editing and embryology. These services are provided by EMBL experts who also share their knowledge with the broader scientific community.

Mission 3: To train scientists, students, and visitors at all levels

EMBL's PhD and postdoctoral research programmes provide world-class training for scientists in a collaborative and interdisciplinary environment. The EMBL International PhD Programme, with over 200 PhD students at any one time, supports students in gaining early independence through a combination of dedicated mentoring and creative freedom. Cofunded by the EU's Marie Skłodowska-Curie Actions, the EMBL Interdisciplinary Postdocs (EIPOD) Programme provides training and career development opportunities for young researchers. A pioneering new fellowship programme, ARISE, was recently launched to train engineers and technology developers to become research infrastructure scientists and leaders.

EMBL's courses and conferences cover a diverse range of topics and bring together experts to share new ideas and techniques, foster collaborations, and develop strategies to drive future research. In 2019, nearly 7,500 participants from 86 countries attended courses and conferences across EMBL's sites. EMBL also promotes scientific excellence through its Scientific Visitor Programme, which provides opportunities for visiting scientists and students to benefit from new technologies and state-of-the-art equipment in EMBL laboratories and core facilities.

Mission 4: To actively engage in technology transfer and industry relations

Scientists at EMBL often seek innovative ways to answer biological questions, frequently developing new technologies and methods in close collaboration with industrial partners as part of the process. EMBL's technology transfer arm, EMBLEM, facilitates the process of identifying and protecting intellectual property, enabling the establishment of EMBL spin-off companies, developing collaborative research agreements, and licensing technologies to third parties. EMBL also partners with industry in large-scale public–private research collaborations, such as Open Targets, which have led to publications and data platforms that advance industry-driven questions.

Mission 5: To coordinate and integrate European life science research

EMBL fosters international collaboration between scientific communities in Europe and around the world by playing a leading role in shaping scientific strategy and policy. EMBL founded the European Strategy Forum on Research Infrastructures (ESFRI) projects ELIXIR and Euro-BioImaging, playing important leadership roles in both organisations and additionally is a member of Instruct-ERIC. Via its Partnership Programme (Figure IN5), EMBL has helped to establish institutes of excellence spanning the life sciences in many of its member states, some with the aim of strengthening less research-intensive regions. EMBL also maintains a strong relationship with the European Commission (EC) and regularly engages with the EC on European science policy issues, thereby contributing to the future direction of European framework programmes. EMBL also contributes to European science policy as a founding member of EIROforum, an alliance of eight intergovernmental research infrastructures in Europe.

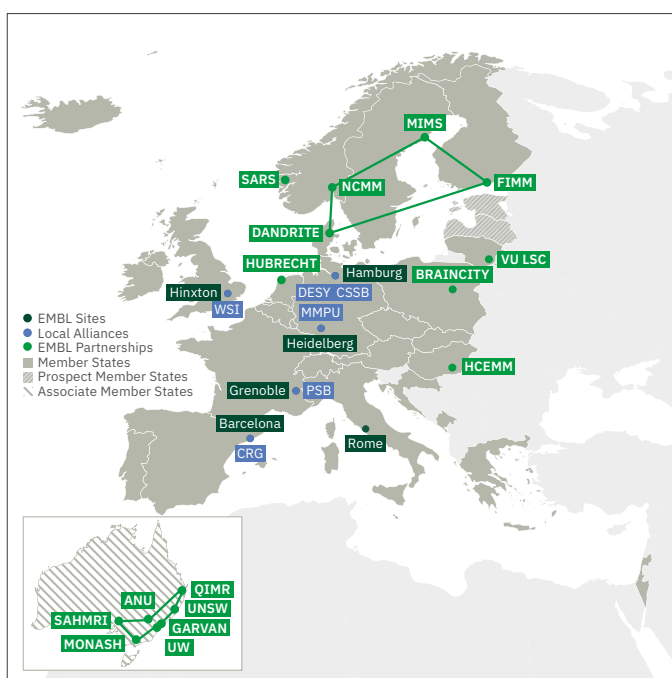


Figure IN5 | EMBL member states, partnerships, and local alliances.

EMBL works to establish links and promote collaborative relationships between EMBL and institutions in the EMBL member states, including EMBL's successful network of partnerships (green dots). The aim of these partnerships is to increase integration and participation of national scientific communities in EMBL's research and activities and also support member states to recruit excellent international talent. EMBL's local alliances (blue dots) represent close collaborations with EMBL's sites (dark green) and are instrumental in the creation of a critical mass of life sciences research within the local area.

Moving Towards the Next Frontier

Although the ambition to gain a molecular understanding of ecosystems may be considered bold, EMBL is well placed to expand its horizons. Over the past decade, EMBL has built up the tools and knowledge to begin to address the molecular basis of life in context, and is ready to do this in partnership with its member states and collaborators. EMBL is recognised for excellence in structural biology, genomics, developmental biology, cell biology, bioinformatics, and instrumentation (Figure IN6). The core strengths of EMBL, its expertise in studying life at multiple scales, its provision of biological data and key services, and its solid history in technology development, training, and innovation, make EMBL an ideal organisation to lead and coordinate new scientific enterprises to investigate life in the context of its environment.

EMBL researchers will build upon experience and models from previous research projects. EMBL has initiated and participated in collaborative expeditions with the Tara Oceans Consortium to monitor oceans around the world, revealing an extraordinary and previously unexplored biodiversity. Studying the impact of climate change on biodiversity is vitally important, given the role many ecosystems play in regulating climate. Researchers at EMBL have shown that the relative abundance of bacteria and fungi in terrestrial topsoil is critical for nutrient cycling and may also have a profound impact on the appearance of antibiotic resistance genes. Understanding the molecular basis of antibiotic resistance transmission in natural contexts is a key endeavour for EMBL scientists. Computational tools, including lightweight apps, have also been developed by EMBL scientists to enable genome-based surveillance of infectious diseases. Scientists across EMBL and its member states are carrying out a range of other research to study the emergence and spread of pathogens. This is highly relevant, given that biodiversity loss facilitates the emergence of new pathogens and frequently increases rates of transmission. Finally, in the context of human health, there is now a plethora of studies demonstrating correlations between gut microbiome composition, nutrition, and common disorders such as metabolic disease or cancer. The hope is that these correlations can be translated into therapies using mechanism-based approaches. Ultimately, **EMBL's future discoveries, made through a combination of research in the field and in the laboratory, will provide key mechanistic insights with potential applications.**

Given the scale and pioneering nature of the scientific research needed to tackle such challenges, success is only feasible with an international effort involving collaboration between world-leading scientific institutions. EMBL's unique international position enables it to coordinate and lead such efforts for the benefit of science and society, while also mediating joint community standards and promoting open science. Harmonised efforts will ensure informed and impactful national and international research and policymaking, especially concerning ecological issues.

The powerful combination of EMBL's experience, expertise, and position prime the organisation to lead this revolutionary endeavour in science. EMBL will remain a stronghold of molecular biology research and will utilise its strengths in this field to drive scientific discovery in new areas. Research in molecular biology will always be EMBL's focus and area of excellence. These strengths, paired with EMBL's collaborative, flexible, and inquiring scientific culture provide the ideal launch pad for the Molecules to Ecosystems Programme and create an ambitious new era of life science research in Europe.

Why Now? An Optimal Time for Investment

Measuring and understanding community dynamics, interactions including infections, the principles underlying phenotypic variation, and the responses of organisms to changing environments are challenges that have hardly been tackled at the molecular level.

Early attempts to measure and understand the reciprocal interactions between environment (nurture) and genotype (nature) to explain phenotypes were made by developmental biologist Conrad Waddington, following Darwin's explanation of evolution through heritable variation and natural selection, and Mendel's laws of genetics. For much of the 20th century, genetic determinism prevailed, and the impact of environmentally induced phenotypic variation of any kind was recognised, but largely ignored experimentally. This was due to the limited quality and quantity of biological data available, and to limited experimental possibilities at the time for perturbing living systems. The additional levels of complexity brought about by the integration of environmental factors at the molecular level have meant that much of molecular biology has focused on reducing environmental fluctuation to a strict minimum, in order to study the impact of genetic and stochastic variation on phenotypic variation.

Over recent decades, remarkable molecular insights have been made in model organisms at EMBL, from bacteria to animals, using defined lab conditions. These have paved the way for EMBL to tackle the urgent need to understand life in context. Emergent technologies are now being developed to collect measurements of ecosystem components at unprecedented volumes, from molecules to cells, organisms, populations, communities, and their chemical and physical environments. Exceptional advances in computational power, AI, and causal reasoning have also enabled the rigorous analysis and creative integration of these data. This tremendous technological progress in the life sciences, coupled with the capacity to gather and analyse data of greater scope, resolution, and quality than ever before, means that measurements of environmental context can be collected in systematic ways, allowing the integration of this new level of complexity into the study of biology. **With these technological and computational advances, Europe – via EMBL – will lead the development of a new way of performing molecular biology: the study of biological components in their natural context.**

Programme Roadmap

Building on EMBL's existing strengths, the 2022–2026 Programme adds several exciting new scientific areas within which life can be explored in a variety of contexts (Figure IN6). Molecular processes and mechanisms underlying responses to environmental changes can be studied at multiple biological scales: from exploring the molecular components inside a cell, to measurements of single cells and multicellular tissues, to whole organisms and studies of populations. In EMBL's new Programme, these different dimensions will be explored, with the selection of specific areas of focus based on their potential for scientific opportunity when coupled with current advances in experimental technologies and data sciences.

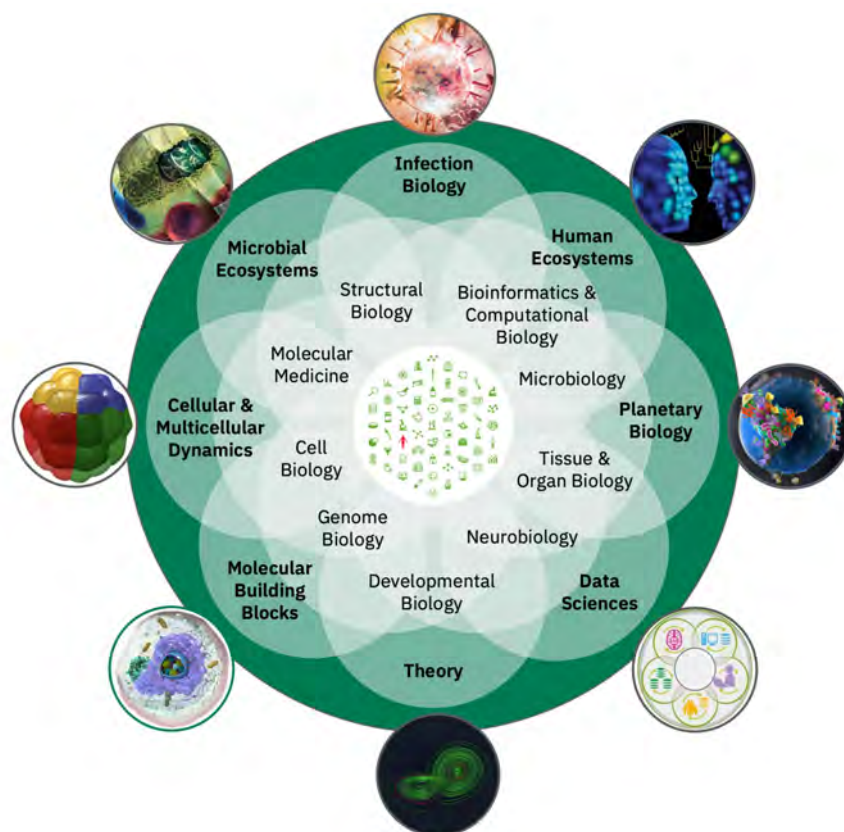


Figure IN6 | The research themes of the Molecules to Ecosystems Programme.

EMBL's current and new research themes are represented by the inner and outer rings, respectively. EMBL will build upon its existing strengths and expertise to conduct and enable collaborative, interdisciplinary research in these areas with scientists in EMBL's member states and beyond.

Section II begins with the first research theme of **Molecular Building Blocks in Context**, which delves into cellular function and subcellular components to determine systematically how responses to a changing environment are mediated at the molecular level. EMBL aims to understand mechanistically how these molecular responses translate into adaptations of cells, tissues, and organisms in different contexts. Understanding how cellular components and processes change over time, how they are interconnected, and how they feed back to one another, lies at the core of EMBL's expertise.

EMBL's approaches for gaining a mechanistic understanding of the genetic and environmental sources of variability in living systems, and understanding responsiveness at the cellular level and in a multicellular context, are highlighted in the theme **Cellular and Multicellular Dynamics of Life**. EMBL will use novel experimental strategies, cutting-edge technology developments, and predictive computer modelling to measure and perturb dynamic living systems and their interplay with the environment. Increasing knowledge about the robustness and plasticity of embryonic cell clusters, bioengineered tissues, and model systems will be essential in revealing the mechanisms that drive normal development and living processes, and the way these processes respond to disruptive environmental changes.

Microbial communities colonise, proliferate on, and impact every surface and subsurface of the planet, even in its most inhospitable corners. To better understand microbial ecosystems, their functional capacities, and their molecular interplay with the environment, the diverse microbial communities residing within the human gut are taken as an exemplar community in the **Microbial Ecosystems** research theme. EMBL aims to use novel computational and experimental methods to understand the functional diversity of individual microbial species and strains, as well as the interactions and properties of gut microbial communities within

the ecosystem of their human host. The ultimate goal is to be able to rationally modulate these microbial communities for the benefit of human and planetary health.

Infection Biology is an area that impacts humans and all life forms on Earth, with pathogens being able to cross species barriers, thereby adversely impacting biodiversity and human health. The current COVID-19 pandemic highlights the urgent need to obtain insight into the emergence and spread of infectious diseases. In the new Programme, EMBL will integrate multidisciplinary experimental and computational approaches to understand how pathogens and their hosts interact. These approaches will aid the development of diagnostic and surveillance tools to prevent the development and spread of antimicrobial resistance, and to work closely with frontline public health agencies to establish genome-based surveillance platforms. This has already begun with the provision of international data hubs for controlled data sharing, to empower scientists at EMBL and around the world to combat the COVID-19 pandemic.

In **Human Ecosystems**, EMBL researchers aim to understand how the environment impacts humans, both as individuals and within populations. A central question is how environmental factors can cause disease and how genotype and the environment influence human phenotypes. In the context of this theme, and in close collaboration with epidemiologists from member states, the environment will be studied through three distinct lenses focusing on the physical, biological, and social environments. Powerful computational, statistical, and experimental methods will address key questions that will bring a quantitative, mechanistic, and molecular understanding of environmental effects on humans.

Spanning multiple ecosystems, the **Planetary Biology** research theme will enable scientists to understand at the molecular, cellular, organismal, and population levels how microbes, algae, plants, and animals interact with each other and respond to natural and anthropogenic environmental changes. The main objectives will be to recognise and understand phenotypic changes that are environmentally induced in nature, using the plethora of tools available for molecular, structural, genomic, cellular, and developmental biology, and the powerful technologies that enable visualisation and perturbation of processes. TREC, EMBL's flagship project to explore European land–water interfaces including coastlines, rivers, and lakes, in partnership with scientists in the member states, is a central part of this theme. By working together and learning from one another, EMBL and collaborators will help to address fundamental and pressing scientific questions about the influence of environmental parameters on biological processes, while also addressing societal questions about the state of ecosystems.

All these research themes will contribute to the growing volume and heterogeneity of the biological and environmental data that are necessary for the study of life in context. To ensure these data are expertly generated, curated, annotated, managed, integrated, visualised, and shared, EMBL will launch a new **Data Sciences** programme, which will lie at the heart of EMBL's research strategy. As part of this strategy, a data science centre connecting all EMBL sites will provide support and training, facilitate research advances in data sciences including novel AI methods, set technical standards, and offer critical public data resources to the molecular biology community, with the overall goal of maximising the value of the generated data. Through these efforts, EMBL aims to be a role model for life science institutions that face similar data-driven challenges.

EMBL aims to create a new and highly integrated **Theory programme** to complement EMBL's research and data-driven methods for studying life in context. The complexity of biology necessitates theoretical approaches. This programme will build up approaches from first principles and will explain biological phenomena using mathematical formalism and models, turn data into understanding, and generate testable predictions. Conceptual theories will be developed and applied to answer specific questions from all six research themes. The interplay between theoretical and experimental research, complemented by a theoretical training programme and visiting theoreticians, will be an integral requirement for achieving EMBL's scientific goals.

EMBL's **Scientific Services** are set up to respond dynamically to the needs of research communities. EMBL's experimental and data services will be developed and integrated to form a central pillar of the new Programme:

- EMBL's cutting-edge technology development feeds into its **structural biology and imaging services**, which enable scientists to visualise molecules across scales. These services include robotically controlled beamlines that provide data on biological structures at the atomic level, and methods to integrate imaging by cryo-EM and light microscopes to show these molecules in their cellular context. By fully supporting the use of complex experimental apparatus, and by interfacing various scientific disciplines, EMBL services enable scientists from EMBL member states to access a range of structural biology and imaging techniques to answer complex biological questions.
- Advances in single-cell genomics and emerging developments in spatial omics will spur a range of new **multi-omics services**, based on new technologies.
- A wide array of **perturbation** approaches will be critical for mechanistic tests. EMBL's ***in vivo* gene editing service** will enhance the study of genetic variation in animal models, and will provide platforms for viral-mediated editing to offer insights into mechanisms *in vivo*.
- New cross-site **chemical biology services** will help scientists explore the effects of environmental factors and novel drug targets.
- In partnership with scientists in the member states, the provision of **mobile services**, spanning imaging, genomics, environmental measures, and data services will enable EMBL to further support research in its member states.
- EMBL's **biomolecular data services** will also see significant enhancements in the provision of reference data, standards, and tools, including bioimage data and human brain and behaviour data, as part of EMBL's data service repertoire. The **Genomic Medicine Platform** will engage with individual national initiatives, advising and proactively transferring technology to EMBL member states that have begun bringing precision medicine into their healthcare systems. EMBL will also provide data portals that can effectively coordinate new data types, which dynamically expand in size and relevance as research communities evolve.

EMBL **Training** activities will embrace the new research themes in providing state-of-the-art scientific training for EMBL fellows, including predoctoral and postdoctoral researchers. The Course and Conference Programme will also reflect the new themes from the EMBL Programme, and the Scientific Visitor Programme will increase the number of scientific visitors to its sites, by offering complementary sabbaticals and secondments. Training activities to strengthen capacity in EMBL member states will also be developed. With remote working becoming a way of life for scientists all over the world, EMBL will build on its success in providing accessible e-learning materials. This will enhance the impact and reach of EMBL's training activities, while also contributing to environmental initiatives at EMBL.

Innovation and Translation at EMBL will be expanded to encompass the translational potential of the new scientific directions in this Programme. In addition to building a portfolio of innovation and commercialisation activities, EMBL will broaden and advance research collaborations and technology development via new public–private partnerships. A range of new activities will be implemented to develop an EMBL innovation culture, empower the next generation of EMBL fellows, and diversify the current instruments for training and knowledge exchange between EMBL and industry partners.

EMBL's mission to **Integrate European Life Sciences** reaffirms its commitment to its member states and associate member states. EMBL will establish new links and initiate collaborative relationships between scientific communities in Europe, especially in the new scientific areas of the Programme. EMBL will foster additional and existing EMBL-modelled inter-institutional research partnerships across Europe, and will

develop a series of initiatives to promote closer collaboration and knowledge exchange. EMBL will continue its key European coordination activities with EIROforum and with the EC, including the EC-led project to establish the European Open Science Cloud (EOSC).

Section III outlines plans for EMBL's **People, Processes, and Places**, which will be pivotal for the implementation of the new Programme, with the establishment of transversal themes across all EMBL sites, in order to launch some of the new research themes. The cross-disciplinary themes will require the recruitment of skilled professionals from diverse disciplines, including engineers, mathematicians, data scientists, theoreticians, physicists, and chemists. The development of an employer branding strategy will support these recruitment efforts. Across EMBL, schemes for career development and the promotion of equality and diversity will also be strengthened. EMBL's processes and systems will need to support modern ways of working alongside expanded IT infrastructures. EMBL's goal is to enable the creation of sustainable campuses across all sites and to firmly embed green working policies and practices. Through local collaborations, partnerships, and engagements, each EMBL site will continue research activities and exchanges with local institutes, regions, and national initiatives.

Given the societal and environmental relevance of EMBL's new Programme, strong **Public Engagement, Communications, and Outreach** will be key. EMBL aims to raise the visibility of its science and technology to inspire, inform, and educate a range of audiences. It will do this by increasing local public engagement at all EMBL sites, multiplying communications activity through collaborations and partnerships, embarking on the TREC outreach initiative in the member states, supporting European teachers and young learners, and engaging with policymakers to improve evidence-based decision making. EMBL also plans to strengthen public engagement and communication skills among staff, and to help increase communications capacity in its member states.

Critical Success Factors

EMBL's history of groundbreaking scientific achievements and other successes comes from a unique blend of ambition, insistence on excellence, cooperation, and openness across disciplines, all of which contribute to a distinctive scientific culture. These factors will embody every part of the new EMBL Molecules to Ecosystems Programme and will be key to its success.

A Virtuous Circle of Novel Research, Technology Development, and Services

EMBL has longstanding expertise in technology development in molecular crystallography, microscopy, genomics, and microfluidics methods, with several important contributions and successful examples of commercialisation. This is fuelled by EMBL's unique capacity to host research groups in proximity to facilities for the creation of instrumentation and software. The collaborative spirit and shared desire to creatively solve problems promotes a virtuous circle (Figure IN7). Research questions drive technological developments that, in turn, drive further research. These technological developments can then be made available to the scientific community via EMBL scientific services. The broad range of research areas and the juxtaposition of curiosity-driven researchers and experienced engineers who are able to build new equipment, often from raw materials or individual components, makes EMBL unique among life science research organisations in Europe. The technology powerhouse that sustains this virtuous circle is essential to EMBL's continued success. EMBL has developed and uses an impressive range of technologies, and the following chapters refer to the application of several technology developments, showcased in **Technology Development Boxes**. In the new Molecules to Ecosystems Programme, many new and ambitious technologies and services are being proposed and will be shared with EMBL member states.

A Guide to Technologies at EMBL (Appendix I) provides brief descriptions of the technologies referred to in this Programme, covering the areas of structural biology and imaging, computational methods, and omics technologies, as well as the convergence of these fields with one another.

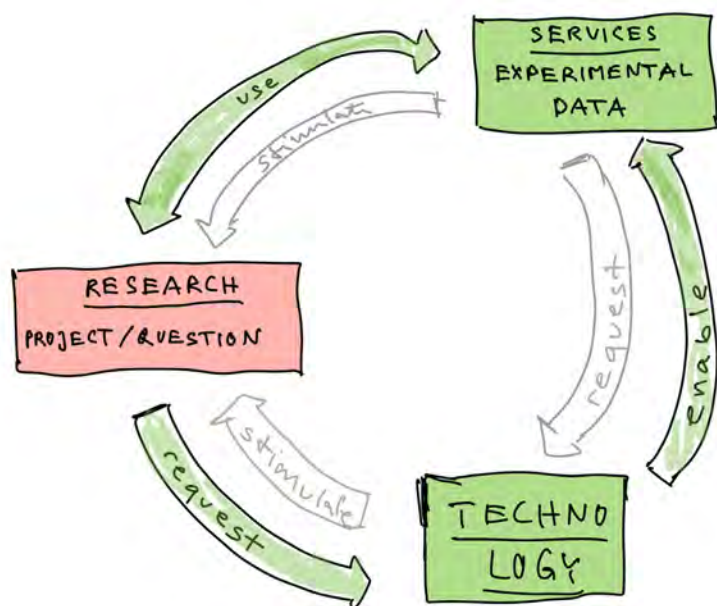


Figure IN7 | The EMBL research–service–technology virtuous circle.

Researchers can use EMBL experimental and data services to obtain answers otherwise not accessible to them. Research questions and service provision also drive novel technology developments. Conversely, novel technologies or services frequently trigger new research questions and projects.

Interconnection Between Experimental and Computational Science

EMBL’s computational research spans multiple areas and all EMBL units and sites. Over 50% of researcher time is spent on computational research, and there are dedicated computational group leaders present in nearly all units. This deep commitment to computational research is important, as much of the integration between research themes is achieved computationally, by analysing the results of well-designed experiments. EMBL’s research expertise is enabled by a clear view of the most important prevailing research questions, a view that benefits from EMBL’s programme of international courses and conferences. Key to EMBL’s future success will be maintaining this close alliance between experimental research, the development of new tailored mathematical algorithms, and AI-based approaches, and nurturing the next generation of scientists who combine experimental and computational expertise in their research.

Open Science

Open science is the movement to make scientific research accessible and transparent, and to remove barriers that may restrict the availability of research to privileged groups or to those prioritising financial gain. Motivated by a desire for science that is high quality, more efficient, and more useful for society at large, open science is promoted by a range of actors, including the EC, EMBL member states, funding bodies, and many scientists themselves.

This motivation and commitment to open science is shared by EMBL. The organisation has a special responsibility to be a leader and innovator in open science, not only in terms of the scientific results it produces, but also in the way research in molecular biology is performed. EMBL also provides scientific facilities and data services that are crucial for all European life science research, thus connecting Europe with global open

science infrastructures. Open science practices at EMBL comprise policies that cover all scientific outputs. This includes the deposition of data in open data resources, publishing preprints and accepted manuscripts to Europe PMC (EMBL's database of open access literature), and making open source software available. The way EMBL delivers open science will be a template for open science practices for life science organisations across Europe.

Interdisciplinary Research

As our understanding of the living world grows and new technologies for the life sciences emerge, there is an increasing need for interdisciplinarity. EMBL has long embraced the integration of researchers from a variety of biological disciplines, as well as physics, chemistry, computer science, and engineering science, into multidisciplinary teams in which traditional silos are broken down. EMBL has always combined groundbreaking curiosity-driven research with innovative technology development: an approach that has been made possible by the collaboration of experimental life scientists, technology developers, and data scientists at EMBL. The new Programme will require an even greater commitment to interdisciplinary collaborative research, with diverse areas such as **ecology, epidemiology, public health, zoology, toxicology, and theory**. Ultimately, a critical mass in many new disciplines will be needed at EMBL for the provision of new expertise and facilities. This will also foster connections with experts in other fields, whose work may benefit from interactions with molecular biology research and its methods.

Intensive Collaboration and Coordination

Collaborative research has become more prevalent globally, across all scientific disciplines, over the past 50 years. Tackling grand societal challenges such as climate change, food security, and threats to human health, as well as dealing with increasingly data-rich and computationally heavy projects, requires intensively collaborative work, which cannot be done by a single individual, laboratory, or company. The new themes of the Programme will build on EMBL's existing strengths and expertise, but will need to bridge disciplines more than ever before. EMBL can succeed in this endeavour thanks to in-depth **collaborations with scientists in EMBL member states** and around the world. Together, these collaborations will enable the development of unexplored concepts, novel tools, and innovative technologies. **EMBL aims to reach out across scientific domains and forge strong links with academics, governments, policymakers, and citizens to co-produce and co-create a new era of biology that will provide a molecular understanding of ecosystems.**

Sustainability Practices as Drivers of Green Research

Given EMBL's scientific commitment to the environment in its new Programme, sustainable practices are being strengthened across all of EMBL's sites, driven by a recent **Green EMBL** initiative. An Environmental Officer was recently appointed to ensure the implementation of these practices. The novel discoveries and tools expected from the new Molecules to Ecosystems Programme should make EMBL a model for research and services in ecologically relevant areas, as well as creating an environmentally friendly workplace. EMBL aims to be a leader in developing new ways to carry out environmentally conscious, responsible research practices.

In the longer term, EMBL aims to set up an Environmental Office to enable the realisation of high-impact global initiatives. By combining resources and expertise and working collaboratively across Europe, research in the new areas to study life in its natural context can stimulate the creation of new funding sources, help to commercialise findings, lead to highly cited publications, and generate outputs that have an impact on

environmental policy, clinical practice, or public health. Supported by trusted research data and critical analysis of validated results, EMBL may be able to target specific sectors of society and provide evidence that supports decisions to strengthen ecosystems through careful, positive change. Depending on the scientific data gained, EMBL could play a role in lobbying for changes in policy or practices. This might involve reducing pollution and thereby slowing detrimental changes to ecosystems. Alternatively, through collaborations with industry, research data may spark innovative and creative solutions to actively strengthen at-risk ecosystems.

Value and Impact

The knowledge and understanding gained from experimental molecular biology has always enabled further discoveries and applications. This knowledge benefits scientists and other members of society in multiple ways; to understand how the rich diversity of life on Earth works at the level of molecules, and to tackle societal challenges and develop solutions.

With this new Programme, EMBL will address fundamental questions about the impact of the environment on biological processes, while addressing societal questions about human and planetary health. Ultimately, this knowledge economy should enable a comprehensive understanding of the molecular and mechanistic basis of life in context, including an understanding of biodiversity and ecosystems. As well as answering scientific questions, EMBL aims to answer societal questions, including questions about the impact of humans on the environment, such as the effects of pollution, climate change, deforestation, and biodiversity collapse; the spread of antibiotic resistance; the emergence of epidemics from zoonotic diseases; the destruction of our soils and oceans; the collapse of natural ecosystems; and human health challenges linked to environmental factors.

EMBL's new Programme aims to push the life sciences into a new era that will greatly strengthen the bridge between biology and disciplines such as epidemiology, ecology, toxicology, zoology, population genetics, engineering, and mathematical theory. EMBL's unique international position enables it to **coordinate and lead** such efforts for the benefit of science and society, while also mediating joint community standards and promoting open science. EMBL will implement and deliver **large-scale collaborative projects** involving molecular profiling in various contexts. With EMBL's philosophy of open science, the framework for such research can be a model for other ecosystem molecular biology research initiatives in the future. Given the scale and pioneering nature of scientific research needed to tackle pressing environmental challenges, success is only feasible with an international effort involving collaboration between world-leading scientific institutions (Chapter 16: Value Proposition).

With the ambitious scientific directions set out in the new Programme, EMBL will strengthen European science, closely connecting our member states and providing **new services, technologies, and multidisciplinary expertise**, from molecules to ecosystems. EMBL will continue to lead the world in the provision of **open research data and standards**, in collaboration with the worldwide scientific community, and to develop the knowledge databases and tools that will enable the study of the complexity of life in the context of ecosystems.

Alongside these, EMBL aims to **train the next generation** of scientists, providing them with an awareness of planetary biology and the need to integrate environmental concepts. EMBL's various external training initiatives will make these opportunities accessible to scientists beyond EMBL, and will increase public awareness of scientific and societal challenges. As a result of EMBL's turnover model, when **EMBL personnel return to member states** their expertise can be leveraged by member state research organisations, which will benefit from having experienced personnel in unique and future-facing roles related to the Programme themes.

EMBL's scientific endeavours will integrate many scientific disciplines in an international, innovative, and interdisciplinary way. These, in turn, will facilitate new ways of bringing technologies to member states with the practical goal of **developing solutions** to mitigate the effects of environmental damage. Harmonised efforts will ensure informed and **impactful national and European policy and guidelines**, especially concerning ecological issues. EMBL's aim is to help Europe become the scientific leader in this new interdisciplinary area, to help both human and planetary health.

The powerful combination of EMBL's experience and expertise primes the organisation to lead this revolutionary scientific endeavour. EMBL's strengths, paired with its collaborative, flexible, and inquiring scientific culture, provide the ideal launch pad for the Molecules to Ecosystems Programme, creating an ambitious new era of life science research in Europe.

SECTION II

2. Molecular Building Blocks in Context

Background

Over the past half century, molecular biology has enabled a deep understanding of the mechanisms that underpin the propagation and evolution of living matter. EMBL has played a major part in many of the advances that have unravelled how the molecular building blocks of life (DNA, RNA and proteins) are used. For example, EMBL research has enabled the scientific community to make significant progress in sequencing and analysing genomes, solving protein structures, understanding processes such as gene regulation and body pattern formation, and deciphering the molecular machinery that viruses use to propagate themselves. These remarkable discoveries have revolutionised our understanding of life. Even so, they constitute only a small fraction of the molecular principles that underlie the diversity of the living world.

A central challenge in molecular biology today is the comprehension of genotype–phenotype relationships. This means understanding not just the contribution of genetic variation, which has been the main focus for molecular biologists to date, but also the environmental factors influencing phenotypes and how life responds and adapts to its environment. To enable life, it is now evident that genetically encoded information must on the one hand be integrated with environmental signals to maintain cellular integrity and activity, and on the other hand to permit adaptation. From unicellular to complex multicellular organisms, living systems must adapt to constantly changing contexts and environments. Most cellular and organismal functions exhibit an inherent regulatory flexibility, which enables them to adapt to external change. This capacity to sense and respond ensures homeostatic function, which is favoured by evolution. How the environment drives the adaptation of organisms and ultimately their evolution remains one of the biggest questions in biology, be this at the level of a microbial community evolving antibiotic resistance, of cells in a tumour responding to their microenvironment, or of changes in an organism’s morphology or behaviour due to the availability of food. Numerous examples of such phenotypic plasticity exist in nature, yet our molecular understanding of how they work and how they evolved remains rudimentary or non-existent.

While the impact of the environment on phenotypic outcome is often well described at the organism and population levels, the underlying molecular mechanisms remain relatively uncharacterised. In particular, the extent to which phenotypic variation is due to genetic variation and/or environmental influences has been difficult to assess. The major challenges today are to **determine systematically which environmental signals influence phenotypic variation, how different responses to a changing environment are mediated at the molecular level, and to understand mechanistically how these responses translate into conditionally adaptive or deleterious phenotypes.**

In this new Programme, EMBL researchers will leverage their strengths in molecular biology to explore the **molecular basis of life in context**. This means investigating how an organism integrates its genetically encoded information and intrinsic and extrinsic environmental signals to produce various molecules, macromolecular complexes, subcellular compartments, organelles, cell states, or cell types, which give rise to different phenotypes. Inferring changes in different contexts requires an understanding of how cellular components and processes change over time, how they are interconnected, and how they feed back on one another. This extends from understanding how the genome is differentially expressed in various environmental contexts, right through to how metabolites and gene networks are integrated during metabolic reprogramming of cells (Figure MO1). This knowledge is fundamental to our understanding of life in its natural environment.

In addition to studying how molecular processes are affected by acute changes, whether physiological (e.g. developmental, nutritional, hormonal), toxic (e.g. exposure to chemical pollutants, unusual temperatures), or pathological (e.g. infection), it will be necessary to interrogate how adaptive molecular changes propagate throughout an organism's lifetime, and the longer-term implications of this across generations. In EMBL's new programme, EMBL researchers will take longitudinal approaches to trace adaptive cellular functions in natural settings, in the lab, and upon perturbation. EMBL's current research applies state-of-the-art technical approaches to visualise and measure molecular and cellular heterogeneity, coupled with multiscale technologies to explore molecular processes. This will enable EMBL to tackle the challenge of dissecting how cellular functions adapt to changes in environmental context.

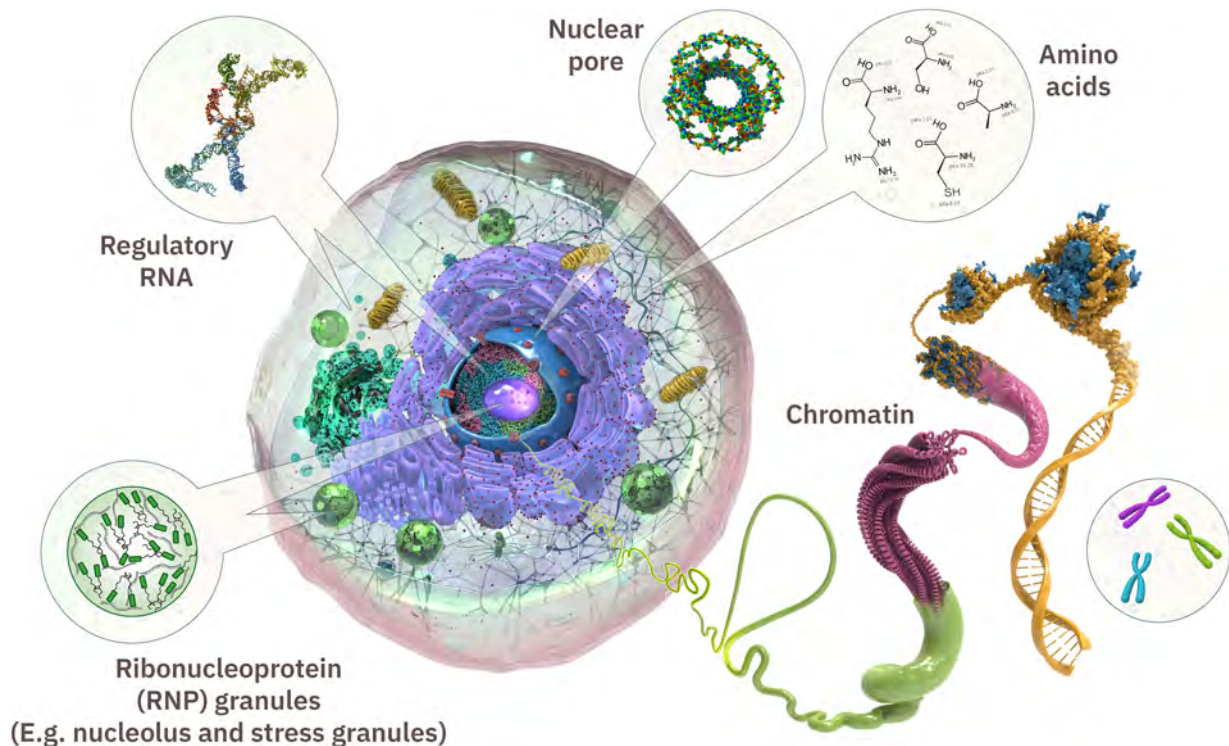


Figure MO1 | A cell contains a diverse range of molecular building blocks.

Illustration of a prototypic eukaryotic cell containing a selection of primary organelles, internal structures, compartments and molecules, which illustrate areas of ongoing research at EMBL that are relevant to the study of responses to biotic and abiotic environmental inputs.

The Opportunity

Building on EMBL's past and current research and strengths in technology development, and taking advantage of exciting new advances in structural biology, multi-omics, genetic engineering, imaging, and computational biology, EMBL can now investigate biological complexity in a way that takes account of environmental variation at the subcellular and molecular levels. EMBL researchers now have the tools to measure inherent variability within and between individuals (e.g. in gene expression), together with the impact of intrinsic signals (e.g. growth factors, hormones, metabolites) and various extrinsic environmental inputs or stresses (e.g. toxicity, nutrition).

At the level of DNA, dynamic protein binding, epigenetic modifications and mutations can be measured at the single cell level; at the level of RNAs, their dynamic processing, modifications, folding, and protein associations can now be captured and linked to physiological and pathological contexts; and, at the level of proteins, their production, folding and allosteric interactions, within macromolecular complexes or subcellular

compartments, such as stress granules, can be followed with a precision that was not even imaginable a few years ago. Furthermore, metabolic and signalling pathways can now be measured dynamically in living systems. These approaches will provide profound insights into the interplay between genetic variation and environmental variables at the molecular level, and will reveal how adaptation to rapidly changing environments is also achieved at the molecular level (e.g. genetic and epigenetic changes, RNA processing, translation alterations, and stoichiometric or allosteric changes in proteins).

Research Aims

In the new Programme, EMBL will strive to provide a deeper understanding of molecular and subcellular processes in the context of the cell, during development and disease, and ultimately in ecosystems. Countless research questions can be asked to investigate the impact of the environment on biological processes. For example, how is the same genome exploited to generate the diverse cell types that make up an organism? How does a single genotype give rise to multiple phenotypes after exposure to environmental signals in the context of phenotypic plasticity? How can exposure to the same environmental factors result in different responses from individuals with different genotypes? How do complex molecular machines orchestrate processes over the lifespan of a cell or an organism, or even across generations? How does metabolism change in response to changes in cellular context, and what effect does this have on subcellular functions? To explore these questions, EMBL will continue to investigate the molecular building blocks of life in different contexts, and will apply systematic and controlled alterations of the environment to study the impact on molecules, their modifications, and the macromolecular complexes they form, as well as on subcellular structures. EMBL aims to:

I. Understand and Predict Function from DNA sequence

The genome is the blueprint of life, but the environment (intrinsic or extrinsic) is key to shaping phenotype. Variations in genomic sequence, epigenetic states, 3D architecture, as well as protein and RNA functions, will be investigated with a view to understanding cellular heterogeneity and phenotypic variation. Specific aims will be:

- To define the physical and dynamical properties of the genome and the factors that influence its functions, particularly gene expression.
- To unravel the structure and function of transcription complexes and other nuclear machineries.
- To explore the many roles of chromatin and the degree to which epigenetic modifications can influence gene expression states.
- To study chromosome folding, its role in genome function, and the mechanisms underlying the folding process.
- To investigate regulatory RNAs at different levels.

II. Develop New Transformative Technologies and Methods

Crucial for the dissection of molecular mechanisms are the technological developments and conceptual models that will lead to a truly molecular understanding of the complex interactions between the environment and biological functions across scales. High-throughput molecular data from (meta)genomics, transcriptomics, and metabolomics, including at the single-cell level, are becoming routine. Following the revolution in imaging technologies and structural biology, molecules can now be visualised with unprecedented spatial and temporal resolution. These technologies must be combined with perturbation strategies (whether genetic, chemical,

or physical) to interfere with the spatio-temporal precision of biological processes. EMBL aims to **develop novel approaches to manipulate molecules and macromolecules**, particularly to enable combinations of approaches (e.g. optogenetics, dCas9, and single-cell omics) to perturb and measure molecules.

For a more mechanistic understanding of the complex interactions between environmental factors and subcellular machineries, molecular structures need to be studied *in situ* to understand how they perform their functions in distinct contexts. The challenge is to capture as much molecular information as possible *in situ* and across timescales while cellular and organismal processes (e.g. cell division, cell differentiation, or reproduction) are actually occurring. In this regard, EMBL aims to develop **new tools for dynamic *in situ* structural biology** to follow genomic information as well as molecular structures and functions over time and in response to intrinsic and extrinsic environmental cues. Machine learning and other AI approaches will be applied to make predictive models of the complex molecular processes that underlie changes in molecular and subcellular structures and functions in different *in vivo* contexts.

III. Understand Subcellular Function in Context

Subcellular systems have been extensively investigated *in vitro*. Many of the specific molecules and complexes responsible for cellular processes have been identified from biochemical purification and *in vitro* manipulations. EMBL scientists will now try to understand the molecules and the subcellular components that they are part of while they are carrying out their functions **in a cellular or *in vivo* context**. EMBL aims to explore several key components and processes to understand their roles in responding to environmental variation:

- **Nuclear organisation**, including nuclear transport, trafficking, and the structure and function of nuclear compartments.
- The **nucleolus** as a key player in sensing and responding to cellular stresses, and **Cajal bodies**, which appear to play a role in key RNA-related metabolic processes.
- Membrane-bound **organelles** such as the endoplasmic reticulum, the Golgi apparatus, vacuoles, lysosomes, and mitochondria play key roles in metabolic pathways and environmental responses.
- **Metabolites** as the functional read-out of cell physiological states and metabolism.
- **Liquid–liquid phase-separated compartments**, such as stress granules, nucleoli, and **intrinsically disordered proteins** (IDPs), which may mediate responses to different types of stress and extrinsic environmental factors.
- The molecular details of **symbiosis**, which may be key to many aspects of cellular function and responsiveness.

Ultimately, EMBL researchers need to understand how organisms evolve and adapt at the molecular level, in the face of environmental variability and in the context of ecosystems. This can only come from a deep understanding of the molecular processes involved. The combination of mechanistic studies, discovery research, and technology, underpinned by a strong interdisciplinary culture, are particular strengths of EMBL that will be leveraged to provide spatial and temporal insight into biological processes in individual cells, organisms, communities, and ecosystems.

EMBL's Approach

Understand and Predict Function From DNA

Even in organisms as diverse as flies, worms, plants, mice, and humans, the number of genes is very similar, ranging from around 15,000 to 30,000 in most cases. What is the basis of biological complexity, if it is not due to the number of genes? One component is the number of different ways the coding genome is utilised within an organism, for example to produce different protein variants through alternative splicing, and different activity states through post-translational modifications. There is also huge variation in the non-coding genome, with humans having 3 billion base pairs in their genome, compared to only 180 million in flies. This vast sea of non-coding DNA is made up in large part of the remnants of transposable elements (mobile genes), which over evolutionary time have provided many of the regulatory elements (e.g. enhancers) that instruct when and where genes should be expressed, as well as structural elements that help to organise the genome in three dimensions within the nucleus. It is becoming increasingly clear from human genome-wide association studies (GWAS) and quantitative trait locus (QTL) studies that most disease-associated variants are in the non-coding portion of the genome, impacting the function of regulatory elements, such as the binding of transcription factors to their target sites.

However, predicting the functional impact of a genetic variant (e.g. a single-nucleotide polymorphism (SNP) or indel) is still incredibly challenging and depends on multiple considerations, including the protein machineries that accompany transcription factors, as well as the chromatin state and 3D organisation of regulatory elements relative to the gene in question, all of which can change in different cell types, tissue types, and environmental contexts. Regulatory elements function in the context of chromatin and the three-dimensional organisation of the genome, folded at various scales, ranging from chromosomal territories to gene loops and nucleosomal arrays. Although each scale is being actively studied, including seminal contributions from EMBL researchers, **scales are typically studied in isolation with little integration across them.** To understand how the physical and molecular properties of the genome are utilised during processes such as transcription, DNA replication, and nuclear transport, a more holistic approach is required, integrating *in vitro* and *in situ* techniques with different resolutions. With cutting-edge expertise and new methods in genomics, cellular imaging, and structural biology, EMBL is particularly well suited to apply such an approach in the context of the next Programme.

The Physical and Dynamical Properties of the Genome

The dynamic, transient, and apparently stochastic way in which transcription factors interact with their cognate DNA binding sites makes **establishing the mechanistic links between transcription factor binding and transcriptional outputs** extremely challenging. This is confounded by the influence of protein–protein interactions, chromatin avidity and function, 3D nuclear ‘hubs’ of activity, and the emerging concept that nuclear factors can change their physical state and undergo phase separations to function. Transcriptional networks, including redundancy and feedback regulation, have clearly evolved to buffer the deleterious impact of genetic variation or environmental changes. Predicting function from sequence is therefore a huge challenge, and it is a pressing, open question as to how much can be predicted from sequence data alone.

EMBL researchers are tackling this challenge from multiple directions. Approaches from structural biology are being used to understand how transcription factors recognise DNA. This is complemented by high-resolution genomics approaches pioneered at EMBL, such as single-molecule footprinting. This method reveals which base pairs are occupied by transcription factors, and – importantly – which combinations of transcription

factors are bound to the same molecule of DNA at the same moment in time. The integration of these two approaches will facilitate more accurate modelling of DNA–protein interactions and how they are modulated in different contexts, and will form a basis for new machine learning approaches.

Transcription factors also bind to each other, to co-factors, and to multi-protein transcriptional complexes, such as the Mediator and SAGA complexes, forming very large protein–nucleic acid assemblages that are estimated to be in the megadalton range in some cases. Importantly, the makeup and interactions of such complexes are thought to vary dramatically at different stages of development or in the context of environmental fluctuations such as hormonal responses or nutritional conditions. Such complexes are now visible due to recent advances in single-particle cryo-EM and cryo-ET, which EMBL researchers are optimising to **visualise transcriptional complexes in the context of the nucleus**. Some of these large assemblages can form biomolecular condensates, and have the physical property of being able to phase separate *in vitro* and to form punctate areas of increased concentration *in vivo*. EMBL researchers are dissecting the molecular properties of phase separation using biophysical techniques (e.g. optical tweezers, atomic force microscopy), biochemical approaches (e.g. solubility mass spectrometry) and high-resolution fluorescent imaging approaches (e.g. fluorescence correlation spectroscopy, fluorescence recovery after photobleaching, or lattice light-sheet microscopy). The dynamics of these large condensates, or transcription factor ‘hubs’, is being quantified using live-imaging approaches, for example in the context of embryonic development. Such microenvironments contain localised concentrations of transcription factors and cofactors, and are thought to support robust transcription at enhancers that contain low-affinity sites. In the next EMBL Programme, researchers will test this hypothesis by exploring the function of microdomains and phase separation in gene expression or other genome functions (e.g. DNA repair), by integrating genetic engineering and optogenetics with cutting-edge lattice light-sheet microscopy and mass spectrometry.

Transcription Complexes

Eukaryotic transcription initiation requires the recruitment of DNA-dependent RNA polymerases to the transcription start site to form very large, multi-subunit transcription initiation complexes. EMBL researchers have contributed to unravelling the molecular mechanisms of eukaryotic transcription initiation by determining structures of RNA polymerase I and III, and providing molecular insights into how viral RNA polymerases – in particular influenza RNA polymerase – interact with and hijack the eukaryotic transcription machinery to transcribe their own genes (Chapter 5: Infection Biology). The next challenge is to **locate eukaryotic and viral RNA polymerases in their cellular context and study their function** using a combination of high-resolution imaging approaches (e.g. *in situ* cryo-ET, correlative light and electron microscopy (CLEM), and super-resolution microscopy) and complementary genomics approaches. Using such an integrated approach, EMBL researchers will also study how chromatin modifiers and remodellers affect chromatin architecture and contribute to gene activation or repression. Using structures of nucleosome remodellers bound to single nucleosomes (such as the INO80–nucleosome complex) as starting points, the interaction of remodellers with the chromosomal landscape, for example around gene promoter regions, will be explored first in reconstituted systems, and subsequently *in vivo*, in the nucleus.

Chromatin States and Functions

Alongside transcription factors, chromatin plays a key role in regulating gene activity and helps to maintain expression states during development, conferring an epigenetic cellular memory. Epigenetic mechanisms can also respond dynamically to external cues to guide genome outputs. Chromatin-based systems can thus represent molecular mediators of cellular responses to intrinsic and extrinsic signals. The future challenge

is to determine whether epigenetic information can be instructive in creating specific gene expression states, and to understand how epigenetic information interacts with genetic information, particularly in the face of environmental change. EMBL scientists are utilising multimodal synthetic approaches to model the functional consequence of precise chromatin perturbations. For example, by exploiting epigenome editing and chemical approaches as perturbation tools, the hierarchy and causal function of chromatin changes will be investigated and used to understand the extent to which chromatin forms an allosteric network of interactions that encode regulatory information. This has direct implications for understanding cellular plasticity and information processing within biological systems, and will make it possible to test hypotheses about the direct functionality of chromatin and the extent to which it acts as an effector of past and present cellular context. These strategies are complemented by approaches to understand allosteric networks using whole-genome reconstitutions in conjunction with cryo-EM, integrating the wealth of sequencing-based data with atomic resolution information from structural studies. EMBL scientists will also develop tools to probe next-generation molecular dynamics, allowing allosteric communication of factors embedded in native chromatin environments to be studied.

One pioneering project aims to systematically determine the functional role of chromatin-based information *in vivo* by generating an extensive resource of precision-edited mice programmed with specific functional mutations in a large cohort of genes involved in epigenetic processes, which are of particular interest in the context of responsiveness to environmental change (Chapter 10: Scientific Services). This exploits EMBL's optimised pipeline for generating mammalian CRISPR–Cas9-edited genomes, which is available to scientists within all EMBL member states. This pan-EMBL project enables researchers to move beyond conventional loss-of-function studies, which typically obscure whether molecular consequences are due to the absence of a protein or complex or to its enzymatic activity *per se*, to provide a deeper mechanistic understanding of the contextual function of chromatin states *in situ*. EMBL has a longstanding track record of pioneering studies in multi-omics platforms, and will integrate these platforms with emerging areas of mathematical image processing. The engineered mice and reagents will form a framework platform to further elucidate the interactions between epigenetic and genetic function in specific cell types and cell contexts, and will be freely accessible to the scientific community as an extensive resource.

Chromosome Folding

The genome is non-randomly organised within the nucleus. Genome folding can create proximity between genes and regulatory elements, leading to the accumulation of regulatory factors and the formation of local microenvironments. Short-range chromatin looping occurs in the context of larger chromatin domains, which operate in chromosome territories. Understanding the roles of the complex and dynamic organisation of the genome for its various functions, and the **interplay between DNA sequence variation, genome folding, and gene expression**, will be key to understanding phenotypic variation and DNA damage responses in the context of environmental change.

Recent years have seen a true revolution in the understanding of higher-order genomic structure, due to the development of new molecular approaches such as chromosome conformation capture (3C and genome-wide Hi-C) and advances in super-resolution microscopy. Recent structural, perturbation, and imaging approaches have provided profound insights into the macromolecular structures involved in genome folding in well-studied model organisms such as humans, mice, flies, yeast, and bacteria. Examples of such pioneering research from EMBL include solving the structure of chromosomal replication domains *in vivo*; elucidating the dynamics of mitotic and meiotic chromosome formation; understanding the structure and timing of chromatin domains during early mouse and *Drosophila* embryogenesis; dissecting the functional roles of chromatin topology during X-inactivation, and of chromatin looping during developmental enhancer–promoter communication; and the first demonstration of DNA loop extrusion by condensin. In the next

Programme, EMBL researchers from diverse disciplines will collaborate to address longstanding fundamental questions related to chromosome organisation and function. High-resolution imaging (e.g. cryo-ET, CLEM, super-resolution light microscopy) and complementary single-locus localisation techniques (e.g. OligoDNA-PAINT, DNA FISH) will be combined with genomics approaches to gain structural insights into specific chromosomal loci that are well characterised at the genomic level.

Such an interdisciplinary approach will be used to elucidate the native structure of activated and repressed genes, the molecular mechanisms controlling 3D organisation and gene expression changes, and their interplay during epigenetic processes such as X-inactivation, as well as the nature of interactions between enhancers and promoters. The interactions of promoters with their cognate enhancers and, more generally, between remote chromatin sites will be studied in the context of embryonic development using model systems. Building on seminal contributions from EMBL researchers in elucidating when, where, and how these contacts are first made, future directions will focus on their real-time kinetics and physical properties, and their regulation and functional role in transcriptional initiation. Whole-genome reconstitutions in conjunction with cryo-EM will enable EMBL researchers to link the wealth of sequencing-based chromatin folding data with atomic-resolution information. This integrated approach will be combined with perturbations to the system, either mutants in *cis* or depletion of proteins in *trans*, to determine the contribution of specific chromatin modifiers to changes at the structural and transcriptional level.

Regulatory RNAs

The functional consequence of transcription is to produce RNA, but there is accumulating evidence that RNA plays diverse roles in the regulation of transcription. Not all RNAs are translated – non-coding RNAs (ncRNAs) play fundamental roles in the regulation of chromatin structure, epigenetic processes, and translation. EMBL scientists recently discovered that a huge repertoire of proteins bind to RNA, and that this RNA binding may have a regulatory role in altering the functional state of the protein in many cases. Many of these interactions are through non-conventional RNA binding domains, and may involve intrinsically disordered regions. Future research at EMBL will include the study of the **functional role of these new RNA–protein interactions in the context of transcriptional regulation in the nucleus and regulatory processes occurring in the cytoplasm**, including autophagy and cellular metabolism. Emerging evidence suggests that ncRNAs are also involved in mediating a cell's ability to respond to environmental perturbations. This may be mechanistically linked to their role in epigenetic processes such as gene silencing, by the regulation of chromatin structure and DNA methylation. EMBL researchers study ncRNAs in the context of Polycomb-dependent gene silencing, X-inactivation, embryonic development, p53-dependent stress response, and response to viral infections. Solving the 3D structure of long non-coding RNAs will provide key information for understanding the functional mechanisms of this unexplored group of non-coding transcripts (Chapter 5: Infection Biology).

In addition to DNA and chromatin epigenetic modifications, RNA is also post-transcriptionally modified, and recent studies indicate that there are hundreds of different chemical modifications in various RNA species throughout all domains of life. RNA modifications thereby add an additional, potentially reversible, and dynamic layer of information that can fine-tune the functional properties of RNA in response to environmental cues. To understand the principles of **RNA editing and its functional significance**, researchers at EMBL are using crystallography and single-particle cryo-EM to study the structure–function relationships of complexes involved in RNA editing and processing. They are also using genomics-based approaches to uncover all genome-wide RNA modifications and perturb them through genome engineering. RNA editing complexes such as APOBEC1 will also be exploited using synthetic biology approaches for ‘transcriptome editing’, akin to CRISPR–Cas systems, which is one exciting avenue EMBL researchers are exploring.

Develop New Transformative Technologies and Methods

Novel Approaches to Manipulate Molecules and Macromolecules

Many tools are available to manipulate genomes, RNA, or proteins under different conditions and to visualise and measure the impact of these changes both *in vitro* and *in vivo*. Manipulation of the genome and other molecules can be achieved with unprecedented precision and ease, thanks to CRISPR–Cas9, transcription activator-like effector nucleases (TALEN), and other technologies. EMBL researchers can explore higher-order genomic structures thanks to super-resolution microscopy and molecular approaches such as DNA FISH and chromosome conformation capture (3C and genome-wide Hi-C). These technologies can be applied at the single-cell level and provide spatio-temporal and allele-specific information on genome dynamics. Researchers can genetically engineer genomic tags for imaging or barcoding DNA, RNA, or proteins in cells or whole organisms. Introduction of precise genetic variants, and of DNA sequences that enable the inducible expression (drug-induced) or deletion (LOX-, viral-, or CRISPR-mediated) of a gene, or the degradation of a protein (via a degron) are possible. Optogenetic approaches allow light-mediated manipulation of proteins *in vivo*, often translocating a protein from one subcellular compartment to another (Tech Dev Box TD1_MD). Furthermore, the integration of high-throughput data in (meta)genomics, transcriptomics, and metabolomics facilitates the move from structure to function, and enables biological discoveries and the modelling of processes in various areas of biology. EMBL's unique strength in **technology development** will allow the generation of the tools needed to drive discovery.

New Tools for Dynamic *In Situ* Structural Biology

EMBL will continue to develop the tools that enable dynamic structural cell biology *in situ*, providing scientists with an understanding of how molecular structures perform their functions in distinct contexts. This will include new correlative methods and approaches for dynamic super-resolution microscopy that can be applied to larger, more physiological specimens. New technologies are also needed to image the physical and chemical properties of biological matter and its environment, including forces, tension, density, and ionic strength, and EMBL needs to combine measurements with quantitative sampling of dynamic 3D cellular morphology. For many biological models, penetrating deep into the tissue is key to studying cells in their native environment, so EMBL researchers will continue to advance deep tissue imaging with subcellular resolution and molecular and physical readouts (Tech Dev Box TD2_MD). Many molecular processes happen on very short timescales that cannot currently be sampled. High-speed live imaging with a time resolution on the order of nanoseconds will open a completely new world, allowing scientists to watch molecular mechanisms unfold in live cells. The next generation of detectors, required to make this a reality, is now starting to emerge. Underpinning many of these imaging technologies are new chemical biology approaches to label and highlight molecules, their modifications, and activities. EMBL will also continue to build tools to bridge spatial scales. EMBL has world-leading instrumentation and tools for crystal harvesting, along with advanced synchrotron X-ray technologies. These also include integration of imaging approaches from *in vitro* molecular imaging by nuclear magnetic resonance (NMR), X-ray crystallography, and single-particle cryo-electron microscopy (cryo-EM); and *in situ* imaging using cryo-electron tomography (cryo-ET) of macromolecular complexes, subcellular structures, cells, and organisms.

The technologies applied will generate vast amounts of various data types, which will need to be integrated into predictive models that can be tested and refined (Chapter 3: Cellular and Multicellular Dynamics). To process multiple biological data types on a large scale, a variety of computational approaches will be applied, including molecular dynamics, dynamical simulations, advanced data science, and machine learning. These models will be further refined and validated through an iterative cycle of experimental perturbations and

modelling to reach accurate, predictive models at different scales. By integrating models, a comprehensive atlas of cellular processes can be established with the goal of understanding the cellular basis of life. The Theory at EMBL theme will be particularly relevant for developing the conceptual frameworks that will underpin these models (Chapter 9: Theory at EMBL).

Understand Subcellular Function in Context

Eukaryotic cells contain subcompartments and organelles that carry out distinct functions. Organelles are generally surrounded by membranes and contain specialised transport systems for the exchange of macromolecules and metabolites between the organelle and the cytosol. The nucleus is the largest organelle in eukaryotic cells and, despite intensive research, is in many ways still poorly understood. Multiple cellular compartments also exist throughout the cell, which facilitate spatio-temporal regulation of biological reactions but do not possess membranes. These compartments include stress granules in the cytoplasm and nucleoli, and Cajal bodies in the nucleus. Understanding the functional role of nuclear architecture and the dynamic structure of nuclear membraneless organelles will be major topics in EMBL's next Programme.

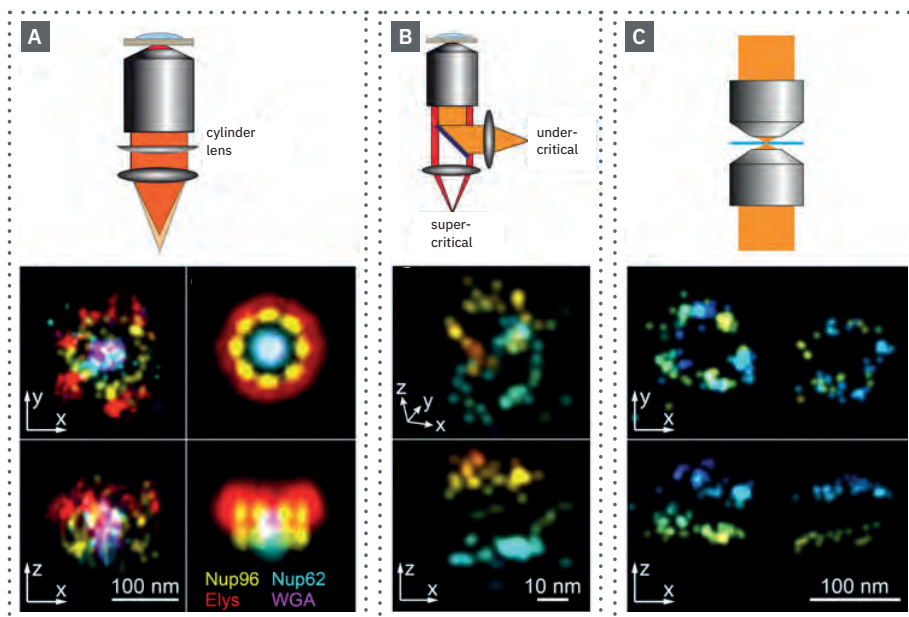
Nuclear Organisation

The nucleus not only contains the genome of the cell, but is also the environment within which several major cellular processes must be orchestrated. These include transcription, DNA replication, chromosome assembly in preparation for cell division, and DNA repair in response to damage. The dynamic organisation of the nucleus is currently a major research focus at EMBL, and several insights into the architecture and dynamics of very large molecular assemblies in their nuclear context have been made. A recent example has been the collaborative work of EMBL researchers to analyse the architecture and dynamics of the nuclear pore complex (NPC), which spans the nuclear membrane and coordinates the exchange of macromolecules between the nucleus and cytoplasm. The elucidation at EMBL of the NPC structure in humans and in an alga, using structural biology techniques, super-resolution microscopy (Tech Dev Box TD1_MO), and modelling, has revealed its intricate molecular structure, with more than 1,000 proteins in the human NPC. EMBL researchers are now exploiting *in vivo* systems to investigate the dynamic assembly and turnover of the NPC in context. Using yeast genetics and *in situ* structural biology, an essential role for a subset of nucleoporins (the protein building blocks of the NPC) in both mRNA export and NPC turnover has recently been defined.

While detailed insights into the structure, dynamics, and function of the nuclear pore at the periphery of the nucleus are emerging, a full understanding of **how multimolecular complexes contribute to shaping the functional architecture of the nucleus** has not yet been achieved. From genomics studies, insights into the molecular profiles of specific chromosomal loci and their changes between active and inactive states are available. In addition, key molecular players in the nucleus and their activities have been defined, and nuclear architecture has been studied extensively at various scales. However, very little is understood about how the **different levels of nuclear architecture affect various functional processes, how functions are controlled across space and time, and how the environment influences function**, including DNA replication, transcription, RNA processing, and nuclear transport. By investigating the interfaces between these processes using a combination of *in vitro* and *in situ* techniques across scales, major breakthroughs in understanding structure–function–environment relationships can be made.

Technology Development Box TD1_MO | 3D Super-resolution microscopy.

Single-molecule localisation-based super-resolution microscopy (SMLM) reaches nanometre resolution and can provide structural insights into cell biological questions. However, such high resolution has been limited to 2D measurements. To enable 3D super-resolution imaging with structural resolution, the Ries Group is developing three complementary approaches: **(A)** A new data analysis workflow that takes into account specific optical properties of the microscope allows the extraction of precise 3D positions from the shape of the single-molecule images. This approach works in up to four colours simultaneously, can capture entire cells, and is compatible with automated high-throughput SMLM. **(B)** The group is developing ‘supercritical-angle localisation microscopy’ in which the near-field emission of fluorophores is evaluated to obtain a z-resolution of a few nanometres in the vicinity of a glass cover slip. **(C)** 4Pi-SMLM is being implemented collaboratively and interferometry is being used to obtain nanometre isotropic resolution, including in thick samples. These technologies allow the Ries group and collaborators to extract precise positions of proteins in complex molecular machines in all three dimensions, which cannot be achieved with other imaging methods. This will help to unravel the structural arrangement of the proteins driving clathrin-mediated endocytosis and, in collaboration with the Ellenberg group, the structure of the nuclear pore complex.



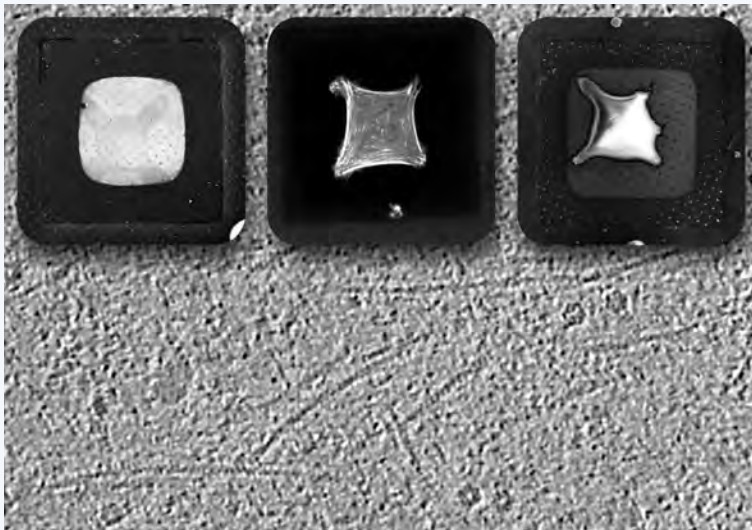
Nucleolus and Cajal Bodies

The nucleolus plays a key role in sensing and responding to cellular stresses such as hypoxia, pH fluctuations, and DNA damage, as it pursues the energy-intensive process of ribosome biogenesis. Inhibition of RNA polymerase I transcription is the first step of ribosome assembly, and leads to changes in nucleolus morphology. **What regulates the shape and size of nucleoli and their massive reorganisation during environmental perturbations like stress? What are the functional consequences of these morphological changes?** To address these questions, it is imperative to connect live imaging data to structural data. To do this, EMBL researchers are developing a framework to locate the three nucleolar subcompartments that specialise in the various steps of ribosome assembly, so they can track proteins by live imaging. Subsequently, macromolecular complexes (e.g. pre-ribosomes or transcription complexes) will be visualised *in situ* by cryo-EM and CLEM (Tech Dev Box TD2_MO; Tech Dev Box TD_SS2), and their molecular contexts will be probed by proximity labelling.

Cajal bodies have been implicated in RNA-related metabolic processes such as biogenesis, histone mRNA processing, and recycling of splicing small nuclear ribonucleoproteins (snRNPs). Cajal bodies also contain large protein complexes involved in the expression of genetic information, such as the Integrator complex. Paralleling work on the nucleolus, EMBL researchers aim to understand the *in situ* structure of macromolecular complexes such as Integrator, and how the processes leading to snRNP assembly are linked to the subcellular architecture of Cajal bodies. Ultimately, this approach aims to provide a **comprehensive and dynamic picture of pre-mRNA processing machineries at the multiscale level** as they carry out their functions in their native cellular environment. Finally, as both nucleolus and Cajal bodies are membraneless, phase-separated organelles, the extent to which phase separation contributes to and influences the molecular functions of each compartment will be explored.

Technology Development Box TD2_MO | Photo-micropatterning for cryo-EM.

The Mahamid Group, together with collaborators in France, have developed a new technique called photo-micropatterning for applications in molecular-resolution cryo-EM of intact cells. Photo-micropatterning allows spatially controlled cell adhesion and the manipulation of cell shapes on cryo-EM grids. This technology overcomes one of the technical challenges presented by cryo-ET, whereby only cells that are positioned in the centre of a grid square are available for processing and therefore imaging. Micropatterning enables the position of cells on the grid to be controlled with a high degree of spatial accuracy, increasing the number of cells available for imaging from only a few to about 30. Micropatterning can also be used to manipulate the shape of cells and study their mechanical behaviour. What's more, numerous different patterns can be generated on the same grid, allowing direct comparison of different intracellular architectures, including the cytoskeleton, the nucleus, or the Golgi apparatus. By understanding the three-dimensional architecture of macromolecules, the collective behaviours that give rise to new mechanical



properties can be explained. For example, in the image below, cryo-ET of an RPE1 cell (background image) reveals branching actin filaments and hexameric densities related to cell adhesion. The insets show a cell grown on a cross-shaped micropattern. This technical advance will help to streamline the cryo-EM pipeline and facilitate automation of the process. It's a significant bridge between structural and cell biology that will enable this technology to become a routine method in the future.

Membrane Trafficking and Organelle Biosynthesis

Beyond the nucleus, the eukaryotic cell contains many membrane-bound organelles, including the endoplasmic reticulum, the Golgi apparatus, vacuoles, lysosomes, and mitochondria (Figure MO1). The basic processes involved in membrane trafficking and organelle biogenesis in the secretory pathway, such as vesicle budding, fusion, or transport, are well defined. **How these processes are integrated and give rise to the size and shape of specific organelles and how this relates to diseases such as cystic fibrosis, lung fibrosis, or cardiovascular disease-related cholesterol homeostasis** remain major questions in the field. EMBL researchers use and continue to develop live-cell imaging and systematic genetics on a systemic scale to achieve an understanding of these processes. Quantitative data can then be used to test mechanistic models. This has already revealed a number of interesting factors that respond to stimuli such as cellular cholesterol depletion, growth factor-induced growth control, DNA damage, or cell differentiation, by changing the localisation of membrane proteins or membrane traffic-associated proteins from the cytoplasm to the nucleus or vice versa.

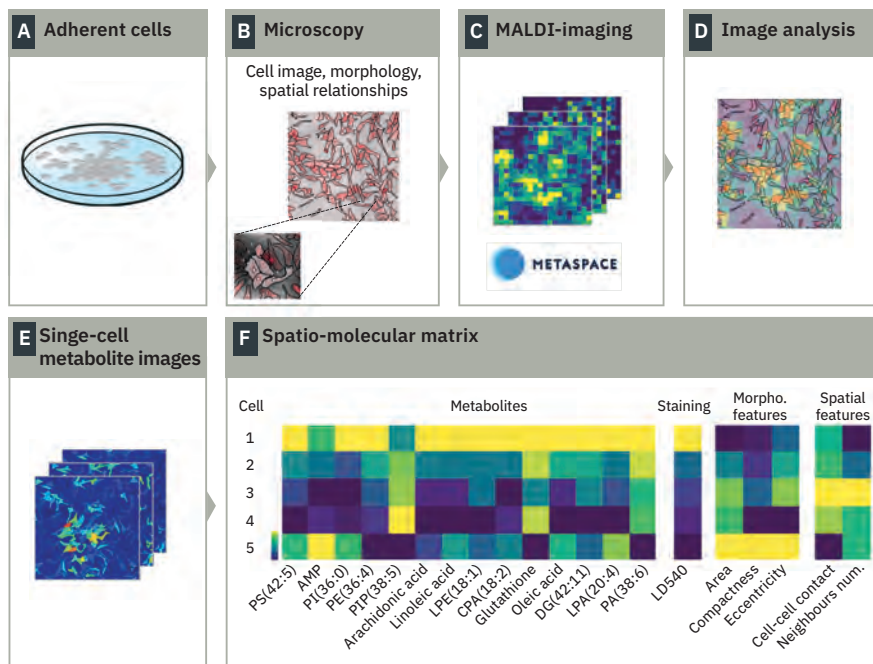
Metabolites

The metabolic profile of a cell provides a functional read-out of its physiological state. Cells and tissues have unique metabolic fingerprints that provide organ or tissue-specific information. This also holds true for many disease states. The subcellular distribution and function of metabolism and metabolites are still not completely known, and many mechanistic links await discovery. Metabolism also lies at the core of organismal survival in the environment. In the next Programme, EMBL will study the **links between intracellular metabolomes, cellular phenotypes, and spatial organisation of cells**, and the influence of the environment on metabolic profiles at the subcellular, cellular, tissue, and organ levels will be explored.

To map metabolic pathways, EMBL researchers will exploit thermal proteome profiling (TPP). The power of this technology is that it makes it possible to detect the effects of metabolites on proteins that manifest in changes in protein properties, and hence opens a new window on allostery. EMBL will also use its in-house expertise in spatial metabolomics, based on MALDI imaging mass spectrometry, to investigate the metabolome across scales, from single cells and tissues to whole organisms. This technology enables the detection of metabolites including amino acids, lipids, fatty acids, and the products of glutaminolysis and glycolysis. Many recent studies have explored the adaptability of metabolic pathways in animals based on comparative analyses of networks aided by theoretical and computational predictions. It should now be possible to integrate this comprehensive background knowledge with direct experimental observations. At EMBL, MALDI imaging mass spectrometry (Tech Dev Box TD3_MO) will be used for the first time for high-throughput metabolic screening of thousands of *Drosophila* embryos to characterise how their metabolism is reprogrammed upon genetic and environmental perturbations. This approach will allow the network structure at genomic and population-wide scales to be studied. Metabolic profiles and changes will be detected and linked directly to the uptake of exogenous, environmentally critical molecules such as pollutants, herbicides, and unwanted drugs. This will serve as a model to evaluate how organisms, including humans, are affected by the molecules in their environment. The high-throughput nature of this technology will allow EMBL scientists to set up screens for thousands of exogenous molecules that pose risks to humans.

Technology Development Box TD3_MO | Spatial single-cell metabolomics (SpaceM).

Measuring levels of metabolites in single cells is crucial for understanding metabolism, its heterogeneity, and its links to cellular phenomena and to cellular and transcriptional programs. However, conventional metabolomics is only feasible for bulk analysis where tissues need to be homogenised. Recently, imaging mass spectrometry (MS) was shown to be a successful method for acquiring spatially resolved metabolic fingerprints from tissues and cell cultures. These fingerprints can be measured directly from each probed location with high molecular specificity and sensitivity, and with a spatial resolution as small as 10 μm . The Alexandrov Team has a particular focus on furthering the development of novel imaging MS methods by using **matrix-assisted laser desorption/ionisation (MALDI) MS**. The team has recently developed a workflow for spatial single-cell metabolomics and lipidomics by co-registering microscopy and MALDI imaging MS data. **(A)** Cultured cells are **(B)** first imaged using bright-field and fluorescence microscopy. **(C)** MALDI-MS is performed in a raster pattern, and **(D)** the resulting laser ablation marks are imaged and analysed separately. **(E)** The microscope image is integrated with the MS data, and metabolic profiles are normalised based on the overlap of laser ablation marks with cells. **(F)** The resulting spatio-molecular matrix integrates information about metabolism and phenotype of the individual cells and enables single-cell metabolic investigations of cell types, cell states, and cell–cell interactions.



Signalling from the Environment to Cellular Compartments

Liquid–liquid phase-separated compartments, such as stress granules, nucleoli, and Cajal bodies, are increasingly being recognised as dynamic entities that may mediate responses to external factors. Their highly dynamic nature and the fact that the components within them are in constant exchange with the surrounding cytoplasm or nucleoplasm may allow a cell to rapidly respond to its environment. Consistent with this view, the organisation of the cytoplasm has been shown to change considerably in response to environmental stimuli such as stress, with many proteins and RNAs being sequestered to phase-separated structures. EMBL researchers have recently shown that, upon food deprivation, P-bodies containing certain

RNAs and proteins form in the ovarian germline cells of *Drosophila* during oogenesis. When this happens, protein synthesis and egg production halt, indicating environmental sensitivity (Figure MO2). Future studies will investigate how the properties of liquid–liquid phase-separated entities relate to their dynamic biological functions in response to their environmental context and the underlying mechanisms involved.

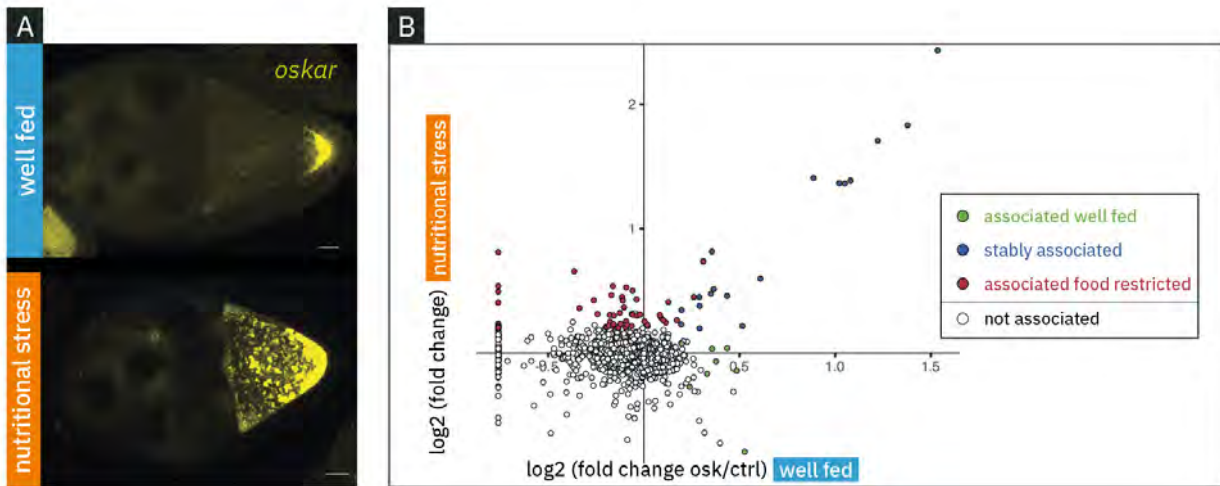


Figure MO2 | Changes in RNA-associated proteome upon nutritional deprivation.

(A) Fluorescent *in situ* hybridisation showing the distribution of *oskar* mRNA in an egg chamber of a fruit fly, either well fed or subjected to nutritional stress. Scale bar 10 μ m. (B) Analysis of the changes in mRNP composition upon nutritional stress. *oskar* mRNPs from well-fed flies and flies that were deprived of protein-rich food for 4.5 hours were analysed by quantitative mass spectrometry.

EMBL will address one aspect of this by interrogating **intrinsically disordered proteins** (IDPs), which are implicated in a wide range of cellular functions, including the generation of membraneless organelles formed by liquid–liquid phase separation. This phenomenon is directly related to the intrinsic properties of IDPs, and has far-reaching consequences for cellular regulation that are as yet barely explored. IDPs are able to gain functional advantages by remaining natively unstructured, either completely or partially. EMBL is well positioned to contribute significantly to this area with its existing research and technologies, including investigating the relationship of IDPs to one another, their biological significance, and their response to environmental cues, especially in organisms in different environments as part of projects under the Planetary Biology theme (Chapter 7: Planetary Biology). EMBL aims to use a variety of approaches to characterise IDPs, such as small-angle X-ray scattering (SAXS), fluorescence resonance energy transfer (FRET), various NMR-related techniques (paramagnetic relaxation enhancements, residual dipolar couplings, chemical shift perturbations, and relaxation analysis). Cellular cryo-electron tomography yields *in situ* information across various scales on both assemblies and individual macromolecules. In-house expertise in TPP approaches will also form an integral part of EMBL's strategic approach. EMBL researchers will use TPP to explore protein–protein solubility and phase separation properties on a genome-wide scale, providing insights into the roles of enzymes, metabolites, post-translational modifications, and phase transitions in the cell under changing environmental conditions. Bioinformatics and sequence analyses will allow the prediction of disorder and also generate ensembles of conformers, which can be assessed in light of the experimental evidence. Finally, as the (fully or partially) disordered models must be validated, annotated, stored, and disseminated, proper curation and presentation of these ensembles via EMBL's data resources is important. This will ensure that the results are discoverable and interpretable by a broad user community.

Looking Ahead: Delving into the Molecular Details of Symbiosis

Beyond intra-organism and biophysical context, inter-organism symbiosis is central to many aspects of cellular function, and indeed is at the root of eukaryotic evolution via mitochondrial co-option. EMBL researchers are engaged in large-scale projects to study the molecular underpinnings of symbiotic relationships, for example by perturbing individual species in gut microbial communities and engineering synthetic and *ex vivo* communities to dissect functional interactions (Chapter 4: Microbial Ecosystems). These approaches will guide and empower further studies by EMBL to investigate the complex molecular interface between (gut) microbial communities and host physiology over an organism's lifetime, and the intergenerational or evolutionary consequences. The recent construction of two complementary EMBL gnotobiotic facilities will support this. Similarly, mapping and observing host–pathogen interfaces to engineer the next generation of sophisticated experimental models to study host–commensal–pathogen interactions coupled with host genetics will be pivotal in understanding susceptibility to infections (Chapter 6: Human Ecosystems and Chapter 5: Infection Biology). The technology to integrate imaging and molecular profiling data will also play an important role in the identification and molecular understanding of symbioses as they occur in marine and coastal samples to be collected by the TREC project (Chapter 7: Planetary Biology). EMBL already collaborates to understand photosymbiosis between single-celled hosts and microalgae in oceanic plankton, applying a combination of quantitative single-cell structural, chemical, imaging, proteomics, and metabolomics techniques to study cell–cell interactions and subcellular mechanisms. The genetic and epigenetic basis of such interactions can be tested using model systems that extend to animal communities. For example, the genomics, transcriptomics, and epigenomics of *Platynereis* and *Drosophila* and their symbionts will be coupled to environmental context to characterise cellular responses. Working across models and scales, EMBL will focus on symbiotic connections between organisms and on the common molecular underpinnings that explain them.

Impact

EMBL has a strong track record in the discovery and mechanistic dissection of the molecular foundations of biological complexity. EMBL researchers have made seminal contributions to understanding the differential usage of DNA, the dynamic 3D organisation of the genome, and the roles of RNA processing and localisation, post-translational modifications, and protein transport and structure. The new Programme will leverage and significantly extend this expertise, opening up new areas of discovery by gaining a holistic view of the dynamics of large macromolecular assemblies in their *in vivo* context, while integrating information across scales to understand the complex interactions between all of these processes. A universal approach, integrating multiscale *in vitro* and *in situ* information with real-time kinetics is imperative to understand how our genome is used and how genes, proteins, and metabolites interact. But genetics is only one part of the equation. An understanding of the relationship between genes and their environment is fundamental to dissecting and modelling biological complexity.

The need to bridge the anatomical, molecular, cellular, and environmental scales requires not only data integration but a true merging of disciplines to form a new era in the molecular life sciences. EMBL has the expertise and tools to make these fundamental links and to understand the molecular basis of biological complexity and how it is influenced by context. For example, major EMBL projects will generate freely available resources, such as cohorts of mouse models created with precision genome engineering, which can serve as valuable resources to member state scientists, enabling them to tackle central questions in genomics and epigenomics that link development and disease. Using such resources and tools will enable scientists to monitor the nature of responses to environmental fluctuations (e.g. human exposomes in human cohort studies, or soil microbial communities) and will allow the formulation of hypotheses that can be tested in the laboratory context. A deeper molecular understanding will also pave the way for precision

interventions to mitigate the harmful effects of our ever-changing world on living systems, including humans, and will help inform policymakers on issues relating to planetary and human health. With its well-established expertise in molecular biology and technology development, combined with truly interdisciplinary and innovative thinking, EMBL is exceptionally well placed to pioneer technology development programmes, both in modelling the molecular impact of environmental change and in discerning its phenotypic consequences. These developments can also be leveraged by member states and the wider scientific community by sharing knowledge and technologies.

The influence of environmental context on cellular and organismal biology, from molecular processes occurring on timescales of nanoseconds to the lifetime of whole organisms or even multigenerational timescales, is only just beginning to be understood, yet has enormous repercussions. Environmental influences permeate all aspects of life, including both our susceptibility to disease and that of our offspring. EMBL scientists are in an excellent position to discern the underlying mechanisms of **contextual responses** at the molecular level, connecting responses across subcellular scales and compartments. This will provide a basis for decoding the complex interactions across molecular layers (for example, genetic and epigenetic layers) that manifest as specific outcomes in natural or artificially perturbed environments, and which are crucial for almost all aspects of life – from bacterial ecosystems, to organism development, to human health. This research will create a roadmap for understanding how molecular architecture is responsive to context.

3. Cellular and Multicellular Dynamics of Life

Background

For centuries, scientists have been fascinated by the remarkable precision with which organisms develop and the stability yet plasticity that they can show in response to environmental signals. To achieve an accurate and predictive understanding of a complex living system, we must understand how cells respond to intrinsic and extrinsic signals to become organised in time and space. There are immense opportunities to achieve a molecular understanding of how living systems function, respond to ever-changing environments, and evolve.

Many cellular and multicellular processes only make sense in the context of their natural environment inside a cell or organism, or at the interface of these settings with the outside world. To survive and reproduce, organisms – and their underlying molecular, physical, and chemical properties – are highly attuned to these environments, and can respond rapidly and adapt to environmental cues. At the same time, mechanisms have evolved to produce a stable set of internal physical and chemical conditions that can resist environmental fluctuations and ensure the survival and propagation of a species.

Living systems emerge from a web of complex dynamic interactions across scales, and rely on integration of molecular, chemical, and physical cues of different types and origins. These include biochemical interactions to make reactions occur more rapidly or build higher-order structures from the molecular building blocks of life (Chapter 2: Molecular Building Blocks in Context); metabolic or nutritional cues that change the internal and external environment; mechanical interactions to push or pull; and geometric properties such as size or shape. While the impact of the environment on phenotypic outcome, referred to as phenotypic plasticity, is well described at the organism and population levels, the underlying molecular mechanisms remain relatively uncharacterised. The key challenge now is to **reveal the molecular and cellular mechanisms that underlie developmental and phenotypic plasticity**. For a given genotype, the questions are: how much variation is possible, what is due to intrinsic variation or response to extrinsic signals, and how is this achieved?

A major goal is therefore to **unravel the genetic and environmental sources of variability in living systems, to understand responsiveness at the cellular level and in a multicellular context**. This will require novel experimental strategies that allow us to measure, model, and perturb these complex multiscale networks of interactions to reach a fundamental understanding of living systems: how they form, how they respond, and how they evolve.

EMBL is in a unique position to address these challenges, with its foundations of truly interdisciplinary and fundamental research, its strength in technology development and innovation, and its unique collaborative culture. The study of the cellular and multicellular dynamics of life in context will require development of theoretical concepts for living systems, alongside experimental strategies with the right level of granularity to answer specific questions. Developing these experimental and theoretical underpinnings will enable scientists to address key conceptual issues in modern biology, and will have a fundamental impact on the life sciences in general.

The Opportunity

Living systems are extremely complex, not only due to the large number of components they contain, but especially due to their highly interconnected, dynamic feedback regulation, which integrates different modalities (e.g. chemical, physical) and bridges many scales in space and time. Of particular importance is the way in which the different levels of organisation in living systems are interlinked, influencing and feeding back

on each other in a reciprocal manner. In other words, organisms both **shape their environment** and, at the same time, **respond to it**. For example, the molecular structure of cells influences the mechanical properties of the tissues they form. In turn, tissue mechanics impact on cellular signalling and cellular structure. This creates a self-organising and self-referential system of extreme complexity, which raises the question: **what are the logical principles that underlie the emergence of life?** Key considerations include:

- **Robustness.** A fundamental property of living systems is their ability to maintain their function and structure despite internal and external fluctuations in a large number of components and environmental parameters. For instance, how do cells scale their structures to different sizes to maintain order and function in time and space? How do embryos develop normally despite changing metabolic and nutritional cues? How do environmental cues influence the ability of tissues to regenerate? Therefore, a key question is: what mechanisms underlie robustness at the cellular, tissue, and organismal scale? Conceptually, researchers need to understand what determines the limits of robustness, beyond which living systems deteriorate into disease states.
- **Plasticity.** Living systems also have the ability to integrate their internal and external conditions to dynamically adjust their appearance and function; in other words, to generate phenotypic plasticity. Multiple questions remain unanswered in this context. For example, how do dynamically fluctuating molecular networks allow cells to permanently explore different functions? How do environmental changes lead to a functional adaptation and how are these adaptations locked into cellular programmes and genetic or epigenetic memory? What is the contribution of heritable genetic and non-genetic components to this phenotypic plasticity? How are multiple cues integrated across different modalities and scales? How does tissue geometry impact biochemical intracellular signalling?

The dynamic responsiveness of living systems to their environment, linked to **robustness** and **plasticity**, can be analysed across different temporal and spatial scales of life – from the organismal to the cellular and subcellular levels. For the first time, an array of technologies exist to obtain high-quality, quantitative, and dynamic molecular data across scales. Nevertheless, novel experimental strategies and technology development will be required to obtain direct readouts and means of perturbing in a spatio-temporally controlled manner. These perturbations must be made not only at the molecular level, but also by manipulating environmental cues including physical and chemical parameters. Such a combined approach, which integrates new experimental model systems, quantitative measurement, precisely controlled perturbations, and theoretical modelling, is now possible across scales. To test theoretical predictions effectively, this new Programme also includes Data Sciences (Chapter 8: Data Sciences) and Theory (Chapter 9: Theory at EMBL) to explore the vast amounts of quantitative and dynamic data that will be generated. EMBL will be able to provide conceptually new levels of understanding of the logical principles that underlie cellular and multicellular life, such as robustness and plasticity, in the context of the environment.

Research Aims

In the next scientific programme, EMBL aims to develop:

- I. **New experimental strategies to reveal the mechanisms underlying the responsiveness of living systems to their environment.** EMBL's goal is to develop strategies to address questions about robustness and plasticity in response to environmental cues in a wide range of contexts,

including embryonic development, regeneration, and disease states. Of particular importance is the use of suitable, genetically tractable experimental model systems, both *in vivo* and *in vitro*, which will enable a systematic and functional interrogation of the interplay between environmental cues and their effects on living systems, at the mechanistic level. Current research and pilot projects (see below), which span the cellular to the organismal scale, will be the bases to address these questions. EMBL will apply a highly interdisciplinary approach, combining new technology development for quantification (ii), perturbation (iii), and theory and modelling (iv).

- II. **New technologies to generate, integrate, and share quantitative dynamic data on relevant molecular, physical, and chemical parameters.** Living systems depend not only on molecular cues, but also integrate mechanical, geometrical, metabolic, nutritional, and other cues across scales in time and space. One key challenge that must therefore be tackled is the ability to generate **quantitative** and **dynamic** measurements of these **multimodal** parameters in living systems. EMBL will continue to pioneer technologies to allow integrated data capture, spanning multiple types of measurement in space and time. Simultaneously, it is critical to further develop methods that allow the integration, visualisation, and sharing of these multimodal and multidimensional datasets. This will enable them to be used, in combination with functional experimentation, to answer fundamental biological questions.
- III. **New technologies for precise perturbations of biological systems and environments.** Probing, disrupting, and perturbing biological systems (e.g. through classical genetics, or chemical and physical means) have always been at the heart of experimental biology. A vast array of sophisticated and powerful perturbation technologies now exist with increasingly high-throughput approaches enabling multiple components of a modality to be systematically addressed. These range from the panoply of CRISPR/Cas9 or small molecule screens, to highly specific optogenetic and degron approaches that can modulate gene expression and protein function in space and time, especially in *in vitro* models. However, perturbing the physical or chemical parameters inside and outside biological systems in a precisely controlled manner is technically more challenging. In addition, for physiological, cellular, and multicellular systems like primary cells, organoids, or the developing embryo, perturbations are often only low-throughput. A key objective is therefore to develop the next generation of automated technologies to perform systematic molecular, physical, and chemical perturbations of physiological biological systems. Moreover, as key regulatory processes usually occur at precise moments in time and/or positions in space, a new generation of perturbation techniques that can be spatio-temporally controlled in living organisms will be developed at EMBL.
- IV. **Extracting correlations and carrying out predictive computer modelling.** Mathematical models of biological processes, along with data-driven computer simulations, make it possible to create testable predictions about mechanisms. The study of biology has always been driven by a continuous interplay between hypotheses and experimental observations, but is classically based on intuitive logical reasoning. Today, the level of complexity often requires non-intuitive approaches, largely due to non-linearities, complex feedback mechanisms, and the sheer number of interacting components. The formalisation of hypotheses into mathematical models is thus an essential way forward to develop a deeper understanding of biological mechanisms. Furthermore, creating data-driven computer simulations makes it possible to test specific predictions about the underlying mechanisms, explore large sets of possible solutions, and determine the key parameters that distinguish between these solutions. When based on quantitative measurements of the real biological system, sets of alternative predictions can

thus drive the next round of experiments, which can verify or refute the predictions of the model. This important activity will rely on an increased number of theoretical biologists, and close links with the new Theory at EMBL theme (Chapter 9: Theory at EMBL).

The combined experimental and theoretical strategy outlined above forms a continuous scientific cycle where hypotheses, mathematical models, and computer simulations continuously improve and integrate knowledge, driven by new quantitative measurements and targeted perturbations of key parameters. Conversely, the experimental design is continuously refined to test more specific and precise questions driven by model predictions. In this way, EMBL will move towards a fully quantitative, dynamic, and predictive understanding of the cellular and multicellular dynamics of life.

EMBL's Approach

New Experimental Strategies to Study the Mechanisms Underlying Responsiveness to Environmental Cues

EMBL's approach is to develop a cutting-edge experimental and theoretical strategy that combines measurements and controlled environmental parameters, while performing rigorous multimodal quantification of dynamic cellular responses across scales. The development of new model systems and interdisciplinary experimental strategies will have a central role. EMBL has ongoing research in this area and has initiated several pilot projects that aim to tackle this challenge in different contexts, such as embryonic development, regeneration, and disease. 🧑🏫 One ongoing pilot project involves a marine model system, the sea anemone *Nematostella vectensis*, which is being developed at EMBL as an excellent model to study the effects of environmental and nutritional cues, for example in the process of tentacle regeneration. Importantly, genome editing possibilities in *Nematostella* have rendered it a powerful genetic model system. Combined with the ability to generate quantitative data across very different modalities, such as spatially resolved transcriptomics and **spatial metabolomics** (Figure MD1), cellular dynamics, and even the quantification of mechanical properties (Tech Dev Box TD2_MD, and Tech Dev Box TD3_MO), this new model system provides an outstanding opportunity to address the mechanisms underlying the integration of environmental cues, such as nutrition, and the way these mechanisms link to cellular programs of regeneration.

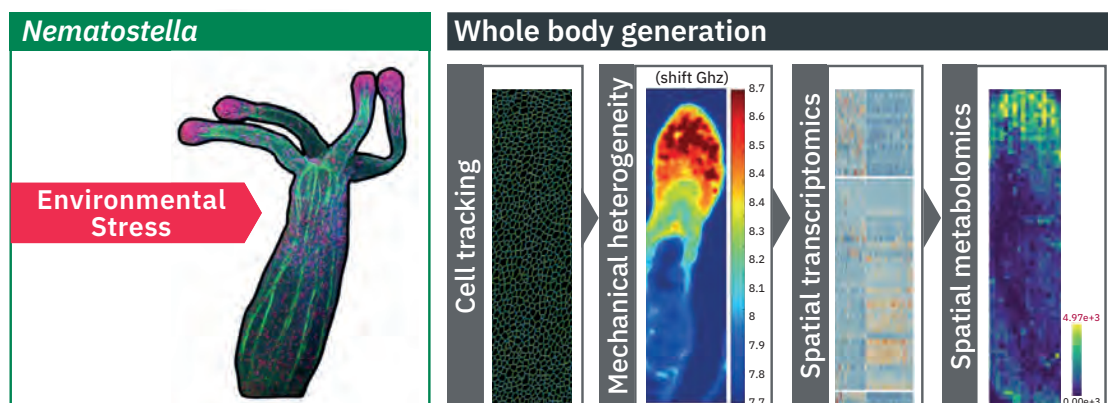




Figure MD1 | The sea anemone *Nematostella vectensis* is a novel genetic model for studying regeneration in the context of organism–environment interactions across multiple scales.

Spatially resolved omics data and imaging-based approaches to probe tissue properties *in vivo* are used to define the multicellular dynamics of regeneration in wild-type and mutant animals when environmentally stressed by injury, heat shock, or chemicals.

 At the organismal and population scales, EMBL researchers will study communities of the fruit fly *Drosophila melanogaster* in controlled laboratory ecosystems to address the effects of changes in environmental conditions (Chapter 7: Planetary Biology). Following defined perturbations at the environmental and metabolic levels, it will be possible to rigorously quantify the effect of these perturbations on gene regulatory networks, signalling, metabolism, and even behavioural responses, both at the level of embryonic development in individual organisms, and at the population level. Again, it is the combination of a genetically tractable model system with the ability to experimentally control environmental cues and to obtain multimodal quantifications of the dynamic responses *in vivo*, that will provide novel avenues and insight into these fundamental questions.

At the cellular level, the influence of the cellular environment is being investigated with a particular focus on the reciprocal interactions between cellular programs and environmental cues such as physical forces (e.g. plasma membrane tension). Altered physical forces are increasingly recognised as playing an important role in allowing cells to sense their external context, leading to dynamic cellular responses including effects on cell fate and cell motility. At the same time, it is necessary to investigate how cells cope with changes in internal or external conditions and yet show **robust** cellular function. By combining newly developed optogenetic tools (Tech Dev Box TD1_MD) and Brillouin microscopy (Tech Dev Box TD2_MD), EMBL scientists aim to address **how physical cues translate into cellular functionality**. Such biophysical analyses will be complemented by integrated informatics strategies that aim to disentangle genetic and environmental factors impacting cellular metrics. For example, a major EMBL study has exploited extensive proteomics datasets to identify the significant impact of altered diet on individual proteotype, particularly on the nuclear pore complex stoichiometry, which in turn has broad cellular influence.

 As well as being impacted by physical and chemical environmental variables, humans can be affected by human behaviour and **social interactions** (Chapter 6: Human Ecosystems). EMBL researchers are developing novel experimental strategies that enable quantification of the effect of social defeat experiences in mice at multiple scales, from neural activity (using calcium imaging) to the detailed extraction of synaptic ultrastructure using high-resolution synchrotron X-ray holographic nanotomography. The pilot project will establish the feasibility of several critical aspects of linking large-scale single unit neural activity recording data to neural ultrastructure and gene expression. It is hoped that, once established, the pipeline could be offered as an EMBL service.

Along these lines, and to investigate how environmental cues impact **disease states**, researchers at EMBL also revert to highly controllable *in vitro* models and drive the development of novel 3D vascularised *in vitro* models of the blood–brain barrier, to study, for instance, malaria pathogenesis and the interaction with brain vasculature (Chapter 5: Infection Biology, and Chapter 6: Human Ecosystems). The *in vitro* strategy complements the *in vivo* organismal studies mentioned above, and could reveal fundamental mechanisms by which environmental and genetic information are integrated and underlie the plasticity and robustness of phenotypic outcome.

New Technologies to Generate Quantitative Data

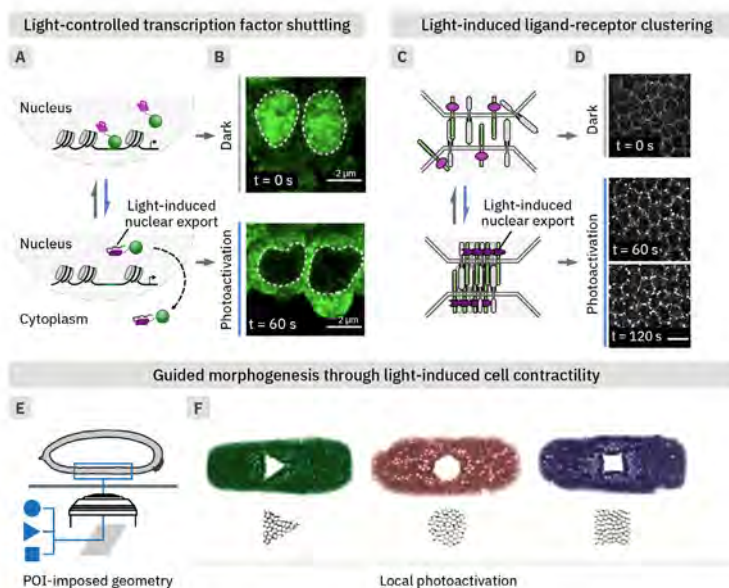
Not all the techniques and tools needed to measure the multiscale nature of life in context, with the required molecular, physical, and chemical modalities, are currently available. For example, techniques to measure changes in physical properties during cellular and developmental morphogenesis and how they are dynamically interacting with, and driven by, the underlying molecular networks, or tools to ask how morphogenesis is affected by changes in the chemical environment are highly needed. To address these challenges, EMBL will develop new technologies that allow dynamic and quantitative measurements of molecular, chemical, and physical properties of cells and multicellular systems. A key goal will be to **develop non-invasive methods**

that perturb the living biological system as little as possible, and seamlessly integrate them with invasive or destructive techniques where no alternative exists to measure the required parameters.

A key set of parameters that currently cannot be measured sufficiently are the physical properties of biological systems. For example, measuring the tension, viscoelasticity, and force anisotropy of the surface and interior of cells and tissues is essential to construct maps of biophysical properties across space and time and connect them with their molecular networks. EMBL is already very strong in biophysical technologies to determine a range of parameters, and has leading expertise in fluorescence correlation spectroscopy (viscosity), laser ablation (tissue tension), atomic force microscopy (AFM), micropipette aspiration (low-frequency surface mechanics, membrane tension), and magnetic droplets (force anisotropy). In addition, EMBL groups have recently developed Brillouin microscopy for non-invasively probing high-frequency mechanics inside cells and organisms (Tech Dev Box TD2_MD). Additional promising methods for future developments include probing single-cell mechanics by acoustic scattering, magnetic twisting cytometry, or optical tweezers.

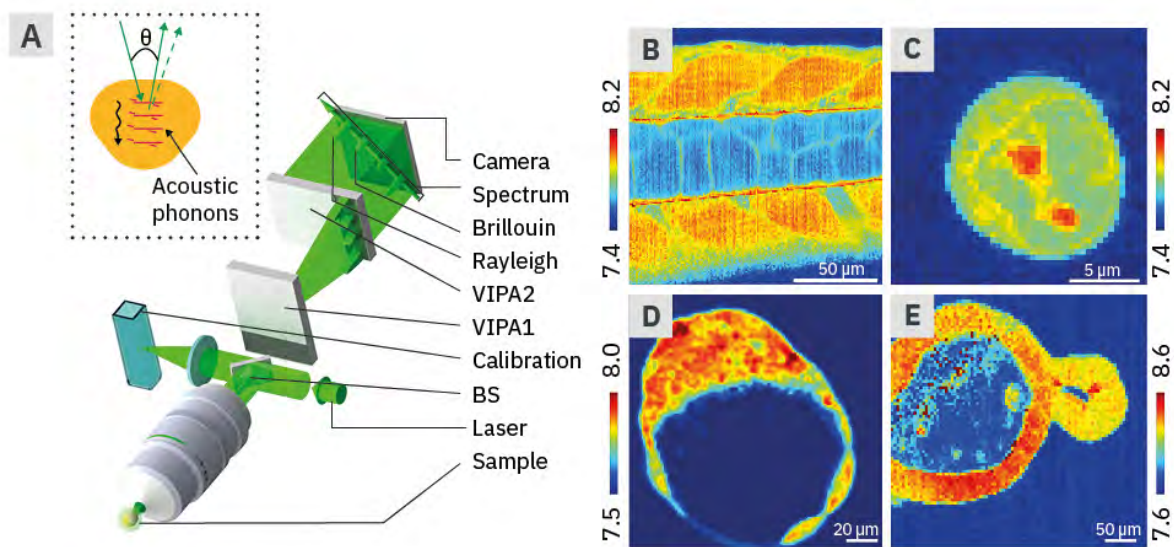
Technology Development Box TD1_MD | Optogenetics.

Optogenetics is a powerful technique that was initially developed to control neuronal activity in living animals. EMBL scientists have pioneered this technology for use in living embryos, with the aim of dynamically controlling protein activity with high spatio-temporal precision as development proceeds. By tagging genes of interest at their endogenous locus with photosensitive protein domains, the activity of endogenous proteins can be controlled at will during animal development. In this figure, a few examples of cellular processes that have been put under optogenetic control are illustrated. **(A–B)** Light-mediated control of the nuclear localisation of the key mesodermal transcription factor Twist during *Drosophila* embryonic development. In less than a minute, the endogenous pool of Twist can be moved into and out of the nuclei. **(C–D)** Light-mediated plasma membrane clustering of the endogenous Delta signalling protein allows fast and reversible inhibition of Notch signalling during *Drosophila* development. **(E–F)** Optogenetic activation of Rho signalling using two-photon illumination allows precise subcellular activation of apical constriction and morphogenesis (tissue invagination) in live *Drosophila* embryos, following the spatial pattern of photoactivation (ROI = region of interest). In the future, the combination of optogenetics with advanced microscopy techniques will allow scientists to elucidate the causal roles of many different parameters during development, giving new mechanistic insights into embryonic patterning and providing the essential data to build predictive models of development.



Technology Development Box TD2_MD | Brillouin microscopy.

Mechanical properties of cells and tissues, such as elasticity and viscosity, are important in determining biological function. However, current biophysical techniques used in the field to assess these biological functions exhibit intrinsic limitations. To enable 3D measurements of viscoelastic properties at high spatial and temporal resolution in biological samples, the Prevedel Group is developing methodologies based on Brillouin light scattering. This new approach, coined ‘Brillouin microscopy’, offers a conceptually novel way to probe elastic and viscous properties of biological materials with subcellular spatial resolution and in a non-contact and label-free fashion. Together with the Diz-Muñoz Group, researchers are applying this technique to questions in cell biology, such as cytoskeletal mechanics during cell division, and are working towards establishing this technology as a more widely used, groundbreaking tool in mechanobiology. EMBL researchers are systematically investigating the relationship between the spectra measured by Brillouin microscopy and common mechanical parameters used in the field, such as the ones derived from atomic force microscopy. Furthermore, techniques to decipher the role of mechanics in morphogenesis and tissue self-organisation are applied by studying early mouse embryogenesis and generating mechanical ‘atlases’ of organisms such as *Platynereis*, which can be linked to spatial gene expression and ultrastructural maps in this animal. This would allow researchers, for the first time, to move biomechanical studies to the molecular regime.



TD1_MD | Brillouin microscopy and its applications in biology. (A) The principle of the Brillouin light scattering interaction with intrinsic acoustic phonons (inset) and a schematic illustration of the imaging system (BS: beamsplitter; VIPA: virtually imaged phase array). Brillouin frequency shift image of (B) a live zebrafish tail tissue, (C) a mouse embryonic stem cell, (D) a preimplantation mouse embryo, and (E) an intestinal organoid. Colour bars denote Brillouin frequency shift in GHz, with higher shift indicating higher elasticity or ‘stiffness’.

Another key challenge in the future is to **correlate multiple methodologies** so that different parameters and data obtained at different resolutions and frequency regimes can be integrated. An example would be a combined Brillouin and AFM instrument, enabling the integration of internal and surface mechanics, as well as high- and low-frequency mechanics. These biophysical methods also need to be combined with the enormous power of fluorescence imaging technologies to probe multiple molecular properties of living systems. Combining Brillouin and light-sheet microscopy, for example, could be used to probe dynamic changes in cytoskeletal structure and changes in stiffness inside an embryo simultaneously. The aim is to make these tools universally applicable and combinable to allow comprehensive measurement of the biophysical state and dynamics of cells and tissues. This would enable EMBL to create quantitative maps of biophysical properties across space and time, and would provide key new parameters to establish accurate physical models of living systems.

Just as important as pushing the biophysical tools, EMBL researchers need to extend the ability to **comprehensively sample molecular properties of living systems in real time and in 3D**. At the comprehensive scale (e.g. using genomics, transcriptomics, proteomics, and metabolomics), this is currently only possible by using snapshots at discrete time points, and thereby precludes the dynamic analysis of live systems. Although some multimodal single-cell methods have been developed, the data are sparse, with complex interdependencies between the quality of the different data types. These single-cell technologies are very recent and constantly improving. EMBL scientists have been at the forefront of their development and use, including the development of computational models to interpret these new types of data. EMBL is also a leader in the development of real-time imaging technologies, ranging from super-resolution, light-sheet, and multiphoton imaging, to alternative approaches such as photoacoustic microscopy, for which molecular reporters are becoming available. EMBL researchers will develop methods to directly correlate and integrate single-cell omics readouts with single-cell imaging. A pioneering step in this direction was recently made at EMBL in another marine model system, *Phallusia*, where single-cell transcriptomics could be integrated with dynamic 4D light microscopy data of the early development from fertilised egg to gastrulation (Figure MD2).

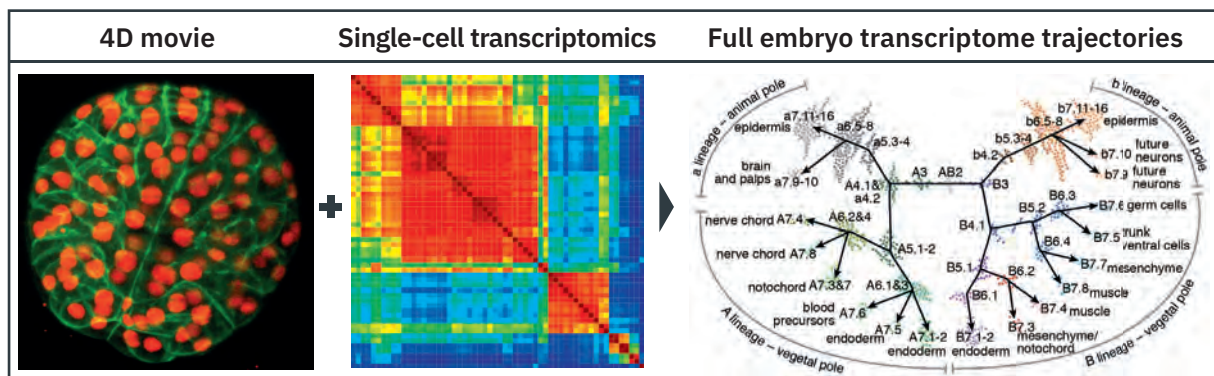


Figure MD2 | Combined 4D real-time imaging and correlative single-cell transcriptome analysis of an embryo of the marine ascidian *Phallusia mammillata*.


3D light-sheet microscopy live-embryo movies of cell lineages were combined with correlative complete single-cell dissection and transcriptome analysis to create complete transcriptome trajectories of all cells in the embryo during gastrulation (<http://digitalembryo.org>).

One key direction for the future includes the development of multimodal spatial omics technologies to measure multiple molecular parameters from the same single cell in developing embryos *in situ*. A second goal is to develop better and more multiplexable labelling technologies for different super-resolution and real-time imaging technologies to non-invasively read out many molecular entities with high subcellular

precision, ideally to the single-molecule level. These crucial labelling and multiplexing tools will largely come from breakthroughs in chemical biology, where EMBL has recently established a new group that focuses on fluorescent dye and novel reporter development. The development of this next generation of new technologies will be essential to generate quantitative data for multiscale biology in context. These technologies can then be made available as new services to member states and beyond, via the new EMBL Imaging Centre (Chapter 10: Scientific Services).

New Methods of Data Integration, Visualisation, and Sharing

A major challenge for multiscale biology is to successfully integrate models and data relating to cellular and multicellular biology from multiple modalities, as well as multiple spatial and temporal scales. Key aspects to overcome include: (i) how to go beyond data aggregation and integration and use the datasets to create useful mechanistic inferences and testable dynamic predictions, and (ii) how to make data openly accessible and reusable so the community can extract new knowledge and create the next generation of models. Interoperability of heterogeneous experimental databases with different modes of information is already possible, and EMBL plans to be at the forefront of this revolution. To bridge the gap between single-cell spatial omics and dynamic, imaging-based molecular and biophysical properties of cells and tissues, data from both domains need to be integrated into unified frameworks, in the form of time-resolved 3D maps (or ‘4D atlases’), which can be interactively browsed and annotated by researchers. This will require multimodal data platforms and interactive data visualisation interfaces for virtual reality to be developed.

At the cellular scale, an initial project was accomplished at EMBL with the dynamic protein atlas of human cell division, which integrates dynamic concentration and subcellular localisation data of the proteins driving cell division into an interactive 4D model that can be browsed in virtual reality (Figure MD3).  At the multicellular or organism level, a pilot project is currently coming to fruition at EMBL on the marine worm *Platynereis*. In this project, a cellular gene expression atlas is correlated with serial block-face electron microscopy data for a specific developmental state of the entire organism. This enables molecular markers to be registered with segmented nuclei and cell types for every single cell within the organism (Tech Dev Box TD2_SS). The ultimate goal for such multiscale organismal atlases is to have longitudinal data over developmental time – or the lifespan of the organism – to capture dynamic changes in cellular transitions.

The development of multiscale and multimodal organismal atlases provides in-depth curation, annotation, standardisation, re-analysis, and integration of independent datasets: for example, molecular data with morphological, metabolic, and mechanical measurements. These types of knowledge bases can provide broad user communities with access to high-quality biological information and analytical tools for data discovery. Here, the unique expertise of EMBL in data warehousing and data sharing will be leveraged to develop the future 4D tools to visualise multimodal biology in context, across spatial and temporal scales. The recent launch of the BioImage Archive at EMBL-EBI (www.ebi.ac.uk/bioimage-archive), together with the unified metadata concept in the BioStudies database, already provides a foundation for linking molecular and spatial data. Equally important is the strong expertise of EMBL research groups in computational image analysis, using artificial intelligence approaches as well as morphometric shape models. The biological atlases they develop will help scientists to achieve one of the ultimate goals of the field: formulating predictive mechanistic models of the dynamics of these systems. Such computational models will link causal relationships at different scales to the emergent dynamics that explain the growth, movement, morphology, and behaviour of cellular and multicellular systems.

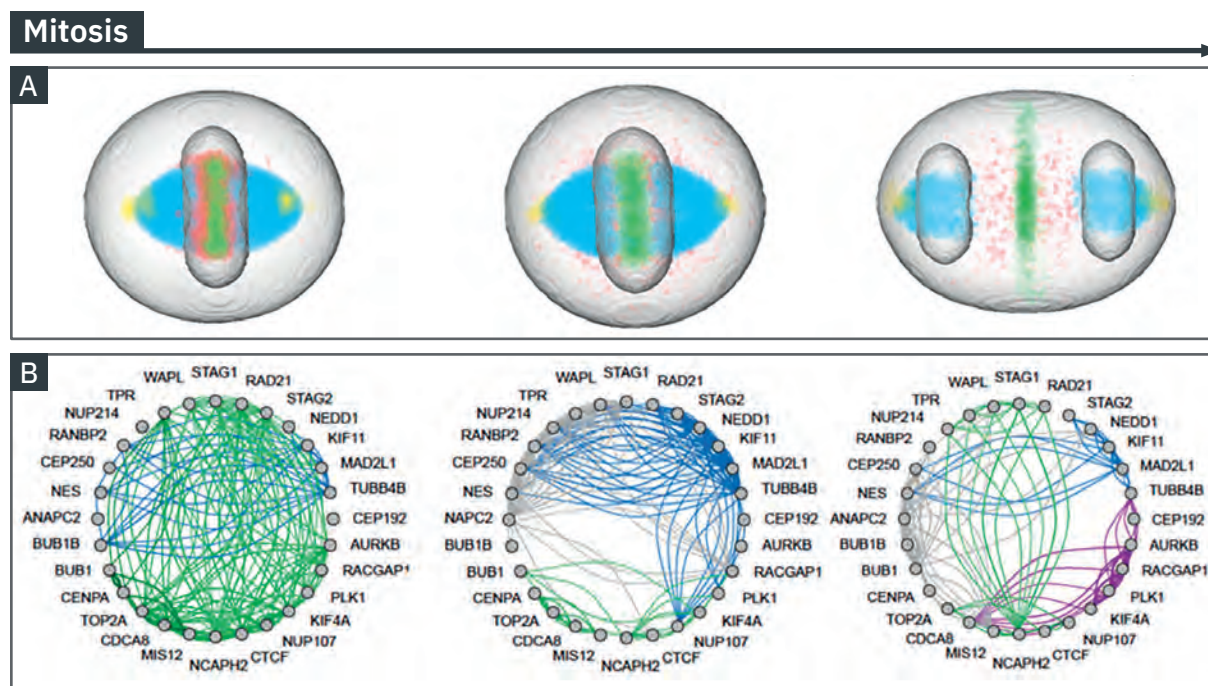


Figure MD3 | Dynamic protein atlas of human cell division.

By integrating experimental and computational approaches, EMBL researchers identified approximately 600 proteins that are needed for a human cell to divide. In an ongoing effort, the timing and subcellular location of those proteins and their interactions are currently identified, enabling the formulation of a data-driven computational model that could predict this dynamic molecular network and relate it to changing cellular boundaries. **(A)** Image analysis and mathematical modelling created a computational mitotic cell model that integrated and visualised protein locations. **(B)** Machine learning was employed to compare the dynamic subcellular fluxes between proteins, allowing the prediction of protein complexes, the temporal order of their (dis)assembly, and the abundance of their subunits. The established approach is generic and can be conceptually transferred to other cellular functions (www.mitocheck.org).

New Technologies for Precise Perturbations of Biological Systems and Environments

Creating precisely controlled perturbations at large scale is essential, both in constructing models and testing their predictions, and to acquire a deep understanding of biological functions to address longstanding mechanistic questions. Single-cell atlas projects are currently generating a wealth of information about which genes are expressed in different cell types. These expression states result from highly interconnected and combinatorial transcriptional networks. Dissecting their regulatory input and functional output in terms of cell function and tissue formation is a pressing current challenge within multiscale biology, which requires rigorous and often systematic experimental testing, combining systematic perturbations with single-cell genomics and other methods.

Recent advances in CRISPR technology have enabled high-throughput screens in cell culture models, either by deleting genetic sequences or by silencing or activating them using dead Cas9 (dCas9) linked to specific domains. Both these techniques are currently impractical to perform at large scale in tissues or embryos. Fundamentally, these ‘genetic’ perturbations are often too crude. This is because a mechanism typically unfolds only at a specific time, such as a critical stage of the cell cycle or development, and/or in a specific place, such as a subcellular compartment or cell type. Beyond qualitative on/off (present/absent) types of manipulations, more tunable and dynamic control needs to occur in order to engineer quantitative changes (e.g. reducing the gene dose or protein affinity) to test quantitative predictions in cellular and multicellular systems.

There is therefore an urgent need for **new technologies that allow systematic perturbations that can be precisely controlled in time and space**. Such new types of perturbation technologies will allow completely new types of questions to be addressed, such as: how are cell-state transitions modulated by perturbing the dynamics of regulator action, or how do perturbations at different timescales affect tissue morphogenesis?

To address this, EMBL researchers are developing new approaches to perturb cellular and multicellular systems with higher throughput, more quantitative precision, and temporal control. For example, most gene regulatory networks do not follow a simple on/off regulatory logic. To alter the levels of upstream regulators in a quantitative manner, perturbations will need to be made over a broad dynamic range, using synthetic programmable transcription factors (such as TALE activators and repressors) and epigenetic modifiers (such as dCas9 activators and repressors). Both technologies have been validated rigorously and applied in cell culture models (Perturb-seq), but currently only at single loci in multicellular systems. Developing methods to enable the systematic use of such systems in primary cells, embryos, and tissues will provide a comprehensive framework for the quantitative control of gene expression, and allow tunable gene regulatory outputs to be created.

Another major effort of EMBL researchers will be to develop systems to control regulators in time. Complex systems are not static atlases, but are rather highly dynamic multimodal networks that constantly change in both their molecular components, such as their levels, interactions, and regulation, and their biophysical properties. In essence, living systems operate as ‘open’ systems that are out of equilibrium, continuously exploring different states and rapidly transitioning from one state to the next. Although inherent in all living systems, the precise manipulation of dynamic networks in time has remained a huge challenge, given the pleiotropic nature of most essential regulators, and the irreversible nature of most genetic perturbations. To perturb a system in real time, researchers at EMBL are therefore developing multiple methods to control or modulate the timing of molecular components in cellular and multicellular systems. This includes the development of microfluidics-based systems to synchronise and control the timing of the cell cycle or organism development, by modulating, for example, the timing of oscillating systems. Opto- and chemogenetics are another very exciting pair of technologies that EMBL scientists are developing in multiple directions to induce different types of perturbations in live cells and developing embryos. Here, the rapid mislocalisation of key effector proteins within the cell can be induced by a laser beam or the addition of a small molecule that allows proteins or RNAs to switch from an active to an inactive state (Tech Dev Box TD1_MD). The power of such opto- and chemogenetic methods is their highly dynamic and reversible nature. Such precisely controlled temporal perturbations will yield new types of data, which will be complemented by the development of new computational methods to model dynamic trajectories combined with inference-driven experimental designs to select the most informative time points to perturb.

Extracting Correlations and Carrying Out Predictive Computer Modelling

In order to reveal the logic of dynamic living systems and their dynamic response to changing environments, EMBL will combine novel experimental model systems and the rich four-dimensional and multimodal data with theoretical and modeling approaches. Beyond the integrated multimodal databases and portals described above, EMBL will focus on the theoretical foundations of the field, to harness the rich four-dimensional molecular and physical data and discern the underlying logic of dynamic living systems.

A first level of prediction stems from quantitative correlations within the data. If both physical aspects of a cell (e.g. membrane tension) and certain molecular measurements (e.g. protein abundance) are found to be correlated in space across many samples, this can suggest a functional interaction. New sophisticated statistical methods and machine learning techniques have become powerful tools for finding such similarities and patterns within large datasets, and can span both omics and imaging data. Developing these important

approaches further, for multiscale biology, will be pursued together with EMBL's Data Sciences programme (Chapter 8: Data Sciences).

A second level of prediction can be used to define mechanisms. Understanding the dynamic mechanisms behind the behaviour of multiscale biological systems requires formalising hypotheses into mathematical models that can be used to simulate the system computationally. This is because behaviours emerge at higher levels through the nonlinear interactions of numerous components that are structured at lower levels, such as the emergent cellular behaviours arising from the dynamic interactions of regulatory and structural molecules. Mathematical models need to provide non-intuitive but concrete predictions, which can be tested with further experiments. For example, a model should ideally predict the change of a cell's stiffness after receiving an external signal, or the altered morphology of an embryo after knocking down a gene.

EMBL is developing, and will continue to pursue, a variety of mathematical modelling and computational simulation approaches to create **quantitative models** of dynamic 3D cellular and multicellular systems. Useful models will be those that produce alternative hypotheses and distinct concrete predictions that can be tested experimentally. One example of an ongoing multicellular modelling project that EMBL is tackling is mammalian organogenesis (Figure MD4). This data-driven project exemplifies the complete modelling cycle. Such a modelling framework can search through very large parameter spaces by running the slightly altered simulations millions of times; however, simulations are only useful when compared to real data, such as when the outputs of the simulations can be automatically and quantitatively compared to the atlas of real gene expression patterns. Computer simulations can, in principle, take into account different modalities of the properties of multiscale biological systems. These include molecular properties that cross different layers; physical properties such as linear forces, pressures, or tensions; mechanical properties such as elasticity, viscosity, plasticity, or anisotropy; and geometric properties such as morphometry, vectorial orientations of processes, and intercalation.

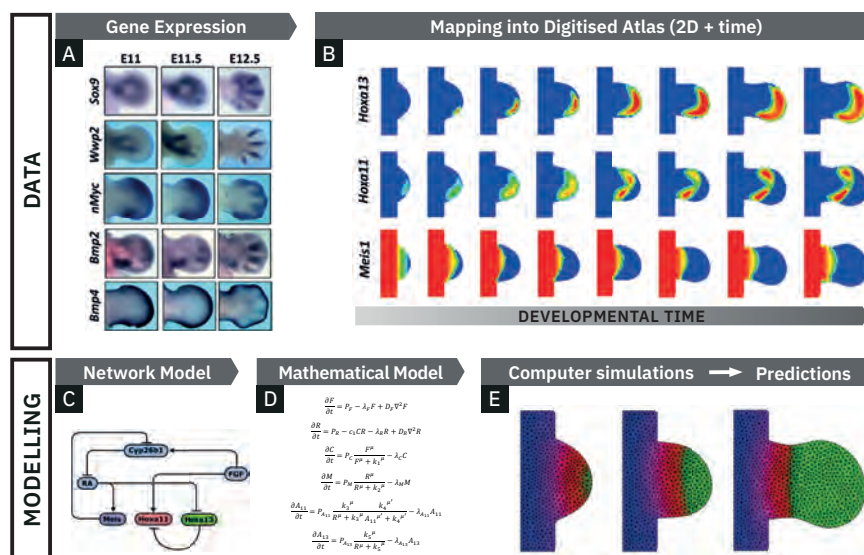


Figure MD4 | Computational modelling of limb development can create quantitative and testable predictions about dynamic gene expression patterns.

(A) The molecular expression patterns for hundreds of genes can be assayed within the whole tissue and imaged. (B) The spatial patterns of genes at many time points during development can then be digitised to create a 2D atlas over time and space. (C) Hypotheses that may explain the gene expression dynamics can be represented as network models, and (D) formalised into mathematical models. (E) These models can then be explored in a computer simulation that makes concrete predictions about the gene expression patterns over all time points during development. Importantly, the quantitative predictions can then be compared automatically with the digitised data (shown in B), to determine how accurate the model's predictions are, and ultimately to guide the researcher towards better hypotheses and more informative perturbation experiments.

Modelling complex systems is non-trivial. In practice researchers have to make choices about how to simplify biological systems to make them amenable to systematic experimental measurement and to computer simulations. A key decision to answer any multiscale biology question is also to find the right level of granularity or level of abstraction, which may range from understanding how every known molecular entity fits into the overall picture to capturing just enough quantitative data to make correct predictions about the functional cell or tissue-level behaviours of interest, such as the overall morphology of an organ. These alternatives are not just a matter of choice, but in fact represent a genuine deep scientific question, at the heart of studying complex multiscale biology. The close integration of **theoretical coarse-graining approaches** that are applicable to multiscale biology is key to guide these decisions. Such an integrated approach will also build on and feed into the new Theory at EMBL theme (Chapter 9: Theory at EMBL). Combining experimental, modelling, and theoretical research is urgently needed to understand how life responds and adapts to its ever-changing environment.

Impact

Understanding the biological principles that underlie cellular and multicellular life and its robustness and plasticity in response to changing environmental context, will not only provide fundamental new insight into what underlies **normal development and healthy life**, but will also be essential to revealing how living systems can deteriorate in ageing, disease, or through disruptive environmental changes. It is important to understand what determines the limits of robustness, beyond which living systems deteriorate into disease states. Thus the novel insights from this research area will also provide a basis for the rational design of cells and tissues, ranging from patient-derived induced pluripotent stem cells to powerful *in vitro* models of disease. EMBL research groups are already moving into tissue engineering; for example, integrating human perfusable vascular networks with other cell types to create *in vitro* tissues that enable the study of placental dysfunction, cardiac tissue regeneration, drug delivery to tumour models, and even the pathogenic processes of malarial infection in a model of the human blood–brain barrier (Figure IB2). The deeper our understanding of multicellular dynamics, the greater our capacity to study human diseases *in vitro*.

The study of cellular and multicellular life in context will address these fundamental questions by providing **the next generation of experimental strategies and technologies** to generate quantitative and dynamic molecular and physical data to perturb biological systems with exquisite control in space and time. EMBL develops technology with the goal of transferring it as efficiently as possible to the wider scientific community. Therefore these new tools will be designed to be generally applicable to many biological systems, and will be shared with EMBL's member state community via new scientific services and training activities as soon as possible (Chapter 10: Scientific Services, and Chapter 11: Training).

Exploring cellular and multicellular dynamics will also provide high-value **data portals for cellular and multicellular models**. EMBL will use its strong foundations in open data provision to share large and multimodal datasets effectively, to set up new core EMBL data services, and in an easily accessible manner to maximally enable community impact (Chapter 8: Data Sciences). In addition, EMBL will develop new mathematical models and computer simulations of dynamic living systems, which will allow the exploration of much larger parameter spaces than is experimentally possible, and again share them with the community to use, build on, and further contribute to. EMBL will abstract from such models general principles and best practices for moving biology from the big data, quantitative discipline it is today, into a predictive science for the future.

4. Microbial Ecosystems

Background

Microbes are the most ancient, abundant, and diverse form of life on Earth. They have co-evolved with and shaped our planet, and without them life as we know it would not exist. Although often considered as simple or primitive, these unicellular organisms colonise, proliferate on, and impact every biotic and abiotic surface and subsurface of our planet, even in its most inhospitable corners. Microbes can be found literally everywhere: in and on animals, on plants, in the soil, in aquatic environments, in food chains, in everything that humans can and cannot touch (from household and medical appliances to the bottom of the ocean, and from microbial mats in hot springs to the permafrost). Microbes do this mostly in the form of complex communities, in which they interact with each other and with their surroundings, forming complex microbial ecosystems. In the past two decades, DNA sequencing technologies have exposed the breadth, complexity, diversity, and ecogeography of such microbial communities, also referred to as **microbiomes**. It is currently estimated that more than 10^{12} microbial species reside in these communities, yet only around 10^4 of them have been isolated or have available reference genomes. Their interactions and functional capacities, as well as the effects they have on their surroundings, are even less well characterised.

Mapping the microbial diversity within an ecosystem is the first step towards understanding the functional role in the ecosystem of individual microbes and their communities. Knowledge of *who* is there paves the way for answering questions on *what* actions they perform, *how*, and *why*. The biological principles, molecules, interactions, and functional outputs will help answer questions on community stability and niche specificity. At present, the only microbiome for which there is a good understanding of its diversity is the human microbiome – the collection of microbes that live on and inside humans. The human gut microbiome in particular has emerged in the past decade as a prototypical microbial ecosystem, due to its accessibility, tractability, richness, and direct relevance to human health. As a ‘human organ’ with roughly 100 times more genes than the human genome, the gut microbiota plays an essential role in host health: helping in host metabolism, immunity, brain function, and response to medication. Numerous studies associate changes in microbiota composition to susceptibility to infections and diseases such as diabetes, colon cancer, cardiovascular and neurological conditions, as well as many others. The development and progress of these diseases is thought to be linked to altered functional outputs of the microbiota as a result of changes in its composition. However, many of these links, as well as the overall view of the gut microbiota and its interplay with the host and the environment, remain largely descriptive, relying on associations between phenotypes of interest and microbial community composition. The next challenge lies in mapping the causal effects and understanding the underlying molecular mechanisms of gut microbiome–host–environment interactions. The knowledge derived can lay the groundwork for understanding other complex microbial ecosystems as they are better characterised.

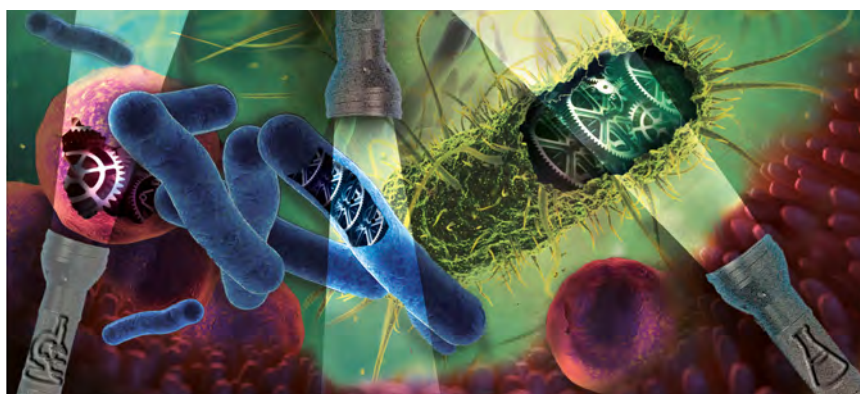


Figure ME1 | Shedding light on the dark genetic matter of the gut microbiome to help establish *de novo* gut model organisms.

The Opportunity

To better understand microbial ecosystems, their functional capacities, and their molecular interplay with the environment, EMBL researchers first need to study the individual players within microbiomes: the microbes themselves. Characterising microbial communities and the interactions within them and with their natural environment, will create an understanding of microbial communities' roles in their habitat, and of the underlying molecular players (genes, proteins, metabolites) and mechanisms. This knowledge will lay the foundations for intervention and rational modulation of microbiomes from dysbiotic states, and/or to help to rebalance or restore the ecosystems they live in. In the case of the human microbiota, it will enable personalised, microbiome-related therapies. In the case of soil and marine microbiotas, it will open unprecedented opportunities for reviving and rebalancing natural ecosystems.

The first step in this process is to collect the available information on microbiomes in one place, allowing for systematic curation and dissemination to the scientific community. This will set standards for future investigations, provide a framework to report and query study results, and facilitate the identification of knowledge gaps to tackle further research. The second step is to develop experimental and computational resources and tools to functionally study isolated microbes and microbial communities in and out of their natural context. These resources encompass strain collections, pipelines for systematic genetic manipulation of the microbes they contain, model communities, experimental platforms to study these communities (e.g. approaches to link genotypes to phenotypes, and to assess the impact of controlled conditions and perturbations), and frameworks to integrate multi-omics and multiscale data. Building from its pioneering work and growing expertise on the human microbiome, EMBL is in a unique position to start dissecting and understanding the underlying principles and molecular mechanisms of the assembly, dynamics, and properties of microbial ecosystems.

Research Aims

Over the past decade, EMBL has been instrumental in developing computational and experimental tools and approaches that have propelled research on the human microbiome, facilitating our current understanding of the microbiome's diversity and its role in human health. These technological advances (Tech Dev Boxes TD1-3_ME) have provided novel insights into human gut microbiome composition across populations, the impact on it of perturbations and age, its links to disease, its interfaces with drugs, its strain resolution, and its encoded functional diversity. In the new Programme, EMBL aims to become a leading hub of microbiome research and resources in Europe by strengthening existing tools and developing new ones, in close collaboration with other leading research organisations in the field, in member states and elsewhere. This will involve the assimilation and curation of microbiome sequencing data, followed by integration with systematic experimental strategies to gain insights into the fundamental principles that shape microbial ecosystems and the functional capacities they encode. **Microbes that colonise the human gut will initially be used as models**, with the focus gradually expanding towards more fastidious and understudied environmental microbiomes related to, for example, soil, plants, marine waters, or other natural environments, with roles in plant growth, carbon cycling, antibiotic biogeography, or pollution degradation (Chapter 7: Planetary Biology). EMBL is in a unique position to address this ambitious initiative, which will **promote the transition of microbiome research from descriptive and correlative to molecular and causal**, revealing the underlying mechanisms and interactions supporting complex microbial communities and their interactions with the environment.

This level of understanding will ultimately enable the rational modulation of these communities, when required. EMBL specifically aims to:

- I. **Understand the functional diversity of individual microbial species and strains.** Current knowledge of bacterial gene functions, pathways, and cellular architecture stems from very few model bacteria, which fail to capture the phylogenetic and genetic diversity of the gut microbiota. As a consequence, the vast majority of genes in the gut microbiome remain ‘dark matter’ with respect to function: that is, of elusive or completely unknown function. EMBL will be the hub for community efforts to **systematically tackle the vast genetic dark matter in the human gut microbiome and to establish new relevant model microbes**. These efforts will build upon existing resources to develop the next tier of microbiome-related computational tools and databases, and the development of automated high-throughput experimental pipelines. This will unravel novel pathways and protein machines, illuminating how microbes produce bioactive molecules, communicate with each other, survive stress, resist or modify xenobiotics, and metabolise nutrients. Importantly, it will also provide a roadmap for generating functional knowledge about key microbes in any microbiome or ecological habitat.
- II. **Dissect the interactions and properties of microbial communities.** Microbes within complex communities, such as the gut microbiota, compete for resources but also cooperate to break down complex food sources, communicate with each other, fend off intruders, and deal with stress or fluctuating environments. Using high-throughput experimental setups at different levels of complexity and tractability (from monocultures and microbial communities to human donor cohorts), EMBL researchers aim to understand the underlying principles driving the organisation, stability, and characteristics of gut microbial communities, and to establish model communities. A specific focus will be on the role of the gut microbiota in containing and combating pathogens, and on the development and spread of antimicrobial resistance.
- III. **Place microbial communities in their ecological context.** Microbes and microbial communities are shaped largely by interactions with their environments. They also have functionalities that only become relevant in their natural ecological context. To study microbial organisms and communities in their natural context, which in the case of the human gut microbiota is the human host, EMBL aims to study the functional outputs and characteristics of such communities in co-culture with their host cells, in organoids, or in gnotobiotic animal models. Within this endeavour lies a unique opportunity to understand the principles and molecular mechanisms by which intrinsic factors (e.g. bacterial genomes, metabolic pathways, and protein assemblies), extrinsic factors (e.g. nutrients, xenobiotics, and host immune response), and evolutionary processes determine the composition and functioning of microbial ecosystems.
- IV. **Modulate microbial communities and their interactions.** The tools and knowledge generated on single species functionalities and on community organisation, interactions, and properties, will serve as a basis for moving towards strategies for rational modulation of the microbiome. An iterative approach, combining high-throughput experimentation with modelling and machine learning, will pinpoint abiotic (pharmaceuticals, xenobiotics, food, prebiotics) and biotic (phages, probiotics, microbial species) strategies to shift or change microbial community compositions in a targeted manner. These modulations will range from precise removal or exchange of a single strain (e.g. a pathogen or a microbe carrying easily transmittable antibiotic resistance) to more radical partial or whole community transplantations. The ability to rationally modulate microbial communities will pave the way for new therapeutics and biotechnological applications.

- V. **Expand and translate knowledge to other microbiomes.** The human gut microbiome will be used as an exemplary microbial ecosystem to chart interactions, understand gene functions, dissect the underlying mechanisms, and probe their impact and role in their ecological context. The accumulated knowledge of microbial function and biological principles, as well as the developed bioinformatics and experimental pipelines, will be used as a springboard to ask similar questions and start work on other relevant microbiomes. Studying other microbiomes in the human body (skin or respiratory or reproductive systems) or ones that humans are exposed to daily in their lives (those found in food and in the natural and built environments in which we live) would be the natural extension to these efforts. In conjunction with plans for field work and *in natura* measurements (Chapter 7: Planetary Biology), EMBL will seek to develop appropriate experimental platforms for cultivating, probing, and characterising microbes and microbial communities from such environmental settings.

EMBL's Approach

Understanding the Functional Diversity of Microbial Species and Strains

Current knowledge of microbial gene functions stems from a few model organisms (e.g. *Escherichia coli*, *Bacillus subtilis*) and pathogens, which bear little resemblance to most species found in the gut microbiota (Figure ME2A-B). Diversity in microbes is high, due to their long evolutionary history and fast reproduction. Hence the phylogenetic distances between bacterial species are much larger than those between eukaryotic organisms, which limits the utility of inferences about gene function and physiological context based on homology to current microbial model organisms. Not surprisingly, more than 60% of the genes in the human gut microbiome remain of unknown function, a **dark genetic matter** that impedes our understanding of their encoded functional capacities and contributes significantly to the outstanding body of proteins of unknown function found across all sequenced organisms (Figure ME2C). To bridge this gap, **EMBL proposes to generate the functional knowledge and resources required to establish new model organisms spanning the diversity of the human gut microbiome.**

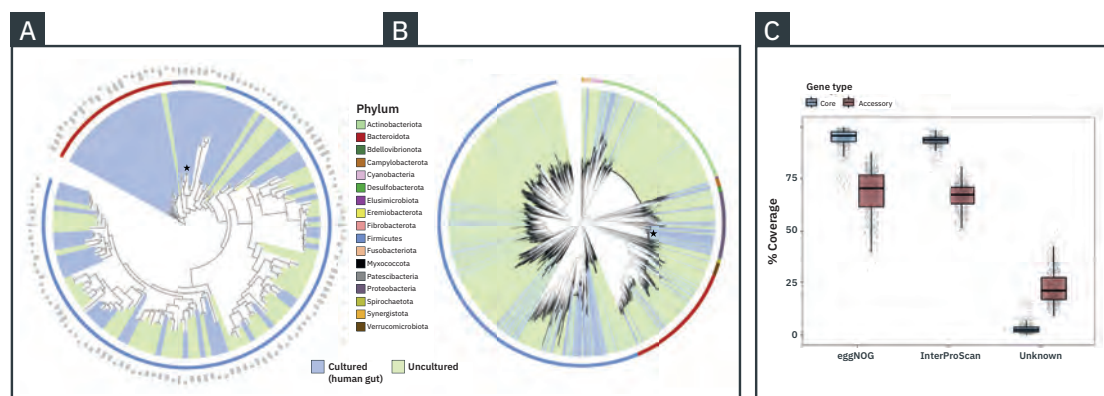


Figure ME2 | The gut microbial species diversity highlights current challenges – the lack of model species and many functionally uncharacterised proteins.

(A) Phylogenetic tree of the most prevalent and abundant bacterial species in the human gut. Species are shown in which prevalence $>10\%$ and relative abundance $>10^{-5}$ in 2,803 healthy individuals across the world. *E. coli* (black star) is the only model organism in the list of species. The phylum of these species is indicated in the outer coloured ring. Most of the species that are cultured (light blue, inner colouring) are part of EMBL's culture collection. (B) Phylogenetic tree representing the 4,644 bacterial species found in the human gut, based on meta-analysis of $>12,000$ human gut metagenomes (Almeida *et al. Nature* 2019). Over 70% of these species are yet to be cultured (green, inner colouring). The location of *E. coli* is indicated as before. (C) Pangenome (the entire gene set of all strains of a species) analysis of the complete set of 205,000 novel genomes produced, following functional annotation by EMBL resources (eggNOG and InterProScan) reveals that, while most core proteins can be functionally annotated, our knowledge of accessory proteins is far more limited. Typically, 20% of the accessory proteins lack any functional annotation, highlighting the current microbial dark genetic matter (Almeida *et al. Nature* 2019).

Establishing Comprehensive Databases and Computational Resources

Metagenomics has enabled the vast functional diversity encoded in the human gut microbiome to be understood. With microbiome sequencing information being released at an ever-increasing pace, and with any two human individuals differing by more than 90% in terms of their microbiome strain content, while being more than 99% genetically identical, there is a need for a single comprehensive catalogue of human gut microbiota composition and functions. Propelled by work led by EMBL groups on sequencing and defining bacterial pangenomes, mapping strain diversity in microbiomes, and assembling genomes from metagenomics data, it has become evident that less than 30% of the approximately 4,500 bacterial species in the human gut currently have culture isolates (Figure ME2B). This makes harmonisation of data-driven efforts crucial to put future metagenomics studies in context, and to guide experiments that aim at dissecting these microbiomes at the molecular level.

EMBL will build on its experience as a leading provider of diverse computational resources in the life sciences, many of them, such as SpecI, mOTUs, MGnify, proGenomes, GenomeProperties, eggNOG, InterPro, STRING, iTOL, and iPATH, being intimately related to analysing microbiome data and/or representing their phylogenetic and functional diversity (Tech Dev Box TD1_ME). These databases will continue to be expanded, and will also serve as the basis for a new integrated resource, which systematically captures and catalogues all bacteria, viruses, and eukaryotes in the human gut microbiome. Information on genomes from isolated species and metagenome assembled genomes (MAGs) will be a particular focus, as they provide comprehensive and uniform taxonomic and functional annotation of genes, genomes, and pangenomes to drive functional hypothesis generation at the level of single microbial species and strains. They also dramatically increase the power to identify functional associations based on genomic context information. Key host metadata such as geographical distribution, age, and health status of associated samples will be curated and linked to these functional resources, to enable associations between functional capacities and environmental factors. The data coordination efforts to build this resource would unite various fragmented efforts in the field, remove redundancy, and ensure consistent quality controls. The large amount of data collected, curated, and managed will provide an unprecedented resource for investigating pangenome information and strain-level variation across microbiomes, not only those living in or on humans but also those found in the ocean or soil.

Developing Experimental Resources and Automated High-throughput Pipelines

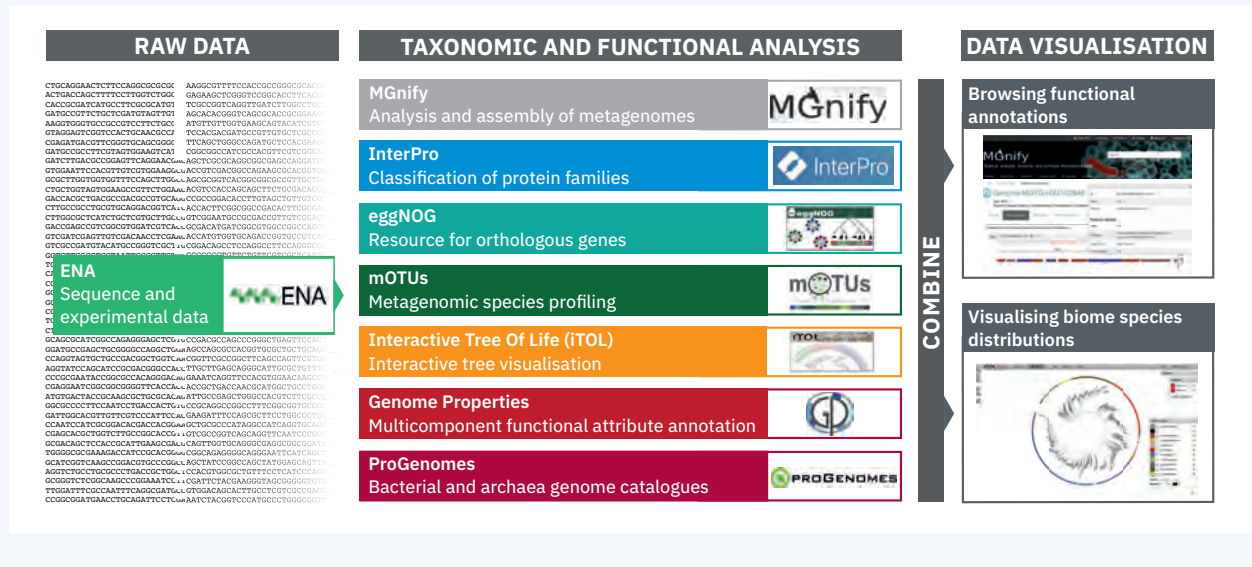
Automated experimental setups to cultivate and study gut microbes have recently started to emerge, with EMBL at the forefront of such developments (Tech Dev Box TD2_ME). Such setups enable the study of these still largely unknown microorganisms at the molecular level in controlled settings. EMBL aims to continue and expand these efforts. First, EMBL will isolate, biobank, and sequence thousands of strains from human individuals to build a repository for functional studies and for understanding the dynamics and properties of different microbial ecosystems. Working in close collaboration with national strain collections, other microbiome cultivation initiatives, and expert labs, EMBL will focus on assembling large strain collections of a selected set of abundant and prevalent gut bacterial species. These collections will be invaluable in efforts to link genes to functions and to other genes, e.g. genes co-occurring or co-evolving and required for a given function or phenotype. Second, EMBL will expand its present capacities to systematically cultivate, perturb, and monitor gut microbes in controlled environments and automated settings. This will involve building new tailored quantitative assays (fitness-, morphological-, genomics-, proteomics-, and metabolomics-based readouts) and pipelines in which microbes can be exposed in parallel to hundreds of ecologically relevant perturbations while their phenotypes are monitored (Tech Dev Box TD2_ME). Third, EMBL will establish genetic tools for synthetic biology approaches and will construct genome-wide mutant libraries, including single-gene knockout, knockdown, and overexpression libraries, for systematic functional studies. These libraries will enable systematic studies of gene function, as well as tailored mechanistic work on human gut microbes and their communities.

Technology Development Box TD1_ME | Microbiome computational tools and databases.

EMBL has pioneered the development of computational tools, databases, and web resources for microbiome research. These include ways to delineate prokaryotic lineages (proGenomes2 – Mende *et al. NAR* 2020), analyse phylogenetic trees (iTOL – Letunic and Bork, *NAR* 2019), profile shotgun metagenomes (mOTUs2 – Milanese *et al. Nature Communications* 2019), and assess eukaryotic microbial genome quality (EukCC – Saary *et al. bioRxiv* 2020). Further integrated computational workflows will increase utility to cater to various research areas and applications, from clinical microbiome biomarkers to diverse environmental microbial communities.

Functional analysis of the genetic diversity encoded in metagenomes is still challenging. Researchers at EMBL were among the first to catalogue the genetic and functional diversity of the human gut microbiome. Databases maintained at EMBL (Pfam, SMART, InterPro, UniProt, eggNOG) are key for functional annotations. EMBL scientists will ensure that these continue to incorporate the latest information, and will transfer annotations to all sequenced organisms.

Recently, over 200,000 bacterial genomes were assembled from human gut metagenomes, and over a billion distinct microbial proteins from diverse habitats were catalogued in EMBL databases (Almeida *et al. bioRxiv* 2020). To increase data accessibility and enable discoveries, integrated infrastructures (e.g. MGnify, Ensembl) and new concepts for data mining will be developed.

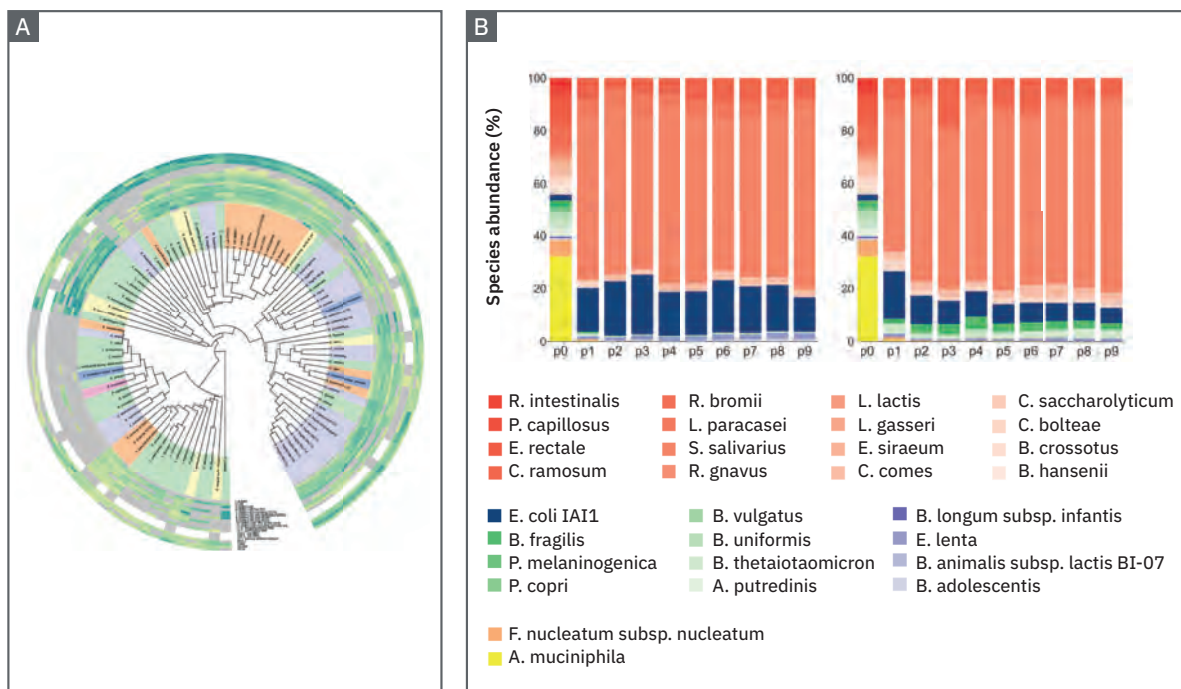


Technology Development Box TD2_ME | Automated microbiomics platforms.

Human microbiome research has been propelled by data-driven science, largely directed by genomics analyses of samples without cultivation. Although these approaches provide unique *in natura* insights into communities and permit associations to health and environmental parameters, they are not enough to understand the underlying causal relationships and mechanisms. EMBL has been at the forefront of establishing experimental setups to investigate the human gut microbiome.


(A) EMBL researchers from the Bork, Patil, and Typas groups have assembled a collection of prevalent and abundant human gut species (now containing >100 species), developed robust and automated cultivation pipelines, and mapped the metabolic capacities of these species (Tramontano *et al. Nature Microbiology* 2018). EMBL researchers have used this collection to systematically profile the interactions of gut microbes with environmental perturbations (Maier *et al. Nature* 2018; *bioRxiv* 2020) and with each other. **(B)** The same pipelines were used to assemble large numbers of complex communities. Here, two stable communities are shown over time, after mixing 32 species in different media, to profile emergent phenotypic traits of communities.

As a next step, EMBL is moving its suite of unique high-throughput reverse genetics (Nichols *et al. Cell* 2011; Kritikos *et al. Nature Microbiology* 2017; Galardini *et al. eLife* 2017; Brochado *et al. Nature* 2018; Zimmermann *et al. Nature* 2019) and MS-based approaches (Tech Dev Box TD_ME_3) from model microbes to less well-understood gut microbes. EMBL researchers are also developing more complex, high-throughput microbiomics setups (Figure ME3), with a special focus on interrogating personalised microbiome communities to empower precision medicine solutions.



Mapping Gene Function, Protein Complexes, and Cellular Pathways

Gut microbes have an enormous capacity to degrade and utilise nutrient sources; produce bioactive molecules, including essential vitamins for the host; sense and transduce signals; respond to and protect themselves from stress; interact with each other; train the host immune system; and fight off pathogenic intruders. To gain insights into this functional diversity and create the foundational functional knowledge required for new model organisms, EMBL will combine its unique experimental resources and automated pipelines, with cutting-edge technologies spanning omics to structures, and with computational approaches for data analysis and integration. For example, using strain collections and genome-wide mutant libraries in novel metabolomics- and proteomics-based read-outs will allow EMBL scientists to link genes to substrates and products, map enzymes and transporters to pathways and to their ligands, and chart the organisation of the metabolic network of these microbes (Tech Dev Box TD3_ME). Similarly, combining hundreds of genetic, chemical, or environmental perturbations with quantitative read-outs (e.g. fitness-based omics) will unravel genotype-to-phenotype relationships, and uncover gene function and organisation en masse (Tech Dev Box TD2_ME). Both the establishment of innovative and diverse quantitative omics readouts and the integration of multi-omics data lie within the core expertise and areas of excellence of EMBL. The data generated from these approaches will be used to chart the main functional units of a plethora of evolutionarily distant gut microbes, as well as provide insights into their function, regulation, and interconnections. The data will also shed light on the microbial dark genetic matter, making it possible to infer the functions of thousands of orphan proteins, and providing leads for further molecular characterisation through biochemical or structural investigations. Here, EMBL will capitalise on its expertise in structural biology (cryo-EM, X-ray crystallography, and NMR spectroscopy) to answer mechanistic questions about protein complexes of any level of complexity or size by determining high-resolution structures. These techniques are optimally suited to map diversity at the level of sequence, function, and physiological relevance onto the atomic level.

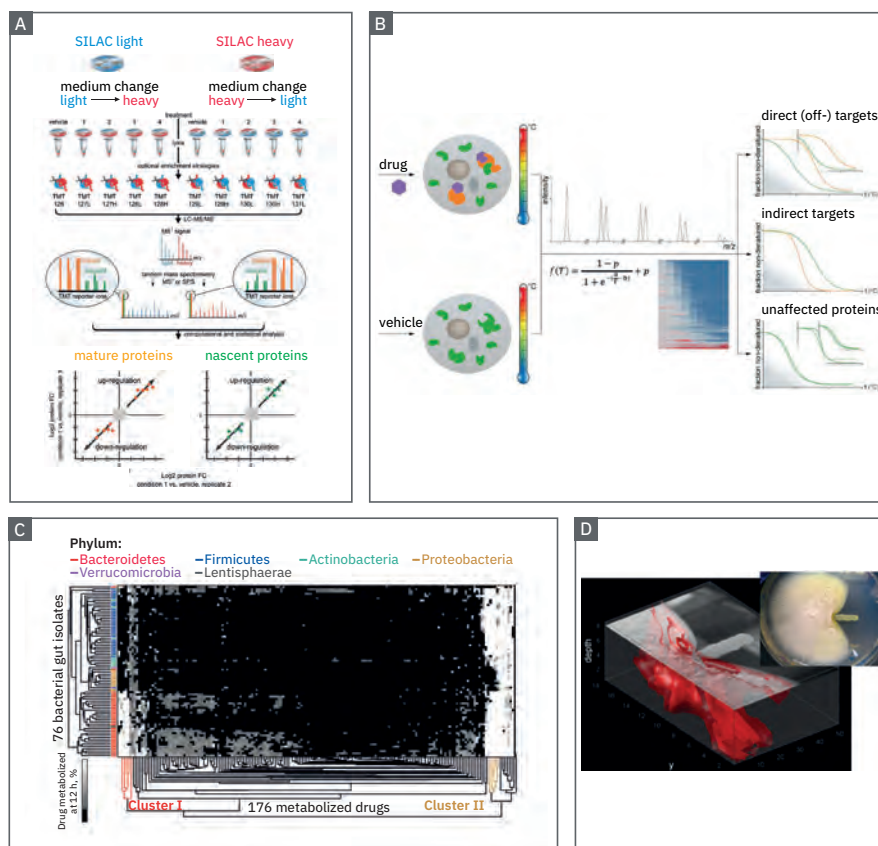
 To get this ambitious plan off the ground, a workshop entitled ‘Unlocking the Gut Functional Diversity’ (Figure ME1) has been organised to bring together key stakeholders, including scientists, journal editors, and funders, from EMBL member states and beyond. The goal is to chart the framework of tools, approaches, and strategies required for generating comprehensive functional knowledge about prevalent and abundant gut species. This knowledge and the available tools will act as a foundation for establishing representative model organisms for this microbial ecosystem.

Technology Development Box TD3_ME | Mass spectrometry-based functional analysis.

Recent technical developments in mass spectrometry (MS) have enabled the characterisation of cells and their environment at the molecular level. Proteomics, metabolomics, and lipidomics measure the composition, interaction, and modification of macromolecules and chemicals to provide direct insights into molecular mechanisms of organismal functions.


(A) The Savitski Team have developed cutting-edge methods for unbiased determination of protein state *in vivo* in microbial and host cells (Thermal Proteome Profiling; Savitski *et al. Science* 2014; Becher *et al. Cell* 2018), and **(B)** for disentangling global protein degradation and synthesis through multiplexed proteome dynamics profiling (Savitski *et al. Cell* 2018). **(C)** The Zimmermann Group uses a combination of high-throughput metabolomics measurements, massive parallelised microbial culturing, and bacterial genetics to process up to 10,000 samples at a time. This unique setting allows the systematic identification of novel functional units of microbial strains and their communities, such as, high-throughput metabolomics analyses of clusters of human gut microbes, based on their xenobiotic-converting activity (Zimmermann *et al. Nature* 2019). **(D)** The Alexandrov Team is world-leading in spatial metabolomics approaches. These approaches can be used to quantify and visualise the chemical environment shaped by microbial colonies and the molecular interactions both between microbes, and between microbes and the host (Watrous *et al. ISME J.* 2013).

EMBL has diverse expertise in MS-based technologies, enabling world-class research to understand life at the molecular level. These technologies are key to moving microbiome research from associations to identifying the underlying molecular interactions in these microbial ecosystems.



Dissecting Interactions and Properties of Microbial Communities

As microbiotas are more than the sum of their individual members, they manifest community properties that the individual species do not exhibit or possess. These collective behaviours are relevant to the stability of the community, and to the way it interacts with its environment – both in terms of how microbiotas perceive and process signals, and how they respond to them. For the gut microbiota, these behaviours are relevant in their responses to the environment (e.g. xenobiotics) and to pathogenic intruders, their ability to degrade food, and their interaction with the host. Yet their prevalence and the molecular underpinnings that lead to these responses are largely unknown.

 For an initial assessment, several EMBL groups performed a pilot experiment by comparing the sensitivity to 30 drugs of approximately 30 gut species in isolation or in a community setting. Communal phenotypes were frequent, with cross-protection (drug-sensitive species becoming resistant in the community setting) being more common than cross-sensitisation (Figure ME3B). Interestingly, these communal traits became less relevant at higher drug concentrations, indicating that communities have a robustness threshold against perturbations. As a proof of principle, EMBL identified the strains and enzymes providing protection against the drug niclosamide. In the future, this setup will be used to systematically probe different perturbations and communities.

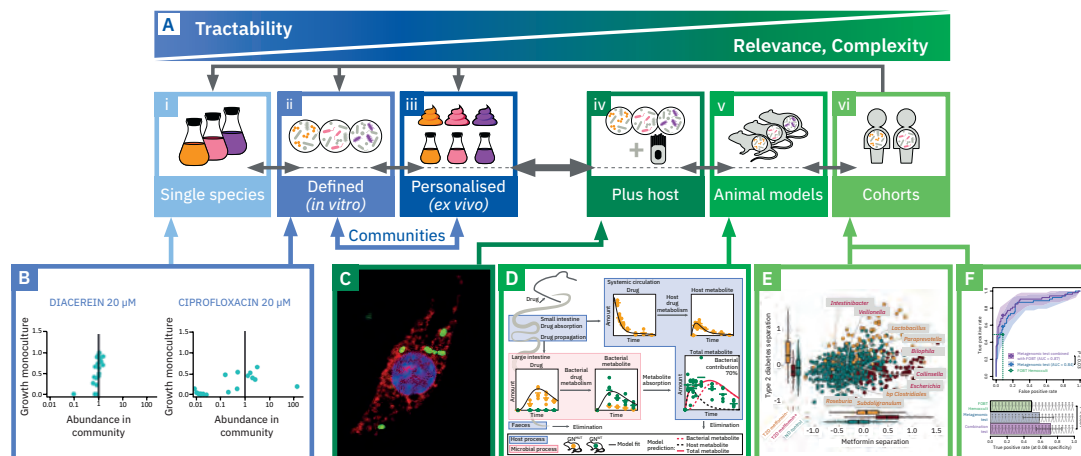


Figure ME3 | Experimental setups to study microbial ecosystems, from monocultures to microbiomes in clinical cohorts.

(A) Experimental approaches to study microbial communities, at various levels of complexity and tractability, that are employed by EMBL researchers (i–vi). (B) Single species behave the same in isolation as in a community (ciprofloxacin), or the community exhibits collective behaviours (diacerein – all strains become resistant) in the presence of a drug. (C) Intracellularly growing *Salmonella enterica* Typhimurium (green) leads to re-trafficking of active cathepsins (red) to the nucleus (blue) of macrophage cells (Selkrig *et al. Nature Microbiology* 2020). (D) Quantifying microbial contributions to drug metabolism *in vivo* using bacterial genetics, gnotobiotic mice, and pharmacokinetic models, which include drug absorption, GI tract propagation, host and microbial drug conversion, and systemic drug elimination (Zimmermann *et al. Science* 2019). (E) Human cohort study disentangles the effect of type 2 diabetes and metformin medication on patients' gut microbiota (Forsslund *et al. Nature* 2015). (F) The use of gut microbiome composition signatures as a biomarker to diagnose human colorectal cancer (Zeller *et al. Molecular Systems Biology* 2014; Wirbel *et al. Nature Medicine* 2019).

Understanding the Stability and Dynamics of Microbial Communities

Community-specific traits emerge through microbial interactions and are challenging to identify and prove solely through bioinformatics analyses of metagenomics data or experiments with bacterial isolates. Therefore, EMBL has recently been developing strategies to study microbial communities at various degrees

of complexity, from synthetic assemblies to *ex vivo* communities (Figure ME3; Tech Dev Box TD 2_ME). The former are artificially assembled communities, derived from mixing together individual strains, and the latter come from cultivating complex individualised communities directly from human stool. EMBL will amplify these efforts in the next Programme, with the goal being to **understand the impact of such interactions, their molecular nature, and their underlying general principles**. This will include studies of how interactions evolve over time and across dynamic environments; what their inter-individual variability is; and whether environments can foster or break communal behaviours, such as their intrinsic stability.

EMBL will focus on investigating the role of specific genetic variants, strains, and species in microbial interactions and community dynamics. Building on its established pipelines to systematically probe the gut microbe–medication interface, EMBL will expand to assessing the effect of nutrition, environmental changes, and xenobiotics such as drugs, excipients, pollutants, and food additives, on the stability and dynamics of different gut microbiome community assemblies. To do this, hundreds of communities in parallel with distinct, defined, and stable compositions will be combined with defined nutrients, environments, and perturbations in a high-throughput manner. These approaches will be coupled with high-content quantitative readouts to provide mechanistic insights into the underlying phenotypes. The results will provide a better understanding of how microbial and environmental factors shape community composition, and how emergent or collective phenotypic traits of these communities feed back to their environments; in this case, the human host. Ultimately this information will advance the ability to interpret microbiome data from clinical cohorts, and to understand the underlying associations to health and disease. It will also enable the development of predictive models of community dynamics and impact.

Dissecting the Role of the Gut Microbiota in Infection and Antibiotic Resistance

A comprehensive understanding of the molecular processes by which benign bacteria within the human microbiome become pathogenic, and the role of these dense communities in the development of antimicrobial resistance and the emergence of difficult-to-treat pathogens, are topics of high interest. The human gut has an intrinsic ability to fend off intruders (colonisation resistance), but at the same time harbours numerous pathobionts (non-harming symbionts that under specific conditions can become pathological) and opportunistic pathogens (pathogens that normally do not cause disease in a healthy host, but can do when opportunity arises) as regular inhabitants. Examples of the latter are enterotoxigenic strains of *Bacteroides fragilis* or *Clostridioides difficile*, a leading cause of diarrheal illness. The interface between the human microbiota in different parts of the human body (gut, skin, mouth, nasopharynx, lung, urinary tract, and vagina) and pathogenic subpopulations is complex and has been largely elusive to date.

A fundamental question is: what are the signals that trigger or prevent the transition of a silent resident pathogen to an active one, causing dysbiosis or even disease? Combining state-of-the-art bioinformatics and molecular biology, systems biology, and structural biology expertise, EMBL is in a unique position to reshape molecular infection biology and place it in the context of the holobiont (microbiome and host) and the environment. Research on the molecular players and mechanisms used by pathogens during infection lies at the core of infection biology and is widespread across sites at EMBL (Chapter 5: Infection Biology). These research themes range from characterising secretion machines and the molecules that bacterial pathogens use to hijack different parts of host biology, to monitoring pathogen physiology and signalling during infection. EMBL will take advantage of its ability to systematically investigate and perturb microbial communities to identify what opens up those windows of opportunity for pathogens, to understand if there are ways to better contain the enemies within, and to discover the effects of removing these species from communities, or exchanging them with avirulent variants. EMBL will also focus on understanding the molecular entities and machines that microbes use to fight pathogenic intruders and that pathogens use to counterattack – from small molecules to secreted proteins and from phages to large protein machines that act as weapons (Figure ME4).

The technologies available at EMBL enable these studies to be carried out across scales, from atomic resolution of isolated molecules to the visualisation of these molecules within whole cells.

Moreover, the high density and diversity of the gut microbiota (10^{13} – 10^{14} regular microbial inhabitants representing 3,000–4,000 species and many more strains, being attacked by 10^{14} – 10^{15} phages) facilitates the exchange of genetic material, i.e. horizontal gene transfer (HGT). This high density and the strong environmental exposures (such as new microbes coming with food or selective pressure from drug treatments) make the gut microbiome the perfect reservoir **for the development and spread of antimicrobial resistance (AMR)**. AMR genes are frequently carried on mobile genetic elements that can move within and between bacteria, promoting the development of difficult-to-treat superbugs in such hotspots of microbial interaction. Understanding AMR transmission is critical for human health and is a core part of research at EMBL (Chapter 5: Infection Biology). In the case of the gut microbiome, EMBL researchers are interested in using their expertise in bioinformatics, microbiology, genetics, and structural biology to chart how AMR spreads in these communities, mapping transmission routes, underlying mechanisms, and possible Achilles' heels (i.e. the fitness cost it confers to the community). This understanding is essential to identify strategies to limit or even prevent the development and spread of AMR to pathogens.

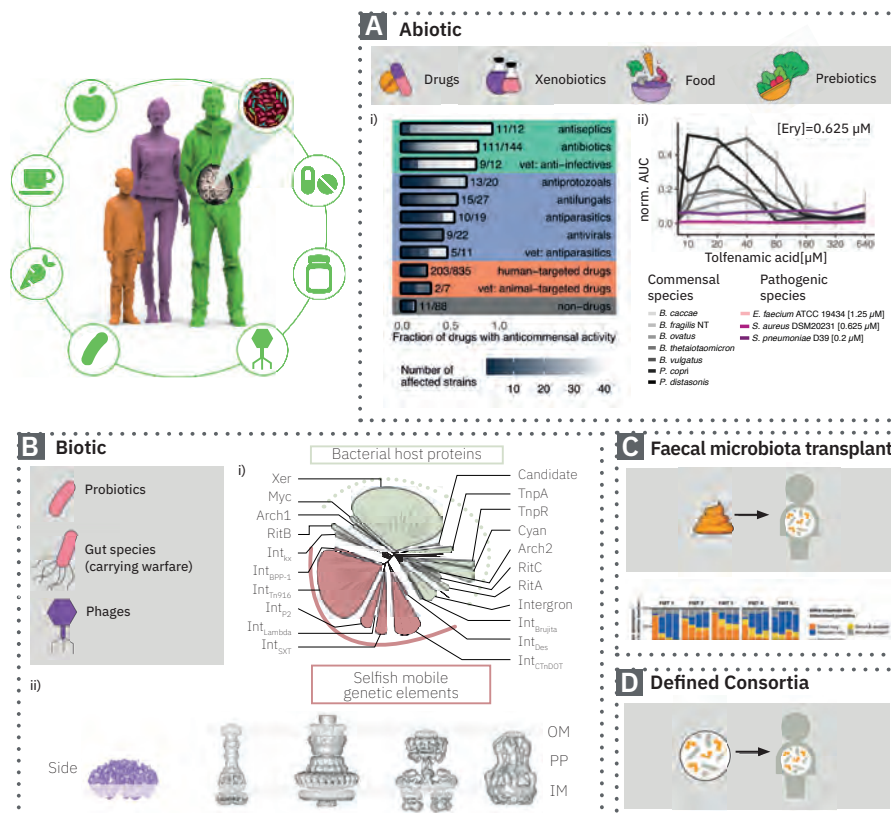


Figure ME4 | Towards a rational modulation of microbial ecosystems such as the human gut.

(A) Abiotics, such as drugs, food, and prebiotics, and combinations thereof, can be used to shift microbiota compositions: (i) many non-antibiotic drugs can directly inhibit specific gut microbes (Maier *et al. Nature* 2018); (ii) species-specific outcomes of drug combinations (Brochado *et al. Nature* 2018) can be used to rescue the collateral damage caused by antibiotics on gut microbes. Tolfenamic acid, a painkiller, alleviates the effect of the antibiotic erythromycin in several commensal species (black traces), but allows erythromycin to retain activity against pathogens (purple traces) (Maier *et al. bioRxiv* 2020). (B) Biotics, such as probiotics, phages, and species carrying secreted proteins targeting other bacteria or the host can be employed to shift bacterial community composition: (i) a systematic survey of genetic elements that can autonomously move between bacterial species reveals their diversity and evolution, and provides new tools for mapping HGT routes and traits (Smyshlyaev *et al. Biorxiv*, 2019); (ii) the first electron microscopy structure of a Type VII secretion system and its comparison with other known secretion systems (Beckham *et al. Nature Microbiology* 2017). (C) FMT can change the microbiota of the recipient: mapping success across individuals (Li *et al. Science* 2016). (D) Defined consortia can be used to enable targeted rational modulation of the gut microbiota.

Placing Microbial Communities in their Ecological Context

Microbial communities are largely shaped by interactions with their environment, with which they form a complex and dynamic ecosystem. To investigate these complex interactions and to separate causes from consequences, EMBL will use experimental settings that mimic the natural environment and probe the effects of the environment to the community and vice versa via multi-level molecular profiling. The goal is to gain mechanistic insights into the functional interplay between microbes and their environment, which in case of the human gut microbiome is the host. The generated results will expand the interpretation of available sequencing data and will eventually enable targeted manipulations of microbial communities. Furthermore, the general framework proposed will provide a roadmap to study other microbiomes, and how molecular signals and pressures within these microbial communities shape their composition and interactions within the ecosystem.

Impact of the Functional Microbiome Output on the Host

To understand how changes in microbial community composition and physiology affect the host, EMBL will use *ex vivo* host models, such as human- and animal-derived cell lines, primary cells, and organoids (Figure ME3). These simplified host systems will be exposed to microbial extracts or communities. Tailored host readouts (reporter assays, high-throughput microscopy, multi-omics, and FACS with immune markers) will be used to assess the impact of synthetic microbial or *ex vivo* communities. This line of experimentation will enable the identification of causal effects of the microbiota on specific host responses, such as neurotransmitter secretion by enteroendocrine cells, stimulation of immune cells and drug responses of different cell types. In addition, it will point to the perpetrator species and strains, and with the help of bacterial genome-wide mutant libraries and multi-omics readouts will reveal the underlying molecules, genes, pathways, and mechanisms of these interactions. The experimental setups developed here will link to complementary approaches used for mapping the effect of the microbiota on drug responses of cancer cells (Chapter 6: Human Ecosystems).

Reciprocal Microbiome–Host Interactions in Complex Models

The host plays a crucial role in shaping its colonising microbial communities and the multifactorial impact of the host on these microbes is hard to fully reproduce *in vitro*. EMBL will therefore employ gnotobiotic animals colonised with communities at various complexity levels (including genetically engineered marker strains or genetically barcoded libraries of a specific species) to carry out spatially resolved assays of compositional adaptation to dietary, chemical, or infectious perturbations (Figure ME3). These data will be paired with omics and physiological measurements, allowing direct comparisons with results from the more controlled *in vitro* settings described above.

The microbiota also influences the host physiology at various levels, which are invisible without the use of whole organisms to determine behaviour, reproduction, and organ function. EMBL will exploit its resources and expertise in mouse biology to re-derive germ-free and/or genetically modified mouse strains. This can include humanised mice (mice with transplanted human gut microbiota), and specific target knockout mice such as those lacking liver CYP450 enzymes, which results in a compromised drug metabolism. Mouse models and readouts will be selected to investigate microbial community influence on drug responses, cancer development, the germline, the liver, and the gut–brain axis. The knowledge from the *in vitro* and *ex vivo* data, especially on causal genes, strains, and communities, will facilitate the prioritisation of experiments in genetically modified mice. The developed models and the data generated will be instrumental to both reproducing interactions seen in *ex vivo* data, and testing associations drawn in large human microbiome

cohort studies. Ultimately, mirroring the effect of microbiota on health and disease in appropriate animal models will not only help scientists to identify causal effects and mechanisms, but will also provide a solid basis for testing intervention strategies for rational modulation of the composition of the gut microbiota.

Modulating Microbial Communities and their Interactions

The microbiome has become a primary therapeutic target, with thousands of clinical trials currently taking place globally. Most of these studies aim to modulate the microbiome towards a health-promoting state for its human host. The means to do so vary immensely, from antibiotics or phages to eliminate community members, to prebiotics to promote certain strains of beneficial bacteria, and from probiotic strain cocktails to complement the community with selected members or traits, to faecal microbiota transplantations (FMT), which involve transplanting a sample of the microbiota from one person to another (Figure ME4). However, these interventions are mostly based on empirical observations of particular microbiome properties, lacking systematic insight into the repertoire of underlying factors and molecular mechanisms that impact the outcome of the intervention and dictate its success. Hence most of these trials are bound to fail or have reproducibility issues when the size of the study increases and the inter-individual complexity and variability of these communities increases. Therefore **the rational design of microbiome modulation strategies remains the pinnacle in translational microbiome research**, which aims to develop therapeutic strategies to treat infections, allergies, inflammatory conditions, metabolic syndromes, cancer, and many more dysbiosis-associated human diseases. A sufficient understanding of microbial processes, in their interactions with each other and the host, are the basis for moving towards more controlled strategies for modulating the gut microbiome. EMBL proposes to combine bottom-up and top-down approaches to achieve this. Bottom-up approaches will be based on mapping and predicting interactions between molecules and microbes within the community context, such as those that occur via mobile elements or phages, and interactions between microbes and the host. Top-down approaches will rely on computational analysis of complex data from clinical microbiome intervention studies to identify patterns associated with successful interventions or to validate findings from defined community settings.

Precision Modulation of Communities

EMBL will utilise its automated high-throughput pipelines to systematically assess the effects of simple **abiotic** modulators, such as chemicals, prebiotics, and food compounds or additives, or combinations of these, on single microbes and communities (Figure ME4A). This will provide foundational knowledge about how to specifically suppress or promote the growth of certain species or to alter community compositions. Such modulations could be minimal, with the overarching goal being to remove specific strains while retaining the overall community composition intact. For instance, replacing an AMR pathobiont by its cured derivative would increase treatment options if this becomes pathogenic for the human host. The use of *ex vivo* communities, strain collections, and genome-wide mutant libraries, together with the ability to rapidly run millions of such experiments, will allow EMBL researchers to systematically assess the role of strain diversity and genetic traits on the outcome of such perturbations. Using machine learning approaches on such rich experimental data will make it possible to predict modulators and responding bacterial species and strains, as a first step towards rationalising more complex modulations.

Investigation of single-agent live microbiome (**biotic**) modulators, such as probiotic bacterial strains or phages, will require first pinpointing the microbial genes and metabolites involved in their interactions with microbial community members or the host. Bioinformatic analyses based on protein homology and genomic organisation can expedite the identification of secondary metabolites, toxins, secretion, and adhesion machines. The increase in genomic sequence information fuelled by large-scale metagenomics assemblies

will further facilitate such strategies, especially ones that use genomic context information to fish out novel microbial weapons that bear no resemblance to previous systems. EMBL has been pioneering this area of research with multiple databases and tools (InterPro, eggNOG, Pfam, STRING) that map protein families or genomic context information and use it to catalogue or identify new functions. These resources and tools will continue to develop in the next EMBL Programme, and aim to take advantage of and cater to the continuing increase of microbiome metagenomics data. *In silico* predictions will directly interface with experimental approaches, which aim to systematically map the effects of bacterial strains, species, and communities on the composition of personalised microbial communities (*ex vivo* communities from different individuals). As bigger parts of the microbial armoury are unravelled, selected molecules, such as proteins or metabolites, will provide candidate agents for microbiome modulation. Experimental follow-up combining omics and analytical technologies will provide mechanistic insights into interesting modulating agents.

Another potential area of study involves **mobile genetic elements** and lytic **phages**, which have been largely unexplored as microbiome modulators. Due to the spread of AMR, the interest in using phages to treat pathogenic infections has increased, with some clinical trials currently ongoing. There are many potential advantages of phage treatment over antibiotics, such as high specificity, low inherent toxicity, and minimal disruption of surrounding tissues or normal flora. However, there are also many issues that remain to be resolved, such as delivery to the infection site or resistance development. What EMBL can contribute to the field at this stage are improved bioinformatics tools for mapping and characterising both mobile genetic elements (Figure ME4B) and the vast phage arsenal present in the human microbiome and in environmental microbiomes. This information will improve understanding of the resistance reservoirs carried within the microbiotas of individuals and the specificity of phage treatments as microbiome modulators.

Whole Community Transplantation

Faecal microbiota transplantations (FMT) are confounded by various factors that affect the success rate of this therapy. EMBL will employ data-driven approaches that involve the analysis of microbiome data from ongoing and past clinical interventions across multiple indications. Previous such work by EMBL (Figure ME4C) of strain dynamics following FMT will be complemented with data from interventions involving complex dietary shifts or phage cocktails. The goal will be to deduce the principles by which gut bacterial taxa or communities are resilient to such perturbations and associate the characteristics of modulations (such as the FMT donor) with success rates. This effort will be empowered by our increased understanding of microbe–microbe and microbe–host interactions. Overall, these approaches will provide a framework for testing any general characteristics that underlie success rates, both in *in vitro* or *ex vivo* communities or animal models.

Understanding how individual human host differences factor into the outcome of FMTs, together with a molecular understanding of microbial communities and ecosystems, will enable predictive modelling of microbiome modulations using defined microbial consortia. These interventions will bestow desired traits and expected molecular impacts on the targeted gut community, such as the elimination of pathogens or microbes carrying AMR. This modelling attempt will not only elucidate what might have to be supplemented to the recipient's microbiome or which pathogens are to be displaced, but will also identify the microbial consortia required to achieve the desired effect. The resulting predictions can be tested using synthetic communities and faecal *ex vivo* cultivation in animal intervention experiments, yielding results that can be used to further improve the models.

Expanding and Translating to Other Microbial Ecosystems

Although EMBL's track record of studying microbial ecosystems is focused primarily on the gut microbiota, diverse microbial communities are found in other parts of the human body, including the entire digestive and respiratory systems, the skin, reproductive organs, and several internal organs. In addition to bacteria, there is a plethora of fungi, phages, and protists within these communities. Over the past few years, advances in sampling and bioinformatics analyses have led to an increased appreciation of the diversity and roles of these less-studied organisms within microbial communities. The computational and experimental approaches pioneered for gut bacteria will act as a roadmap to facilitate the cataloguing and functional dissection of other microbiota members and microbiomes. EMBL will foster research in these areas, aiming to gain a better understanding of the collective role of microbes in human health, the degree of interplay between the different human microbiotas, and the role of the environment in shaping microbiome composition and phenotypic traits.

Humans are exposed to microbes and microbiotas at every step of their lives – through the food chain, social interactions, and via their natural and built environments. Scientists are only now starting to understand the microbial exposome in a quantitative and molecular way, including what it is, its influence on humans and life on the planet, and how much human activities change microbiome composition and behaviours. The computational tools for mapping this microbial diversity and its functional underpinning will be key for moving towards a better understanding.

As EMBL's experience in cultivating microbial species increases, and in collaboration with European institutes that have decades of expertise in specific microbial ecosystems, EMBL will create automated experimental setups to map species and functional diversity for defined microbial communities outside of the human microbiome. This includes building strain collections, frameworks for systematically monitoring perturbations and microbial phenotypes, and tailoring these advanced molecular technologies to be applicable to the study of these systems. Employing these approaches, EMBL and collaborators can start assessing the principles (e.g. stability, communication, competition versus cooperation, emergent behaviours, gene transfer) that define the organisation of these microbial communities and facilitate niche specificity. Integration of this information with geography and ecological measures will be the next challenge to conquer for big data and machine learning-based approaches. The ultimate goal will be to understand the flow of information and the overarching principles that govern stability of microbial ecosystems and (evolution of) life on Earth.

Impact

Of all microbial ecosystems, the human gut microbiome is currently by far the most studied. However, despite intense research there are still fundamental gaps in knowledge of microbial genes, community behaviours, and microbial interactions. EMBL proposes a plan to systematically fill these gaps through a coordinated effort, combining cutting-edge computational and experimental approaches. Many aspects of human physiology are linked to the gut microbiome, including obesity, immune system function, and even mental health. EMBL's unique technology platforms will form essential and foundational resources for the scientific community, in member states and beyond.

The molecular insights gained from the approaches EMBL will undertake in the next Programme will ultimately help scientists understand how genetic and environmental factors shape microbial community composition, the collective phenotypic traits of these communities, and their impact on the environment. Ultimately, this information will enable scientists to rationally design therapeutic interventions targeting the microbiome to provide health benefits for individuals and society. The developed approaches, tools, resources, and data integration frameworks will pave the way for future studies of other microbial ecosystems, such as ones found in the natural or built environment, the composition and diversity of which we are only now beginning to map.

5. Infection Biology

Background

Infectious diseases are among the most prevalent causes of human illness and death in the world. Global warming, a decrease in biodiversity, antimicrobial resistance (AMR), modern lifestyles (travelling habits, high urban density, mass food production, and globalisation), and socio-economic factors, such as disparities in public health, make combating infectious diseases a more urgent challenge than ever. Besides humans, infection impacts all life forms on Earth. Pathogens can cross species boundaries, creating alarming possibilities for future epidemics and adding to rising concerns about changes in biodiversity and planetary health. Further increasing the challenge is the diversity of pathogens, which include viruses, bacteria, fungi, protozoa and other parasites. Within each of these groups, there is staggering diversity, both in their fundamental biology and ecology.

Disease mechanisms and transmission routes are also varied. For instance, both bacteria and viruses use diverse mechanisms to evade their host's immune response and to hijack the host's cellular processes to use them for their own benefit, while residing outside or inside different cell types. Many prominent infectious diseases in humans represent zoonoses showing successive propagation in diverse susceptible new hosts. These include HIV/AIDS, influenza, Ebola virus disease, plague, salmonellosis, and most recently SARS-CoV-2. Other pathogens (e.g. Epstein-Barr virus or *Mycobacterium tuberculosis*) persist in the host and can establish a dormant infection. In all infectious diseases, the host response is key to disease development and manifestation. Pathogens have developed a wide range of mechanisms to modulate both the innate and adaptive parts of host immunity to facilitate infection and transmission. In many infectious diseases, it is the failure of the immune system to respond appropriately that is a substantial cause of mortality.

Human pathogens now pose a greater threat due to their increasing resistance to available therapies. Multidrug-resistant (MDR) pathogens are found in rapidly growing numbers worldwide, increasing mortality caused by once treatable diseases and reducing our quality of life. Bacterial pathogens account for more than 3.5 million deaths per year globally, with an increasing fraction coming from MDR infections. Similarly, in more than 10% of adults living with HIV, the virus has developed resistance to the main first-line antiretroviral drugs (nevirapine and efavirenz), according to the World Health Organization. Recently, both WHO and the UN have declared antimicrobial resistance (AMR) a major threat to global public health. Awareness has finally permeated all layers of society, but appropriate action on the development of effective anti-infective strategies is still lacking. For instance, only one novel broad-spectrum antibiotic class has entered the market since the 1990s. The repeated failures of target-driven drug discovery have led most big pharmaceutical companies to close their antibiotics R&D departments. This, together with the rapid development of resistance to existing compounds, creates a highly challenging situation.

The increased risk of transmission of disease from wildlife to humans and between humans is likely a hidden cost of human economic development and disruption of ecosystems, which has catalysed the (re-)emergence of viral pathogens such as Zika virus and SARS-CoV-2. With our ability to treat these serious global public health threats being challenged, infectious diseases can re-emerge as a major factor of human morbidity and mortality, even in countries with developed public health systems. Insights into pathogen spread, evolution, and transmission, as well as characterising and understanding the host response, are critical to tackling infectious diseases (Figure IB0).



Figure IB0 | Mapping infection biology.

The study of infection biology needs to take into account multiple factors such as the unique mechanisms of various pathogens (e.g. viruses, bacteria, fungi, and protozoa), the host organism (e.g. bats, pigs, and chickens), and situational contexts (e.g. farmed land, food chain, and urban areas). EMBL aims to combat this at several biological scales (e.g. from molecules to communities) with a view to impacting a number of applications (e.g. antimicrobial resistance, diagnostics, drug discovery, and outbreak surveillance).

The Opportunity

With infectious diseases having immediate and devastating consequences globally with long-term repercussions, research on the biology and mechanisms of infection, as well as on diagnostics and treatment, is vital and urgent. Pathogen biology and disease are fundamentally linked to the interaction with the host, which can take on many perspectives, from the conflict with the immune system to the parallel genomic evolution of pathogens on different timescales and under different selective pressures (e.g. drugs, immune system, and competing microbes).

Technology developments over the past decades provide an increasing array of atomic, molecular, cellular, and physiological assays that can be performed on pathogens, on cellular systems modelling humans, or on humans directly. This produces remarkably rich datasets, which can be used to probe the human (or other host) response to infection. The opportunity awaits to understand the molecular machineries of pathogens and the host response in more detail than ever before, to discover new interactions at the host–pathogen interface, and thereby to provide new means for treating infection.

To tackle the AMR crisis, an improved molecular understanding of host–pathogen biology is not enough. This needs to be coupled with more specific diagnostics methods to optimise the use of existing antimicrobials, a better understanding of the mechanisms of AMR selection and spread to devise innovative resistance prevention strategies, and new antimicrobial therapies which are less prone to resistance development. Modern technologies can enable effective tracking of pathogens between hosts, across national boundaries and ecological habitats (Chapter 7: Planetary Biology).

The SARS-CoV-2 pandemic has revealed the urgent need for transparent, secure, and scalable sharing of pathogen data for the global management of infectious disease. Pathogens do not respect national, ecological, or species boundaries. Thus, infectious diseases of humans and animals are studied across a wide range of scales, ranging from longitudinal studies of single infections or individual outbreaks to indefinite global surveillance, host context, and historical and evolutionary perspectives. Currently, scientists have limited tools to access structured infectious biology data. Unified and systematic views into pathogen biology, allowing the user to navigate the many different dimensions of datasets, curated metadata, and literature will be essential to allow scientists, public health agencies, and clinicians to connect, access, and use this information in a meaningful way, so as to fundamentally change the way infectious diseases are understood and treated.

Research Aims

For several decades, EMBL has undertaken significant research activity on selected pathogens and their interaction with their host. Much of this work has been driven by cellular and structural biology, which has unravelled detailed mechanisms of fundamental interest, as well as potential drug targets. More recently, high-throughput genetics- and proteomics-based approaches, as well as cutting-edge imaging technologies, have enabled more systematic views of the host–pathogen interface. In parallel, a concerted effort to tackle the AMR crisis has been forming. In the new Programme, EMBL will integrate multidisciplinary experimental and computational approaches to dissect the complex molecular mechanisms behind the interplay of pathogens and their host ecosystems:

- I. *In situ* information about the cellular organisation of pathogens and the structural basis of their essential biological processes can give insight into new drug targets and inspire therapeutic strategies. With a **focus on pathogen-specific protein machineries**, EMBL aims to extend beyond *in vitro* structure determination into visualising these machineries in their cellular environment, using cryo-electron tomography (cryo-ET) and other techniques that build on EMBL’s strength in integrating molecular information across resolution scales.
- II. The mechanisms of pathogen adhesion, invasion, and effector translocation in the host are still poorly understood. Equally elusive remains the understanding of the targets and mechanisms of effector proteins which hijack diverse host cellular processes to manifest infection. EMBL aims to use its diverse cutting-edge technologies to **systematically map, view, and model host–pathogen interfaces** at the atomic, molecular, and tissue level. This will enrich the understanding that scientists have of diverse infection processes, opening avenues for new anti-infectives and immunomodulatory strategies.
- III. Capitalising on the growing volume of population-scale human biological datasets, EMBL aims to develop sophisticated statistical models to **map the human genetics of infection susceptibility**. These models can be deployed to combat endemic and epidemic outbreaks, for example in the context of national data within EMBL member states. In collaboration with

clinicians, EMBL will also employ multimodal approaches for cellular phenotyping of patient samples and machine learning methods to characterise the human response to infection.

- IV. To improve treatment efficacy, reduce side-effects, and prevent the development of drug resistance, EMBL proposes to develop individualised, microbiome-based diagnostic and surveillance tools. EMBL will combine predictive models, automated screening, medicinal chemistry, and structural biology to identify new anti-infective targets and compounds. Systematic genetics and proteomics approaches will expedite drug mode-of-action identification, which is a current bottleneck in antimicrobial drug discovery. EMBL further aims to **create a better understanding of antimicrobial resistance mechanisms and transmission paths** in patients and in the environment, so as to devise **new ways to delay, prevent, or revert this resistance**.
- V. EMBL has a long history in collating and openly sharing genomic and proteomic data of pathogens to promote fundamental science. In recent years, EMBL has also started working closer to the frontline of public health, collaborating with public health agencies on a platform for secure and controlled data-sharing for pathogen genomics and AMR surveillance. This includes the provision of **international data hubs** for the SARS-CoV-2 pandemic. EMBL will develop further **data platforms, computational frameworks, and infrastructure** to allow scientists across multiple domains to access and process multi-dimensional pathogen data. These resources will support rapid response to future outbreaks of novel pathogens and facilitate target identification for anti-infective research.

Given EMBL's unparalleled resources, technologies, and multinational research networks as well as its broad relevant expertise from informatics to molecular and systems-level sciences, EMBL is in a unique position to provide new knowledge and innovative solutions to combat microbial pathogens.

EMBL's Approaches

Pathogen Specific Molecular and Cellular Machineries

Pathogens employ a range of specific protein machineries in their replication and invasion of host cells. Several examples are described below to illustrate how EMBL aims to move beyond *in vitro* structure determination to **visualise these dynamic machineries** in their cellular environments.

Replication Machinery of RNA Viruses

Many important human viral pathogens are **RNA viruses**, including Ebola, measles, rabies, SARS-Cov-1 and 2, Dengue, Zika, hepatitis C, polio, rhinovirus, rotavirus, HIV, influenza, and Lassa. All RNA viruses, except retroviruses, encode an evolutionarily related RNA-dependent RNA polymerase (RNAP) that is able to transcribe and replicate the RNA genome. As the key viral transcription/replication machine, the viral RNAP is a primary target for antiviral drug development.

EMBL researchers aim to obtain a detailed structure-based understanding of how the **influenza virus RNAP** functions, and specifically the distinct processes of transcription and replication of the segmented, negative-sense RNA genome. In the past, crystal structures of the 270 kDa heterotrimeric RNAP have made it possible to pioneer a new generation of anti-influenza drugs that target unique features of the influenza RNAP. Most

recently, multiple high-resolution cryo-EM snapshots of all stages of transcription by the influenza RNAP – initiation, elongation, termination/polyadenylation, and recycling – have been determined, leading to a model of the complete transcription cycle with numerous novel features (Figure IB1 A–C). Similar methods can be used to investigate the functional dynamics of RNAPs from other emerging viruses, including arenaviruses, bunyaviruses, and coronaviruses.

In the future, the main aim is to go beyond *in vitro* mechanisms. In an influenza-infected cell, for example, transcription and replication of the viral genome occur in the context of viral ribonucleoprotein particles (RNPs), in which most of the RNA is coated by protective viral nucleoprotein in the nucleus, aided by specific host factors. The goal will be to understand the mechanisms operating at these two extra levels of complexity: (i) how are nucleoproteins dynamically remodelled around the RNAP during transcription? and (ii) how is the influenza RNAP able to robustly pirate nascent host transcripts to prime its own transcription, a process known as cap snatching? With the rapid development of cryo-ET workflows, visualising viral RNPs in the chromatin context will likely become feasible. Additional lines of investigation include applying techniques developed to probe the structure of long non-coding RNAs to study the structure of viral RNA such as the 5' untranslated region of SARS-CoV-1, which possibly functions as an internal ribosome entry site to initiate protein translation on the human ribosome.

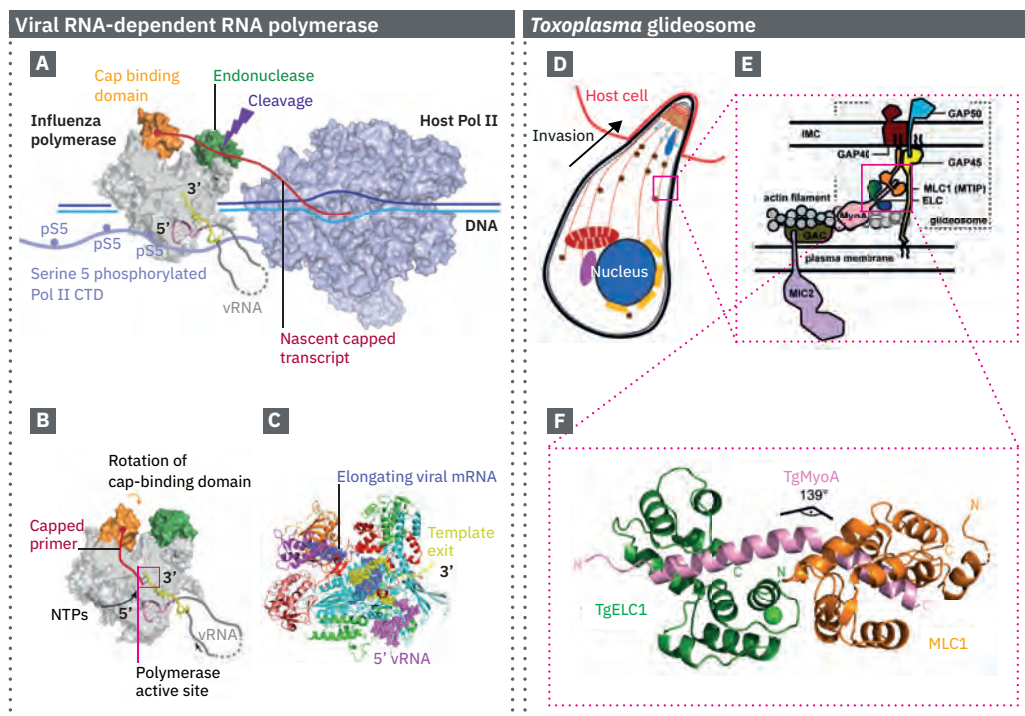


Figure IB1 | Molecular machines: the influenza polymerase and the parasitic glideosome.

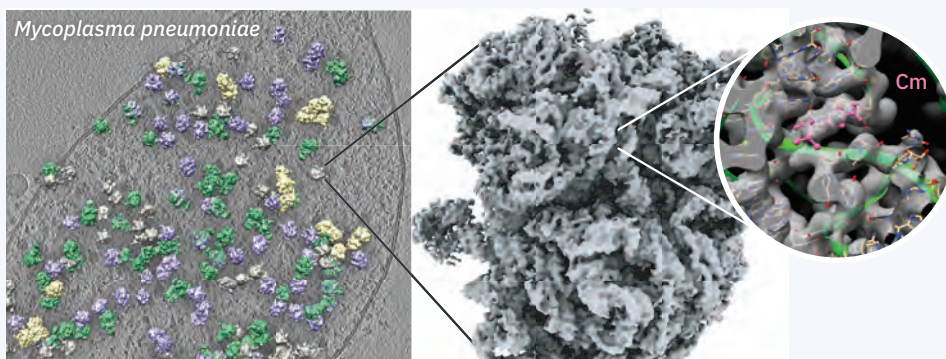
(A) Schematic of ‘cap-snatching’ by influenza RNA-dependent RNAP (Flupol). In the infected cell nucleus, FluPol (grey) associates with host RNAP II (Pol II – light blue) by binding to its phosphorylated CTD. This gives FluPol access to nascent capped transcripts (red) that it cleaves with its endonuclease activity (green domain) and uses to prime transcription of viral mRNAs. The chromatin environment in which FluPol robustly manages to compete at the right moment for capped RNAs can be visualised by cryo-ET in infected cells. **(B)** Schematic of cap-dependent transcription initiation. **(C)** *In vitro* cryo-EM structure of FluPol actively elongating viral mRNA. **(D)** Schematic depiction of the invasion of an apicomplexan parasite into a red blood cell. **(E)** Schematic representation of the current model of the glideosome. Actin is immobilised to the plasma membrane whereas myosin A is part of the glideosome, which binds the essential light chains ELC and myosin light chain MLC1. Myosin A and its light chains further interact with glideosome associated proteins GAP40, GAP45 and GAP50, which anchor the glideosome in the outer membrane of the inner membrane complex. **(F)** High-resolution X-ray structure of a trimeric sub complex (MLC1, ELC1, fragment of MyoA) of the glideosome.

Molecular Machines *In Situ* Within *Mycoplasma pneumoniae*

Pathogenic species of *Mycoplasma* are involved in a number of diseases targeting the human respiratory system. One of these, *M. pneumoniae*, has become a model for systems biology and multi-omics approaches, due to its very small genome size. EMBL researchers have used *M. pneumoniae* to establish a workflow that allows key molecular machineries to be visualised in action at near-atomic resolution inside the cell by cryo-electron tomography (Tech Dev Box TD1_IB). This has yielded the first *in vivo* structural view of a central supramolecular complex that directly couples transcription and translation, a mechanism unique to bacteria. Building on these technical advances, the next tier is to investigate membrane protein regulation and organisation. Although many of the pathogenic activities of *M. pneumoniae* are localised to the cell membrane, the organisation of the membrane proteome is largely unknown. EMBL will aim to structurally elucidate the organisation of the *M. pneumoniae* membrane proteome using integrated in-cell cryo-ET and proteomics approaches. As a similar level of resolution has not been achieved for *in situ* imaging of any organism to date, EMBL's approach will likely reveal general principles of membrane proteome organisation.

Technology Development Box TD1_IB | Towards creating whole-cell near-atomic resolution structural models.

Imaging is a powerful tool for both fundamental research and biomedical diagnostics. The Mahamid Group has been pioneering cryo-electron tomography (cryo-ET) pipelines that enable the visualisation of an entire cell of a bacterial pathogen, *Mycoplasma pneumoniae* (left, background). Focusing first on the ribosome (left, foreground; middle), the protein production machinery of all living cells and a major antimicrobial target, has already yielded fundamental mechanistic insights into its coupled function with RNA polymerase in bacteria (O'Reilly *et al. bioRxiv* 2020). Albeit a model system with a reduced genome, *M. pneumoniae* still encodes 688 proteins. Assigning each protein and DNA and assembling protein complexes like in a puzzle in these electron microscopic images will allow composition of the first full 3D molecular picture of a cell. This is a computationally intensive process but has immense implications both for basic biology and for novel screening concepts. For example, treating this human pathogen with chloramphenicol, a well-known protein synthesis inhibitor, reveals the precise binding site of the drug on the ribosome (right) *in situ*. As more protein complexes are mapped inside the cell, this setup can be used as a screening platform to identify drug targets and their off-targets in the cell, as well as to view the downstream cellular responses they elicit at a nearly atomic resolution.



Cryo-ET slice of a *M. pneumoniae* cell (left), overlaid with annotated ribosomes: grey, single ribosomes; green, ribosomes interacting with putative RNA polymerase; yellow, ribosome assemblies engaged in protein production from a single mRNA; purple, a super complex of ribosomes directly engaged with RNA polymerase, unambiguously demonstrating for the first time transcription-translation coupling *in vivo*. A high-resolution reconstruction of the *M. pneumoniae* ribosome (middle) at near-atomic resolution (3.5 Ångström) with a zoom into the ribosome active site (right) where chloramphenicol binding (pink) stalls protein production.

Unique Gene-expression Systems in Trypanosome Parasites

The majority of human African trypanosomiasis (sleeping sickness) is caused by the *Trypanosoma brucei gambiense* and is transmitted by the tsetse fly. Trypanosomes possess specialised RNA processing pathways, which are distinct from those of their mammalian hosts and hence provide ideal drug targets. EMBL researchers will study the molecular basis and RNA interaction dynamics of two key trypanosomal RNA processing pathways. First, the mitochondrial RNA (mtRNA) editing pathway will be targeted with structural biology approaches and parasite cell biology. mtRNA editing in trypanosomes is conducted by the enzymatic RNA editing core complex together with a second protein complex, the RNA editing substrate binding complex (RESC). No structural information on RESC or any of its subunits is available. The aim will be to use structure-function approaches to understand the assembly mechanism of RESC, its composition, its specificity for mRNA and gRNAs, and its role in mtRNA editing. A second unique process in trypanosomes is trans-splicing, which separates the functionally unrelated coding clusters with several RNA–protein complexes that are unique to trypanosomes. EMBL groups will characterise these dynamic complexes in atomic detail and will explore their function in trans-splicing. This objective is highly relevant with respect to future drug targeting, as trans-splicing and the related complexes are essential for the parasite.

The Apicomplexa Invasion Machine (Glideosome) of *Plasmodium*

Plasmodium species are intracellular, parasitic single cell eukaryotes of the family Apicomplexa and are the causative agents of malaria, which causes ~400,000 deaths per year. Another apicomplexan parasite, *Toxoplasma gondii*, is responsible for toxoplasmosis in humans. Proliferation and transmission of these obligate endoparasites in their host organisms rely on efficient cell invasion. This active process is based on parasite motility that is referred to as gliding and is empowered by an actin/myosin motor. This motor is localised within the intermembrane space between the parasite's plasma membrane and inner membrane complex (IMC), an additional double layer of membranes that is unique for these single-cell organisms. While motility is achieved by the interaction of the myosin with actin filaments, the myosin is linked to the IMC by a membrane-embedded multi-protein complex referred to as the glideosome (Figure IB1D-F). The structures of a few individual glideosome components are known; however, how these proteins assemble into an active complex and perform their function remains elusive. To obtain structural and functional understanding of the glideosome, two approaches will be followed: i) individual glideosome components, subcomplexes of known intermediate assemblies, and the entire complex of all known glideosome members will be structurally and functionally characterised using biochemistry, X-ray crystallography, small-angle X-ray scattering (SAXS), and cryo-EM; and ii) analysis of the endogenous membrane complex will be analysed by mass spectrometry to identify so far unknown glideosome components followed by single-particle cryo-EM. This work will help to understand the underlying mechanisms of motility in apicomplexan, which is crucial for any invasion process of the parasite.

Systematically Mapping and Modelling Host–Pathogen Interfaces

Mapping Host–Pathogen Interfaces

EMBL will combine systems-based approaches, computational biology, and structure–function analyses to map the machines and the effector arsenals of pathogens, their host targets, and the mechanisms of action that pathogens exploit to survive and proliferate in the intracellular context.

The protein machines that bacterial pathogens use to adhere to cells and translocate effector proteins into the host cell have been a focal point of infection biology for decades. Effector proteins are used to usurp host defenses and to hijack various host cellular processes so that the pathogen can survive and proliferate. Past EMBL research has focused on structural analysis of a specialised secretion system of *Mycobacterium tuberculosis*, which secretes a number of key virulence proteins required for infection (Chapter 4: Microbial Ecosystems, Figure ME3B). In the next EMBL Programme, EMBL will continue to contribute cutting-edge structural insights into the **function of diverse secretion systems**, including type III systems, which cross even the host membrane to translocate effectors directly inside the host. These efforts will be complemented by computational analyses (e.g. Hidden Markov Model-based sequence analysis and machine learning approaches) and structure-function follow-ups to identify, classify, and understand the diverse protein domains that pathogens use to adhere to host surface-exposed sugars, lipids, and proteins, the so-called adhesins. The identification of conserved species-specific adhesins can be used as likely vaccine targets. Together with collaborators, EMBL's next goal is to provide a better overview of the structural and functional diversity of secretion apparatuses and adhesion protein families across different pathogens. Going beyond classical structural biology, the aim will be to visualise entire systems *in situ* using cryo-ET and correlative light and electron microscopy (CLEM) approaches.

Furthermore, machine learning and bioinformatics will be combined with proteomics-based labelling approaches to systematically identify the **secretion arsenals** of intracellular pathogens, which can reach more than 300 proteins for some species. EMBL will also employ a suite of computational and experimental approaches to **map the host targets of viral, bacterial, and eukaryotic effectors** (Figure IB2A; Tech Dev Box TD2_IB). Proteomics-based approaches (e.g. affinity purification coupled with quantitative mass spectrometry, thermal proteome profiling (TPP), proteomic phage display, and phosphoproteomics and ubiquitin proteomics) in *in vitro* infection contexts, and integrative computational approaches (homology models, structural information on interfaces, and human protein–protein interactions) will be used to identify host targets of effector proteins. Altogether, these efforts will provide a comprehensive understanding of how diverse pathogens interact with their host, revealing points of convergence and divergence in pathogen biology.

A common strategy for all intracellular pathogens is to **intercept host signalling pathways** and use them for their own benefit, for example to reorganise lipid transport and exploit cytoskeletal elements to build protective intracellular compartments; to avoid lysosomal killing and to move within the cell; to blunt immune responses; or to deregulate cell death pathways. To do this, pathogens utilise a myriad of approaches, from effector proteins with intrinsically disordered regions that mimic binding motifs of signalling proteins to effector proteins that carry their own signalling enzymatic activities (e.g. kinases, phosphatases, E3 ligases, DUBs, Rho GTases, and GEFs). Interestingly, within this vast repertoire of activities, pathogens have come up with novel enzymatic activities for post-translational modifications (PTMs), such as serine ubiquitination and glutamylation, to control the activity of host signalling cascades (Figure IB2B). EMBL will utilise its unique computational and experimental expertise on signalling networks and PTMs (Tech Dev Box TD2_IB), such as phosphorylation, ubiquitination, and Rho signalling, to dissect the mechanisms that pathogens use to intercept host signalling networks and the structural and functional ramifications those PTMs have on the human proteins. Thereby, EMBL researchers will not only gain a better understanding of both host and infection biology, but will also identify new intervention points for drugs and tools for bioengineering.

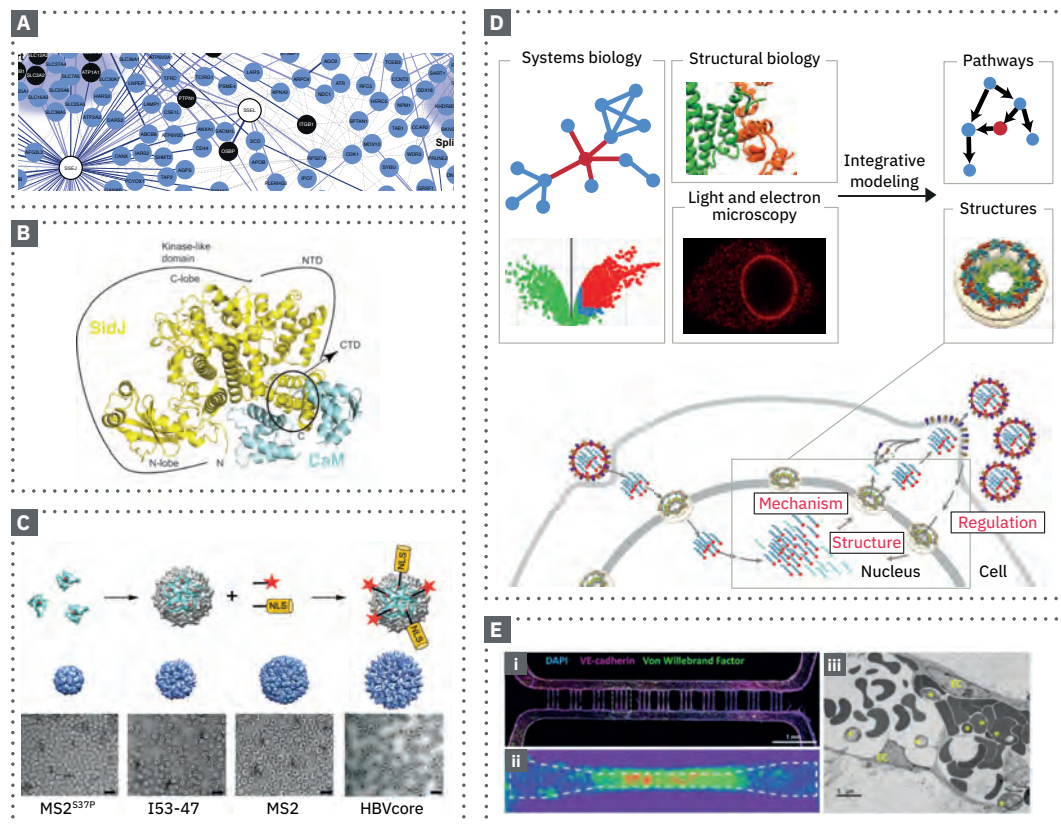



Figure IB2 | Probing host–pathogen interfaces.

(A) Systematically mapping physical interactions of secreted *Salmonella* effector proteins with their host targets during infection; effectors cooperate to hijack host processes (Walch *et al. bioRxiv* 2020). (B) Cryo-EM structure of SidJ–Calmodulin (CaM) glutamylation complex (Bhogaraju *et al. Nature* 2019). Legionella effector SidJ counteracts the action of SidE, another bacterial effector that catalyses serine ubiquitination of host proteins to promote Legionnaires' disease (Bhogaraju *et al. Cell* 2016; Kalayil *et al. Nature* 2018). Bacterial effector repertoire also includes DUB proteins that add another layer of control for SidE activity (Shin *et al. Molecular Cell* 2020). (C) Labelling of different viral capsid proteins with maleimide reactive NLS peptide and reactive fluorescent dye allows for building capsids with different labelling ratio. Capsid structures (top) and their EM images (bottom, scale 50 nm) are then used to define the biophysical parameters for their nuclear entry through the Nuclear Pore Complex (Paci and Lemke, *bioRxiv* 2019). (D) Integrative pathway and structure modelling of the nuclear export of viral genomes during influenza A virus infection cycle. (E) *In vitro* vascular systems, (i) engineered capillary-size vessels can be used to spatiotemporally track *P. falciparum*-infected red blood cell sequestration in the microvasculature. Sequestration in the capillaries can be quantified with (ii) light or (iii) electron microscopy. Sequestration of infected red blood cells is shown in (ii) with a heat map (blue, no cytoadhesion; red, high cytoadhesion) and in (iii) with an asterisk. EC is endothelial cells.

Viewing the Host–Pathogen Interface

EMBL will employ its unique imaging and structural capacities to view intracellular pathogens at different stages of infection and characterise their interactions with the host at an atomic-resolution level. FIB-SEM, cryo-EM, and cryo-ET will be used to capture subcellular changes during the course of infection and to view compartments that intracellular pathogens create or reside in.  Pilot studies are underway, with an initial focus on different viruses, including SARS-CoV-2, and will be performed in collaboration with the University of Heidelberg. EMBL will also exploit its expertise in nuclear pore complex (NPC) biology and couple biophysical measurements with imaging and electron microscopy structural information to gain a better mechanistic insight into how large viral capsids enter the nucleus (Figure IB2C), how viral pre-integration complexes facilitate nuclear import, and how viral ribonucleoproteins (RNPs) exit the nucleus through the NPC at later stages of infection (Figure IB2D). EMBL will also capitalise on its wide interests in nuclear biology and shed more light into how viruses take over nuclear processes. For example, cryo-ET studies of chromatin,

chromatin remodellers, and viral pioneering transcription factors will facilitate an understanding of how host cell chromatin is hijacked by viruses to directly integrate and/or modulate gene expression.


Modelling Host–Pathogen Interfaces

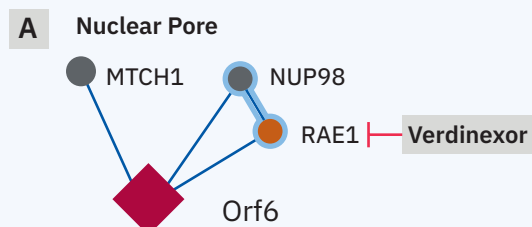
Beyond mapping and viewing interfaces, EMBL will go a step further to develop computational models of these interfaces, integrating diverse data types and engineering the next generation of sophisticated experimental models to study host–pathogen interactions. To create computational models of different granularity, from signalling network models to whole-cell models, EMBL will integrate experimental information on host and host–pathogen protein–protein interactions, genome-wide data on protein expression, localisation, activity, and PTMs during infection, and structural data on interface points. For example, EMBL researchers will integrate structural and systems-based information to model the way the influenza RNPs exit the nucleus via the NPC and assemble viral particles when leaving the cell (Figure IB2D). The models will point to new intervention points for viral and bacterial infections and will guide future experimentation at the host–pathogen interface.

Technology Development Box TD2_IB | Integrated systems-based approaches to tackle emerging pathogens.

Pathogens use a myriad of mechanisms to trick host defences and to hijack host processes for their own benefit. These mechanisms range from intercepting host signalling and epigenetics to directly recruiting necessary protein machines and enzymes to build compartments, protect from host defences, scavenge nutrients, and move within the cell. EMBL is home to a suite of unique systematic technologies that can provide paradigm-shifting insights into host–pathogen interfaces.

Quantitative proteomics-based approaches (see Chapter 5: Microbial Ecosystems, Tech Dev Box TD3_ME), such as thermal proteome profiling (TPP), are unique in providing the state of host and bacterial proteins during infection, illuminating the active interfaces and pathways at different stages of infection, and pinpointing possible drug targets. Systematic protein–protein interaction (PPI) profiling can unravel the direct targets of effector proteins and identify drugs that disrupt these interactions. Recently, EMBL has pioneered methods to map PPIs in the context of infection (Figure IB2A). Profiling proteome turnover (Savitski *et al. Cell* 2018), phosphorylation (Potel *et al. Nature Methods* 2018), and localisation (Selkrig *et al. Nature Microbiology* 2020) can provide further insights into mechanisms pathogens use to intercept host responses. This versatile and systematic profiling of proteome states can be combined with computational approaches to assess the functional consequences of these events at different levels, from an atomic view of a PPI interface (Bradley & Beltrao *Plos Biology* 2019) to dissecting signalling networks (Müller *et al. Nature Cell Biology* 2020; Ochoa *et al. Nature Biotechnology* 2020), to identify downstream

druggable host proteins/pathways, and to predict pathogenic traits, such as AMR (Galardini *et al. eLife* 2017; Bradley *et al. Nature Biotech* 2019).  In collaboration with others, EMBL has piloted research using these integrative approaches to identify how repurposed drugs may block the infection cycle of SARS-CoV-2 (Gordon *et al. Nature* 2020; Bouhaddou *et al. Cell* 2020).



B

Protein	Interaction motif		
SARS-CoV2 Orf6	56	QPM E ID	61
SARS-CoV Orf6	57	EP M ELD	62
VSV M protein	49	DE M DTHD	55
KSHV Orf10	414	EP M QS	418

SARS-CoV-2 Ofr6 protein interacts with an interferon-inducible mRNA nuclear export complex. **(A)** Small molecule inhibitors shown for RAE and **(B)** the NUP98-RAE1 interaction motifs from several viral species (Gordon *et al. Nature* 2020).

EMBL will also engineer novel 3D vascularised *in vitro* tissues and establish relevant organoid systems to simulate better infection contexts and study host–pathogen tissue interactions. These will include models to recreate human cerebral malaria pathology with cutting-edge *in vitro* bioengineering approaches. EMBL researchers will develop 3D blood-brain-barrier (BBB) models (Figure IB2E) with tubular geometry that incorporate multiple cell types (brain microvascular endothelial cells, astrocytes, and pericytes) to study how parasites, platelets, neutrophils, T cells, and cytokines potentially contribute to BBB dysfunction. Future applications of these models could extend to other parasites, viruses, or bacteria that cross the BBB and cause meningitis. EMBL will also develop vascularised cardiac models capable of mimicking endothelial barrier function seen *in vivo* to study viral or bacterial infections linked to cardiomyopathy. Altogether, these models will facilitate understanding of critical, hard-to-simulate host–pathogen interfaces and provide a unique platform for probing relevant host-targeted therapies.

Molecular Biology of the Host Response

Human Genetics of Susceptibility to Infection

EMBL is in an excellent position to use human genetics, coupled with molecular and cellular phenotyping of patient samples collected by collaborators, to characterise the human response to infection. The availability of more than six million genotyped individuals across Europe, many of them also with exome or genome sequences (Chapter 6: Human Ecosystems), enables comprehensive analyses to find host susceptibility loci that are involved in differential response to infection. EMBL scientists are also developing more sophisticated joint geographic and statistical models, which can more accurately integrate over the heterogeneous nature of infection exposure (e.g. by age, social contacts) and both traditional epidemiological and genetic risk factors. These models can potentially leverage encrypted anonymised proximity tracking, either as a way to estimate population contact maps or – more ambitiously – to fully model infectious agent exposure. These models can then be deployed to understand both endemic and epidemic disease outbreaks. The outputs of these models will provide more finely-grained risk population models, and will result in a better understanding of the pathways of infectious disease and potentially new drug targets or drug repurposing opportunities. These research efforts from EMBL complement worldwide efforts to study infectious disease genetics, which are expected to be brought together into the Genome-Wide Association Studies (GWAS) Catalog, a data resource aggregating the results of all published GWAS, which is run jointly by EMBL and the National Human Genome Research Institute (NHGRI), part of the US National Institutes of Health. By integrating these data with an extensive molecular understanding of host–pathogen interfaces, more complex models for genetics are also possible, for example involving epistatic interactions between genes in the human genome. Looking ahead, with a substantial proportion of European populations having genotype or genomic information available, such models will be important in optimising future healthcare delivery in the context of infectious diseases. With each national endeavour in EMBL member states, EMBL scientists will work in partnership with the genomic and epidemiology groups of the member state to best exploit EMBL’s scientific expertise within the context of national datasets and healthcare delivery structures.

Cellular Phenotyping of Host Immune Response

Host responses can also be studied by molecular phenotyping of patient samples, in particular from accessible tissues or samples such as blood, urine, or sputum, usually acquired at the point of hospitalisation. In collaboration with clinical partners, EMBL can provide the technology and analytical components needed for clinical studies. A particular strength of EMBL is multimodal phenotyping integration, such as integrating measurements of single-cell RNA expression, DNA methylation, and cellular microscopy with computational analyses using innovative machine learning techniques.

In patient-derived samples and in model cellular systems, EMBL will extensively study key immunological cell types using various technologies. An intriguing opportunity would be finding evidence of O-GlcNAcylation of local chromatin involved in retroviral and retrotransposon suppression in the host genome. This is one example of how fundamental mechanistic biology can provide new insights into infection biology and, conversely, how the study of infectious agents provides insight into fundamental biology. EMBL will also study the fundamental properties of immune cells, using various genomics tools such as single-cell expression and chromatin accessibility, as well as novel microscopy techniques, such as Brillouin microscopy, to measure surface stiffness in migrating immune cells (Tech Dev Box TD2_MD).

By working with collaborators in the developing world, in particular via the Human Heredity and Health in Africa (H3Africa) Initiative and via global collaborations such as the CRyPTIC consortium for *Mycobacterium tuberculosis* treatments, EMBL scientists will apply experimental and computational techniques in areas of the world with a wider variety of endemic and epidemic diseases. This has many ramifications. Firstly, higher levels of many endemic diseases and greater variation in human genetics, in particular in Sub-Saharan Africa, mean statistical studies can be powered sufficiently to find biological effects and generate associations with specific genetic variants. Secondly, EMBL can work in partnership with scientists in the developing world to deliver innovative field site pathogen genomic testing schemes using, for example, nanopore sequencing.

New Anti-infective Strategies

EMBL aims to generate new knowledge, resolve current bottlenecks, and provide innovative solutions to the rapid increase of AMR, one of the biggest public health challenges of the 21st century. In the next EMBL Programme, EMBL will build on its automated screening platforms, data analytics, and extensive molecular expertise to develop diagnostics and anti-infective strategies, to map AMR reservoirs and transmission routes, and to devise new ways to delay, prevent, or revert AMR.

Developing Diagnostic and Surveillance Tools

To foster antimicrobial stewardship and reduce the risks for development of resistance, EMBL proposes to **develop microbiome-based diagnostic and surveillance tools** for identifying pathogens or pathobionts and their resistance capabilities. The current first-line treatment for infection-associated symptoms is broad-spectrum antibiotics. Only if there is no improvement are more precise but time-consuming microbiological tests performed to identify the pathogen and its drug susceptibilities. To avoid the misuse of antibiotics, which promotes AMR, and to mitigate the collateral damage of antibiotics on commensal microbial species, rapid diagnosis of specific pathogens and their AMR potential is urgently needed. This knowledge can be used to guide tailored treatment options.

Gut microbiome research at EMBL (Chapter 4: Microbial Ecosystems), enables the identification of bacterial pathogens and viruses based on metagenomic or metatranscriptomic fingerprints of known infectants. This includes profiling of non-invasive samples such as stool or saliva, down to the resolution of individual genes and even residues (e.g. those that are known to confer AMR through modification of target). Leveraging EMBL's collections of public data and emerging platforms on individual pathogens, as well as the increasing knowledge of antimicrobial resistance and mobile genetic elements (Chapter 4: Microbial Ecosystems, Figure ME4B), a supervised list of hundreds of infectious agents together with their resistance risk will be developed. This list will be amended with rules derived from machine learning to summarise the worldwide distribution and abundance of pathogens and their associated resistance. This approach will not only provide strain-specific, biogeographic knowledge on infectious agents for more precise diagnosis, but will also reveal the functional repertoire of pathogens, including potential AMR genes and mechanisms.

Devising New Targets, Molecules, and Strategies to Fight Pathogens

To prevail in the fight against microbial infections, new effective targets, molecules, and strategies are urgently needed. EMBL's vision is to build on its multidisciplinary expertise, platforms, and capabilities in informatics, pathogenesis, structural biology, high-throughput screening, and drug discovery to reduce current bottlenecks in mode-of-action (MoA) understanding of anti-infectives and to discover alternative strategies and new targets to combat bacterial and viral pathogens.

A major bottleneck in anti-infective discovery is the identification of the MoA of new compounds. This is vital to improve drug efficacy, reduce the potential for resistance development, guide combinatorial drug use, and mitigate adverse effects of drugs on other microbes or on the host. EMBL has pioneered systematic genetic (chemical genetics) and biochemical approaches such as TPP (Figure IB3B) to **identify the MoA of novel anti-infective compounds**. In the new EMBL Programme, these will be complemented with machine learning approaches applied to experimental data and public chemoinformatics and drug databases (ChEMBL, Open Targets, STICH, DrugBank; many developed or maintained by EMBL), computational docking approaches taking into account natural sequence variation, and *in situ* imaging of drug action (Tech Dev Box TD1_IB).

The second way in which EMBL can contribute is by providing novel cutting-edge approaches in drug discovery. This work will leverage EMBL's comprehensive crystallographic fragment screening and chemical biology platforms (Chapter 10: Scientific Services) to **identify new candidate molecules for rapid optimisation**. These targeted approaches will be geared towards antiviral and antibacterial drugs that specifically target facets of the host–pathogen molecular interface, such as candidate secretion systems, adhesins, and secreted proteins.

Moreover, EMBL aims to **develop new strategies that break with the current trends of antimicrobial drug discovery**. Previous efforts have concentrated on target-driven discovery of broad-spectrum monotherapies and often failed, as identified molecules were prone to rapid resistance development, lacked *in vivo* efficacy, or had adverse effects. EMBL will seek strategies that follow a new path. The first goal will be to identify molecules that target specific pathogenic traits (e.g. biofilm formation and intracellular growth) rather than conserved processes that are essential for the growth of all bacteria. This will yield more specific therapies, less prone to resistance development and less detrimental to the resident microbiota. One such direction will be to identify molecules that ectopically trigger suicidal systems of bacterial pathogens, such as the toxin-antitoxin pairs prevalent in bacteria (e.g. *M. tuberculosis* carries over 100 pairs). These consist of a 'toxin' protein that inhibits bacterial growth and an interacting 'antitoxin' (RNA or protein) that neutralises the toxin. EMBL researchers will strive to uncover the mechanisms of these pairs and the molecules that they sense. The ultimate goal will be to exploit these mechanisms to turn these systems against the pathogens that carry them.

EMBL will also continue exploring **drug combinations**, which remain largely unexplored for antibacterial treatments. Combinations can offer new solutions, allowing for reuse of neglected antibiotics or for repurposing of other drugs and even food additives, to inhibit AMR pathogens (Figure IB3A). More importantly, combinations offer ways to prevent, reduce, and revert antibiotic resistance. Recently, EMBL researchers discovered that combinations have species-specific activity and can thus be used to decrease the collateral damage of antibiotics to commensals (Chapter 4: Microbial Ecosystems, Figure ME4A). This reduces pressure on the entire community to develop resistance. To provide more natural treatment options, the combinatorial effects of bioactive molecules present in food will be explored. Efforts to use phages or probiotics to selectively remove looming pathogens from human microbiome will provide a further innovative strategy (Chapter 4: Microbial Ecosystems).

Preventing the Development and Spread of Antimicrobial Resistance

Technological advances in DNA sequencing, telecommunications, and artificial intelligence provide unrivalled opportunities for monitoring AMR in health systems, industry, agriculture, sanitation, and transportation. EMBL can advance this revolution by developing tools and knowledge for **AMR detection and discovery**, based solely on genomic information. These tools will promote precise diagnostics, guide medical decisions, and support epidemiology and environmental surveillance. EMBL's previous contributions have opened new paths for bacterial GWAS and predictive models. Ensuing goals are to: (i) improve the cataloguing of resistance gene families in databases; (ii) improve genotype-phenotype associations in diverse strains of pathogens and in microbiome communities for confident discovery of new AMR genes; (iii) track the evolution of resistance in infected patients and the environment; and (iv) build robust predictive models for AMR development and evolution. Extending on genomics-based efforts, EMBL aims to integrate omics, imaging, and phenotypic data and to pursue molecular mechanistic and structural insights (Figure IB3C) into the functioning of **new AMR elements** (e.g. transporters and ribosome modulation).

Particular attention will be devoted to **understanding and preventing the development and spread of AMR**. In bacteria, resistance and virulence genes can move within or between genomes carried on mobile genetic elements (Figure IB3D; Chapter 4: Microbial Ecosystems, Figure ME4B). This is a critical issue for human health, as it leads to rapid development of difficult-to-treat MDR superbugs. By combining expertise in bioinformatics, metagenomics, microbiology, genetics, chemical biology, and structural biology, EMBL aims to begin filling this knowledge gap, mapping the transmission routes, the underlying mechanisms, and its Achilles heel (e.g. environments where AMR confers a fitness cost). This understanding is essential to identify strategies to limit or even prevent the development and spreading of AMR, and to revert it in cases where it is already advanced. Here, EMBL aims to systematically explore paths that exploit the use of a second drug or chemical (inducing collateral sensitivity), microbial species interactions, and host responses.

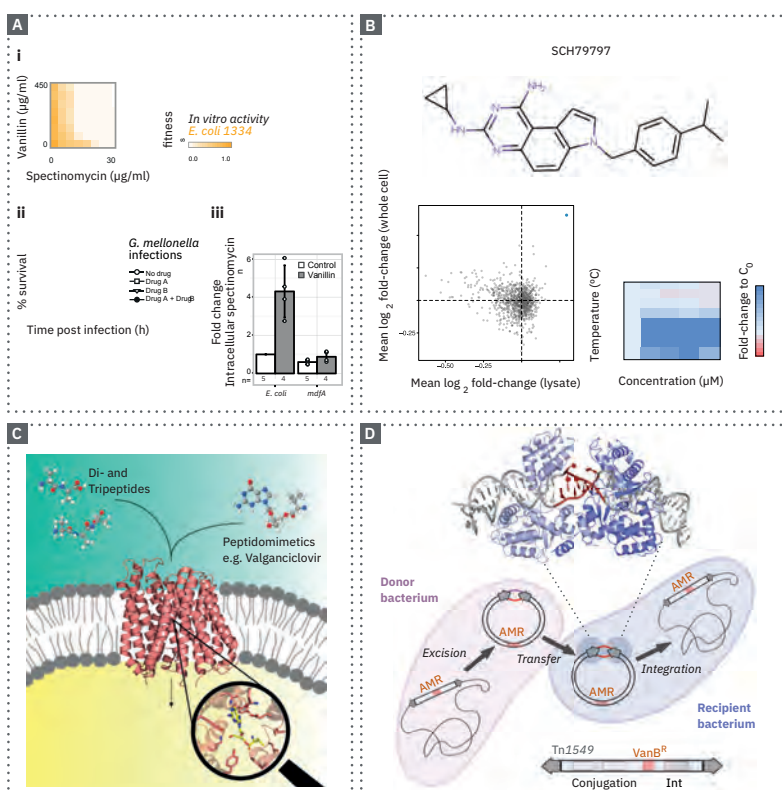



Figure IB3 | Mechanisms for new anti-infectives.

(A) Drug combinations provide new solutions – a neglected antibiotic, spectinomycin and a food additive, vanillin act synergistically against MDR *E. coli* **(i)** *in vitro* and **(ii)** *in vivo*. **(iii)** Vanillin promotes spectinomycin uptake by *E. coli* via the MdfA transporter (Brochado *et al. Nature* 2018). **(B)** Thermal proteome profiling (TPP) can be used to identify the MoA of antibiotics (Mateus *et al. Molecular Systems Biology* 2018; Imay *et al. Nature* 2019). Here, an example of a new molecule with dual activity is shown. One of its moieties targets the essential dihydrofolate reductase (FolA) in *E. coli*, causing its strong thermal stabilisation (Martin *et al. Cell* 2020). **(C)** A structure of the peptide transporter DptA with the antiviral drug valganciclovir provides insights into how the drug enters human cells and what sequence variants may lead to resistance (Ural-Blimke *et al. JACS* 2019). **(D)** Transposase-DNA complex structure reveals mechanism for mobile genetic element-mediated AMR transfer (Rubio-Cosials *et al. Cell* 2018). The Tn1549 transposon transfers vancomycin resistance from a donor (pink) to a recipient bacterium (blue), using self-encoded transposition (Int) and conjugation proteins.

Data Platforms

Platforms for Outbreaks and Monitoring of Global Threats

EMBL has a track record of curating, annotating, and sharing multi-omics data from pathogens, along with contextual metadata, primarily from scientific experiments. In recent years, collaborative projects with European public health agencies have aimed to develop platforms for data-sharing and analysis of outbreaks. The experience in the provision of data platforms and portals (e.g. the Pathogen Portal and the European SARS-CoV-2 Data Platform for the coordination of data from the current pandemic) has emphasised the need to **consider public health as a further arm of the scientific community that provides and uses molecular data**. Extending EMBL's operations to support public health brings important data for EMBL's traditional scientific services and research, but also allows for EMBL to engage in new science to predict, understand, and monitor outbreaks and global threats.

EMBL researchers have also developed novel genome workflows and DNA search tools which enable real-time global surveillance based on clinical sequence-based diagnostics for tuberculosis (TB; Tech Dev Box TD3_IB).  Current collaborations with the WHO supranational TB laboratory in Argentina will pilot this platform as an early-warning system for national public health organisations (e.g. automated outbreak detection or cross-border transmission alerts) and as a means for measuring prevalence of AMR to current and repurposed drugs. Furthermore, this collaboration is intended to facilitate the adoption of sequencing by low- and middle-income countries.

Public health microbiology is now a major global force in terms of generating sequences and associated metadata, and in terms of its importance for human health. Currently, the uses to which public health agencies put genomic data are relatively limited, and not all data are deposited openly. EMBL will deliver platforms and analytics that make pathogen genomics data and microbiome metagenomics data more accessible, facilitating global collaboration. These data will be integrated with existing EMBL open data resources and will incorporate new types of data (e.g. socio-economic, travel, meteorological, climate change, food chain, social media, healthcare, and biodiversity).

EMBL will leverage existing pathogen data platforms to deliver key global infrastructure to address future rapid outbreaks and surveillance of pathogens via the **4D pathogen genome, variation, and phylogenetic map**. This will be directly accessible by users via a navigable map, underpinned by structured omics datasets, tools to manage and support data sharing, and an analytical machine to render raw sequences into genomes, variations, and phylogenies. The system will connect pathogen data to host data, such as vector distributions; host transcriptome, immunome, and microbiome; and clinical or epidemiological data. The system will drive open data sharing in infection biology and will enable early controlled access, allowing public health agencies to collaboratively pre-analyse data before making it public, and offering dedicated cloud computing.

Technology Development Box TD3_IB | Real-time genome-based global epidemiology service for TB.

There are more than 10 million new cases of TB and 1.5 million deaths every year. One in five deaths is due to MDR *Mycobacterium tuberculosis*, the causative agent of TB. Solving two key problems would greatly aid global management of TB. First, comparing a bacterial genome from a patient with a live database of global infections is currently not possible, partially due to the sheer challenge of building such a system in a scalable manner. The second challenge is that the majority of the TB burden is borne by low- and middle-income countries, who have many challenges to the adoption of routine sequencing of TB and whose frontline staff have limited incentive to share data. Towards these ends, the Iqbal group is developing a platform, Mykrobe Atlas TB. The first problem is solved by developing a DNA search index capable of storing millions of genomes. The second is solved by providing state-of-the-art TB diagnostics connected to upload of data (if consented, and with embargo). Upload is rapidly followed by information on outbreaks, drug resistance, and monitoring, providing an incentive. An early version is being trialled in Argentina at the Malbrán Institute in 2020, with potential for follow-up in Madagascar, India, and South Africa.



Data upload

M. tuberculosis raw sequence data
Geography (fuzziness supported)
No patient identifiable data

Intended users

National TB reference labs
TB control officers
W.H.O

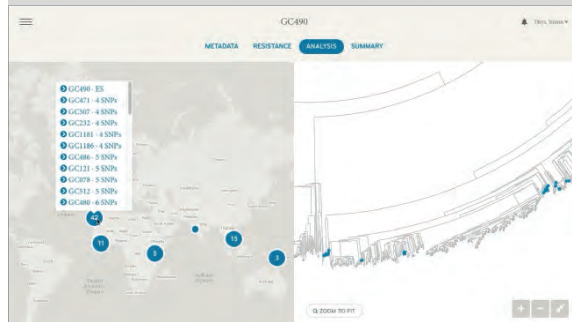
Features

Secure access
Sharing within/ between organisations
Customisable dashboards
Filter by time/space/genetics/resistance

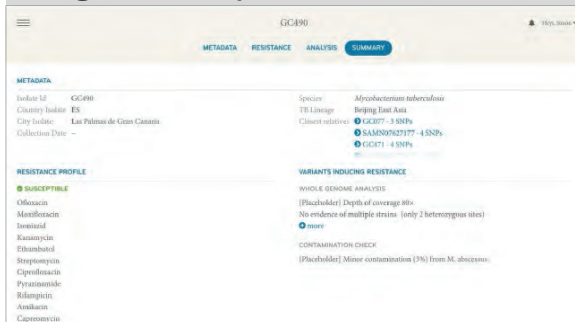
Benefits

Outbreak detection
Cross-border movement detection
Fast analysis: incentive for data submitter

Real-time surveillance



Drug resistance profiles



Open Targets for Microbes

The Open Targets Platform is an integrated EMBL data resource that allows users to identify and prioritise targets for the development of safe and effective medicines for human disease (Figure IT4 in Chapter 12: Innovation and Translation). The platform integrates evidence from various databases at EMBL's European Bioinformatics Institute (EMBL-EBI) and elsewhere to identify and score likely targets, and provides additional data to allow prioritisation based on target tractability, known safety liabilities, and tissue selectivity. Currently the platform focuses on chronic human disease with a genetic contribution. Building

on its experience of developing the Open Targets Platform, EMBL will develop a sister platform for microbes, serving both academic and pharmaceutical communities, to tackle the challenges of infection biology. EMBL plans to liaise with relevant European institutes and industrial partners to develop an informatics platform that can be used to identify protein targets for anti-infectives, and to quantify disease relevance, tractability, and resistance evolution potential. The resource will promote the prediction and modelling of small-molecule target engagements (on pathogen or host), foster the prediction of resistance profiles, and help scientists to understand the MoA of novel compounds.

The infrastructure already available through the Open Targets Platform could be restructured to include pathogen proteins as targets, alongside the host proteins with which they physically interact. This information would be extracted from published experimental research, or would be predicted based on data from biological pathways and known interactions. Additional descriptions of disease ontologies and of AMR pathways would also be necessary. Such a database could be provided from the cloud as an open access resource to allow the following questions to be answered:

- What virulence proteins do pathogens encode? What information or data are available on these? What are their close orthologues from related pathogens or human genes?
- Which human proteins or receptors interact with the pathogen or its proteins? What networks are they part of? What drugs or compounds are available? How do these active compounds work? What bioactivity and AMR data are there?
- What clinical trials exist for compounds binding these targets? Are there drug repurposing opportunities? If drugs are being used in clinical trials, what are the known targets of these drugs? Are there known safety risks? What classes of mechanisms do these drugs fall into?

Impact

The recent SARS-CoV-2 pandemic has exposed the gaps that exist not just in our knowledge of pathogen biology and human susceptibility to infection, but also in our abilities to rapidly detect infections, monitor transmission routes, and develop therapeutics. EMBL in close collaboration with its member states can help address many of these limitations.

Firstly, there is a large societal impact that stems from the lack of broad and integrated approaches for mechanistic analyses of the biology of pathogens and their interactions with humans and other hosts. The use of cutting-edge technologies such as proteomics, structural biology techniques, high-resolution imaging, and computational methods to map host–pathogen interfaces across different levels and scales stands to provide a deeper understanding of pathogen biology, new therapeutic intervention points, and new means and mechanisms to modulate host cell biology. Within these approaches lies the opportunity to gain new knowledge of the principles by which pathogens invade cells and evade the immune system. From these principles, the global community will have a far better knowledge to react to new emerging pathogens. Cross-disciplinary collaborations such as the Centre for Structural Systems Biology (CSSB) at EMBL Hamburg, a joint initiative comprising three universities and six research institutes, will be important to focus on mechanistic questions in infection research, including concrete applications for drug discovery and new therapies.

Currently, AMR is shaking modern medicine to its foundations: invasive surgery, transplantation, chemotherapy, premature infant care, and care of the critically and chronically ill are some of the areas that depend on the ability to control infections. MDR pathogens are a major threat to global public health and a priority to be urgently addressed, both according to the WHO and the UN. The current SARS-CoV-2 pandemic might further expose the negative consequences of our shortage of working antibiotics as hospitalised

patients are regularly treated with multiple antibiotics to prevent pneumonias, which will in turn promote further AMR development. Building on the largest genomic databases, unique tools, multidisciplinary skills, and international research networks, EMBL will have unmatched possibilities to make a difference towards combatting resistance dissemination and the threat of MDR infections. The knowledge and intervention approaches devised at EMBL will benefit the entire European research landscape and society with a long-term medical and fiscal benefit.

The new infectious disease-related computational tools will improve pathogen and AMR detection. Dedicated data platforms are intended to deliver immediate utility to the EMBL member states, focusing on the needs of public health users. By gaining broad engagement, these will amass hugely valuable open datasets for science. For fundamental scientists in the member states, the 4D pathogen genome, variation, and phylogenetic map will provide a uniquely rich view into public data, integrating pathogen, host, geography, and time. Finally, Open Targets for Microbes will be a foundational resource that can underpin the discovery of future anti-infectives.

6. Human Ecosystems

Background

Human ecosystems research at EMBL aims to understand how the environment interacts with humans during development and adult life. A central question in human ecosystems research is understanding how environmental factors can precipitate disease, and more generally, how genotype and the environment influence human phenotypes. The question concerns not just an understanding of how the environment impacts the individual, but also how that individual changes its environment, from its own intimate biological environment of commensal microorganisms, through to the large changes humans make to the physical environment.

In the context of human ecosystems, the term ‘environment’ can be separated into three distinct components – the **physical, biological, and social environments** (Figure HE1). The physical environment includes factors such as pollutants, chemicals and nutrition. The biological environment encompasses the organisms interacting with humans, which particularly includes symbiotic, commensal and parasitic microorganisms. The social environment describes the way in which humans can be affected by human behaviour and by social interactions. These three environmental components can interact with one another, for example, nutrient intake can be affected by the social environment, and by individual genotypes (that is, by genotypes of humans, and by those of other organisms, present in the biological environment). All these components, along with fundamental stochastic events in each person’s life, determine the complex phenotypes of an individual.

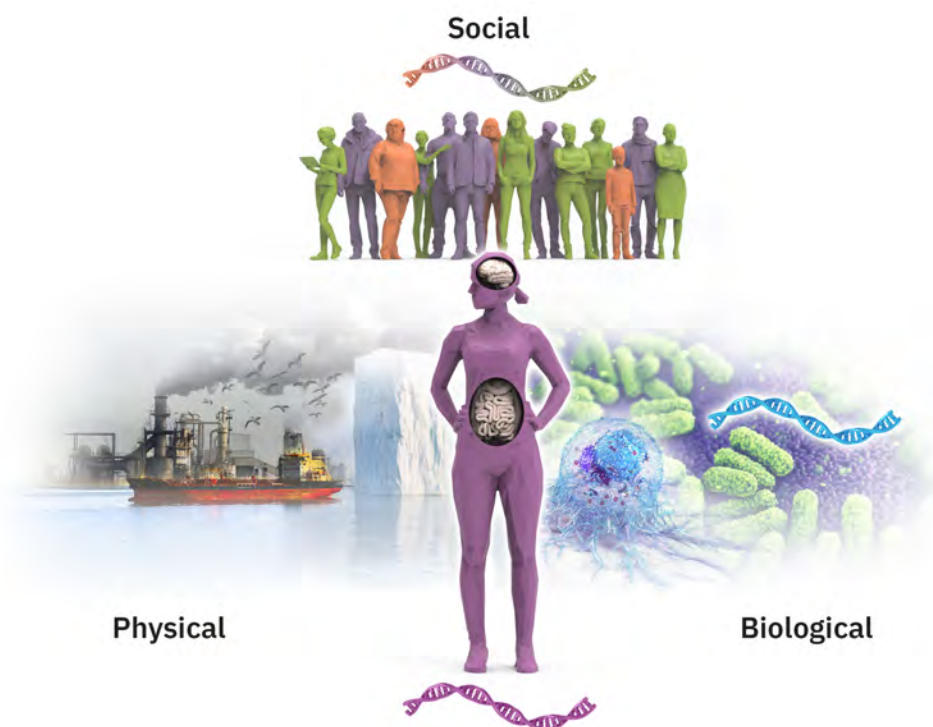


Figure HE1 | Human ecosystems.

Human phenotypes are impacted by the physical, biological, and social environment. Host genetic variation moderates this impact, and biological and social environmental effects can be studied by examining indirect genetic effects – such as the impact of one organism’s genetic variation, on the phenotype of another.

It is widely documented that environmental factors, for example, pollution, or lifestyle-related factors such as smoking and diet, are leading determinants of human health. Estimates from the World Health Organization suggest that at least one-quarter of the disease burden worldwide is attributable to modifiable environmental factors, and the majority of this burden relates to non-infectious diseases. Many of the diseases which are increasing in prevalence, such as cardiovascular disease, diabetes, cancer and mental illness, are often complex combinations of the three types of environmental components described above. Understanding the way in which environmental factors can contribute to phenotypes, and in exacerbating human disease at a mechanistic level, can facilitate more accurate disease intervention and prevention strategies. Such an understanding is, therefore, critical to human health and wellbeing.

Science and society are experiencing the beginning of an era of unprecedented accumulation of new knowledge about humans and their environment. In the coming decade, hundreds of millions of individuals will have their health data recorded. Databases will capture detailed information on physiological and molecular measurements, utilising technologies such as genome sequencing or wearable devices, to collate health outcomes or environmental data from a variety of sources. This information has the capacity not only to capture genetic data, but also the 'molecular fingerprints' of the environment, that can be obtained by analysing these data. This is explained in more detail below.

Environmental molecular fingerprints of the physical, biological and social environment can be measured in different ways, including, but not limited to:

- **Population-scale genotyped human datasets:** often coupled to Electronic Health Records (EHRs), as key resources for research. These datasets, containing genotype and environmental data, can be instrumental when investigating diseases and healthy aging.
- **Metagenomes:** the human microbiome is both a part of, and an important mediator of the environment. Microbiomes may exert disease-promoting or disease-mitigating effects, and can act as a biomarker for disease. Unraveling host-microbiome interactions can, thus, provide insights into microbiome-dependent mechanisms underlying health and disease
- **Epigenetic profiles:** environmental exposures often leave epigenetic markers (such as DNA methylation) in accessible tissues, such as blood, which can be recognised later in life. Other high-dimensional molecular data (for example, blood and urine metabolomics) can provide new insights into the homeostatic state of any particular individual.
- **Mutational signatures of tumours:** can be derived from cancer genomic sequencing (to explore genomes, transcriptomes and epigenomes) or other high throughput omics technologies, such as metabolomics, which can capture the effect of physical and biological environment exposures (eg. tobacco or UV-light exposure, or the presence of tumour-associated microorganisms or viruses) on the cancer's somatic genome.
- **Brain activity:** is an important data type with particular relevance to the social environment because it reflects the individual's immediate and past sensory environment and because it drives behaviours that alter the social environment. Brain activity can be quantified by direct physiological measurements, such as EEG/MEG, or inferred by imaging methods, such as fMRI.

Collectively, these data can illuminate not only the crucial influence of environmental components, but also interactions between genetic and environmental factors, which in turn can reveal causal molecular and physiological insights into human phenotypes and disease. The scientific community, through collaborations with other sectors and with citizens, can harness these data to promote and enable healthy living. EMBL proposes to lead revitalisation of the mechanistic understanding of human ecosystems research across Europe.

The Opportunity

Advancing human ecosystems research is timely for two reasons. First, environmental challenges are increasingly impacting human health, both as humans deplete and pollute natural resources, and as they are exposed to dietary excesses and stressors, brought on by wealth and urbanisation. Second, science is experiencing an exponential rise in the data available on human molecular and related environmental measurements, including genetics coupled to phenotypic variation as part of healthcare. The current pandemic has emphasised the critical importance of understanding healthy human ecosystems. New knowledge is needed not just of infectious agents but also more comprehensive understanding of the impact of human behaviour. Fundamental molecular biology research is an essential component in responding intelligently and ethically to this challenge. Research insights generated through integration of these data are likely to yield novel opportunities for disease intervention or even prevention, and promote healthy ageing.

Extensive human cohort datasets are primed to become key resources for researchers. These are not only emerging across Europe, but also more globally. They facilitate the development of hypotheses about the impact of physical, biological, and social environmental effects. Sophisticated statistical and computational methods can infer environmental effects based on population cohort data, in particular with molecular measures, broadening the types of environmental effects accessible to research. The impact of such population-scale studies will be enhanced by integrating and standardising data across country borders and by connecting them with molecular data resources, such as those hosted at EMBL.

EMBL's unique position to lead in human ecosystem research can be attributed to a number of its current strengths:

- EMBL plays a leading role in developing tools and access platforms for the **harmonisation and integration of biological data** worldwide.
- EMBL is a European inter-governmental organization with a **research-enabling and policy-guiding mission** that can facilitate the leveraging of human cohort datasets across borders in a manner that magnifies national investments in this area.
- EMBL is in a position to **test hypotheses deriving from human ecosystem studies in a mechanistic, intervention-oriented manner** because it has longstanding expertise in a wide range of model systems, combined with expertise in cutting-edge multiomics, imaging, and phenotyping technologies.

Research Aims

In this Programme, EMBL will explore quantitatively the effects of the environment and its influences on human biology. EMBL aims to integrate two approaches – statistical discovery from diverse, large-scale cohorts, and laboratory-based interrogation of specific molecular processes – to bring a quantitative, mechanistic, and molecular understanding to environmental effects. The influence of the environment, whether physical, biological, or social, will be approached mechanistically using these two complementary methods.

- I. **Statistical discovery based on cohorts.** In collaboration with member state scientists, EMBL aims to leverage large-scale population cohorts, where both the influence of the environment, or environmental component (**E**), and phenotype (**P**) are measured in a large collection of individuals. By studying associations between **E** and **P**, it is possible to derive hypotheses about

environmental effects by considering a wide range of exposures, including lifestyle factors and long-term exposures. However, as human cohort studies cannot be controlled, moving from associations to understanding causal relationships can be challenging. In settings where genetic information (**G**) is also available, this can help scientists to derive more mechanistic hypotheses. As any environmental change must interact with the molecules generated from the genome (often proteins), there will be scenarios where two different genotypes (**G**) (alleles) producing two variants of a protein or different levels of it, will respond to an environmental variation (**E**) in different ways – resulting in different phenotypes (**P**) in a given environment. This is known as genotype–environment interactions (**G×E**). Large-scale human cohorts comprising genotype and phenotype information, and ideally as many environmental measures as possible, can be computationally leveraged to uncover **G×E** interactions and to generate hypotheses about the molecular pathways that mediate the impact of the environment on human phenotypes. It should be noted that environmental effects can themselves be dependent on genetic variation that can be measured (for example, the genetic makeup of microbial populations that reside in humans, or of the infectious agents that challenge humans throughout life).

- II. **Laboratory-based discovery.** A range of experimental subjects, including human (healthy and patient) primary samples, biological models such as organoids, engineered tissues, and microbiome samples, as well as model organisms such as mice and fish, can be used to interrogate the influence of the environment. Experimental setups where the environment of human-derived biosamples is directly controllable – for example cell- or organoid-based systems with chemical or infectious perturbations – are one possibility. Another possibility is the study of organisms or combinations of organisms (model systems) that scientists are confident recapitulate key aspects of human biology within a broad range of controllable environmental factors. These studies can be conducted in highly controlled environments, ranging from Petri dishes to animal husbandry systems, or in systems with controlled experimental conditions where multiple organisms can be introduced, such as mesocosms and ecotrons (Chapter 7: Planetary Biology).

In undertaking human ecosystems research, EMBL will partner with experts from other fields such as epidemiology, population health, data science, and healthcare, to leverage expertise and drive understanding in this field beyond the research context. In addition to expertise in data science and experimental molecular biology, EMBL serves as a neutral hub for data – through both data services and credible analysis; this is paired with a cross-border scientific network, and international objectives, which are not determined by individual national priorities. Thus, EMBL has a unique configuration of trans-national context, skills and technologies which enable it to investigate environmental effects on humans and to explore them at a molecular and mechanistic level.

EMBL’s Approach

Understanding the impact of the environment on humans is complex, but not unsolvable. The application of innovative statistical approaches to the rapidly expanding large-scale human cohort data available worldwide, will lead to the generation of credible hypotheses, linking environment and genetic variation to human phenotypes, and testing under controlled laboratory settings will reveal the molecular mechanisms involved. The potential also for hypotheses to be generated in human cellular systems or model organisms which are then explored in human population data also stands to generate novel discoveries in the potential causes of human disease. EMBL is uniquely positioned to perform and enable these roles.

Statistical Discovery: Hypothesis Generation in Human Cohorts

By virtue of both expertise in human molecular data management and analysis, and because of EMBL's inherent pan-European role, EMBL has access and deep technical knowledge of key **population-scale human biological datasets** (below referred to as **cohorts**) in Europe and around the world. EMBL researchers estimate that the data from more than 30 million people will be available in such cohorts throughout the world within the next five years, and these cohorts will include diverse genotypes and environmental measures.

The publicly available large (>50,000 subjects) human biological datasets emerging across Europe now contain data from more than six million human subjects (Table HE1). These datasets typically have some level of genotype information – such as whole-exome or whole-genome data, or concrete plans to generate these – and provide some direct or indirect information about environmental exposures (e.g. occupational, diet, lifestyle), social measures (e.g. mobile phone use, location), and biopsies (e.g. blood). In some cases these datasets are linked to individual health records. The advent of such datasets has triggered considerable excitement in the human genetics field, and exposome research, aimed at extracting correlations between genetics, environmental exposures and phenotypic variables, is starting to generate novel hypotheses about health outcomes. This is a new field bringing together epidemiologists, geneticists, psychologists, toxicologists, and molecular biologists to tackle these problems. The aim is to provide mechanistic insights into human biology, as well as potential solutions to prevent or treat disease, such as diagnostic or prognostic tools, novel therapeutics, or lifestyle guidance.

There are several ways in which EMBL will participate in this endeavour. First, EMBL will collaborate with member state and other international scientists to maximise the utility of these data. Second, EMBL researchers will exploit these data to carry out data science investigations on the genetic and environmental risk factors underlying human disease phenotypes. One product of this research will be the development and validation of innovative new tools and methods for analysing large human datasets, which will be made available to the wider research community.

Table HE1 | National and European research projects with large cohorts and data relevant to environmental variables.

Name	Country	Example Environmental Measures	Cohort size
EU Child Cohort Network	Pan-European pregnancy & early life cohort meta project comprising existing cohorts (incl. ALSPAC, Born in Bradford, SWS)	Socioeconomic status, Migration, Urban Environment, CVD, Respiratory, Mental Health, Linked medical records	250,000 children and parents
UK BioBank	United Kingdom	Current address, Residence at birth, Occupation, Workplace factors, Passive smoke exposure, Indoor air pollution, Mobile phone use, Linked medical records	500,000 middle aged adults
Danish EHR	Denmark	Education, Income, Occupation, Housing, Residence, Birthplace, Pets, Linked medical records	>5,000,000 whole population >500K genotyped
Estonian Biobank	Estonia	Occupation, Education, Place of birth, Place of residence, Linked medical records	>50,000
LifeWork	Netherlands	Occupation, Place of birth, Place of residence and residential history, Air pollution, Noise, Mobile phone use, Shift work, Occupational chemical exposures, Linked medical records	>88,000
CONSTANCES	France	Place of residence, Social and demographic Socioeconomic status, Life events, Behaviours Regarding Occupation, Environment chemical, biological, biomechanical psychosocial lifelong exposure and follow-up	200,000
COSMOS*	Pan-European	Environmental exposures, Mobile phone usage, Health registry data	250,000
German National Cohort	Germany	Place of residence, Occupation, Geocoded exposure data, Education Income, Psychosocial factor, Linked medical records	200,000

**Application through individual country cohorts.*

The Development of Innovative Bioinformatics Tools and Methods

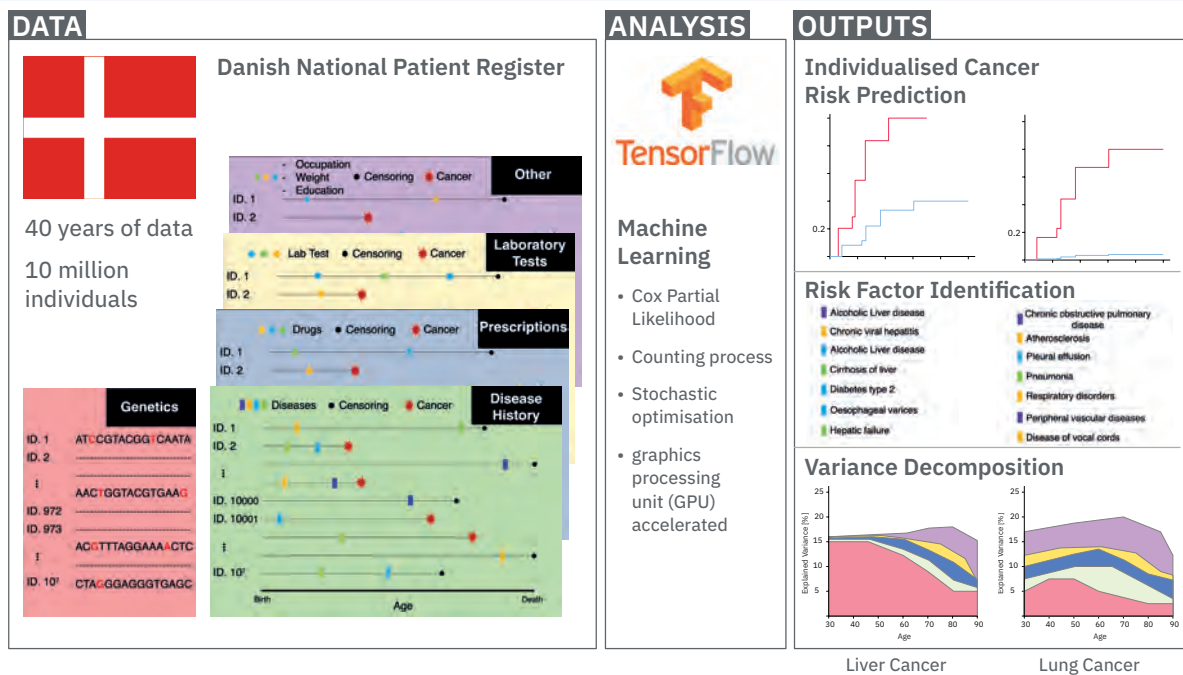
Novel statistical approaches can be applied to these data to enhance the quantitative understanding of diseases. This is critical in order to extract meaningful signals from big molecular datasets, such as genomics, epigenomics and transcriptomics, as well as large longitudinal records for (thousands to millions of) patients, which can be accessed by combining cohort datasets. This work will drive iterative improvements in computational methods which can better model and predict environmental and GxE effects, requiring improvements in methods to collect high quality data about human phenotypes. Such phenotypes that would enable the inference of environmental exposures (the exposome) that are typically difficult to measure in human populations. These methods need to operate in a distributed manner, to combine data across sites and cohorts, many of which will not be accessible within the same compute environments due to ethical and legal barriers.

These methods will build on previous work by EMBL researchers to explore which environmental factors are drivers of observed GxE such as the StructLMM algorithm to jointly assess the impact of hundreds of environmental factors (such as diet, physical activity or living conditions) on genotype-phenotype relationships. Previously, such analyses required a narrow hypothesis choosing a specific environmental factor, such as physical activity, and testing for interactions with genetic variables to understand the impact on phenotypes. The model can be, and has already been broadly applied to various areas, including BMI in UK BioBank data, and eQTL in whole blood, to identify gene expression changes associated with cellular context.

Developing innovative methods can be instrumental in, for example, **predicting a person's risk for cancer based on their medical history**. EMBL researchers investigated cohort data using statistical approaches to enhance the quantitative understanding of cancer, applying machine learning to estimate cancer risk based on recorded antecedent health data (Tech Dev Box TD1_HE). The developments of methods such as the Multi-Omics Factor Analysis method, are posing further exciting challenges in the development of methods for multi-modal data integration, semi-supervised learning, patient stratification and target identification. EMBL's future efforts to strengthen its expertise in machine-learning methods and compute infrastructure (Chapter 8: Data Sciences) as well as theoretical research (Chapter 9: Theory at EMBL) will drive further innovation in this area. EMBL will also leverage its partnerships with expertise in cohort analyses such as the Nordic EMBL Partnership for Molecular Medicine and the Molecular Medicine Partnership Unit (MMPU) to develop statistical methods and scientific infrastructure for population data.

Technology Development Box TD1_HE | Machine learning: Cox partial likelihood model to calculate cancer risk.

There are more than 17 million cancer cases per year worldwide, and the lifetime risk of developing cancer is nearly 50%. To better understand cancer risk factors, and an individual's cancer risk, EMBL researchers developed machine learning algorithms for mining electronic health records from millions of individuals. These algorithms were used to analyse health registry data covering nearly the entirety of the Danish population during the past 40 years, with data from 10 million individuals with 236 million clinical diagnoses. The algorithms are implemented using the TensorFlow AI backend, to enable an efficient analysis of large data volumes. Such inference reveals not only a very large variety of medically-assessable cancer risk factors, but also how these factors, taken together, change each individual's cancer risk. These findings have potential implications for developing more efficient and effective risk-stratified cancer screening. The analysis also provides summaries of how different factors contribute to cancer risk at different stages of life, which helps describe the natural history and typical exposures related to the disease.



Data Harmonisation, Hosting, and Coordination

Big data brings challenges in terms of data access, coordination, handling, and integration. Health data are longitudinal, with early or late manifestation of disease phenotypes. There are imperfect data linkages within countries, incomplete cross-country and cross-cohort replication, and variable data quality and assurance. To inform this Programme, EMBL hosted a human population cohort workshop in March 2020, bringing together researchers in bioinformatics, epidemiology, population health, psychiatry, human genetics, and statistics to discuss the challenges of gathering environmental variables, cohort harmonisation, and computational analysis. Researchers from all disciplines agreed that, as the number and size of human cohorts expands, so will the need for FAIR (findable, accessible, interoperable, and reusable) data standards. Capturing current and new environmental factors in human populations is a complex effort that requires a multidisciplinary

approach, with input required from many members of the research community, including sociologists, epidemiologists, geneticists, neuroscientists, and microbiologists.

Medical data cannot be shared in the same way, or at the same scale as biological data due to ethical and practical limitations. EMBL has extensive experience in handling large complex and cross-referencing datasets, and is well suited to work as a neutral broker across national borders, to facilitate and promote standard practices, and overcome obstacles to advance cohort analyses.

EMBL has been a key player in developing standards for genomic medicine, and is also a founding member of the **Global Alliance for Genomics and Health (GA4GH)**, in which EMBL-EBI leads the overall scientific direction of the project, as well as a number of technical work streams, and key real-world driver projects. GA4GH is a policy-framing and technical standards-setting organisation, which seeks to enable responsible genomic data sharing, and involves over 400 academic, healthcare and industry entities. GA4GH standards include the widely-used genomics pipelines (BAM, CRAM and VCF), and GA4GH has developed a variety of service-based protocols, consistent with data governance agreements, to provide access to genomic datasets. These protocols are being implemented by a global community, such as the Swiss Personalised Health Network (SPHN), Genomics England (GeL), the AMED genetics project in Japan, and the All of Us cohort from the National Institutes of Health (NIH).

EMBL also jointly runs the **European Genome Phenome Archive (EGA)** with the Centre for Genomic Regulation (CRG) in Spain, which is the established resource for handling research cohort data deposition on publication, including UK Biobank data. Two recent initiatives aim to extend this resource. First, the EGA is becoming more federated (Chapter 10: Scientific Services), with the admission of local EGA nodes, for example, the development of nodes in Finland and Sweden. Second, the EGA federated infrastructure is a key component of the **1+ Million Genomes** initiative, coordinated by the European Union. One such federated EGA node under development and co-led by EMBL is the **German Human Genome Phenome Archive (GHGA)**. This project aims to create valuable links between the pan-European initiatives for organising biodata and the activities in Germany. The GHGA aims to enable archiving and sharing of access-controlled human genomics data from patients across Germany, to reduce silos in biomedical research, and facilitate collaborative genome science in Germany and Europe.

EMBL is also involved in more recently launched initiatives such as the International Common Disease Alliance (ICDA) and the International HundredK+ Cohorts Consortium (IHCC). The broad reach of projects in which EMBL is directly involved, and the global nature of the standards on which EMBL leads, provide a unique viewpoint on worldwide human cohort studies. Building on these experiences, EMBL aims to facilitate integration across biological datasets, promote efficient and ethical data and cloud technologies, provide expert training, and support member states in their endeavour to maximise research access to clinical data.

Laboratory-based Discovery of Environmental Influences

Hypotheses about putative causal human environmental effects (from cohort data or other experiments) can be tested under controlled conditions, where genotype and environment can be precisely manipulated. Unlike genetic variation, environmental variation is particularly difficult to control in human populations, making it essential to test these effects in cellular or organoid systems of lower complexity, or using animal models in the laboratory. Moreover, harnessing this work to understand how humans respond to their environment will require testing on multiple scales – from molecular machines, to cells, to tissues, and to the whole organism in its physical and social milieu. In each case, the impact of **precisely varying environmental conditions, ideally with fine control of genotype information**, will be the way to provide a mechanistic understanding of environmental effects. Only when researchers are able to piece together environmental effects at various levels will the understanding of humans in the context of their ecosystems be acquired.

The Physical Environment

Mechanistically exploring the impact of the **physical environment** on humans is complex, as it is generally unethical and unsafe to expose human subjects deliberately to potentially harmful chemicals or nutrients. Instead, EMBL aims to obtain a mechanistic understanding of the effects of chemicals on cells, primary tissues, organoids, engineered tissue systems, and model organisms, with some exemplar research questions described below.

Humans are exposed to numerous potentially harmful molecules, including pesticides, industrial pollutants, synthetic molecules used in food packaging, and many others. There is also an increasing awareness that exposure in early life may be responsible for adverse effects in later life. One area where molecular technologies, such as structural biology and imaging, can help to provide mechanistic understanding of environmental pollutants (e.g. bisphenols, phthalates, or parabens) is in the study of endocrine-disrupting chemicals (EDCs), which are suspected to cause a wide range of developmental, reproductive, neurological, and metabolic defects in humans and wildlife. EDCs share some physicochemical properties with natural ligands, allowing them to bind to nuclear receptors (NRs) and activate or inhibit their action. Molecular structure studies are revealing unanticipated mechanisms by which chemically diverse EDCs interact with the ligand-binding domain of NRs. EMBL's expertise in high-throughput structural biology screening could provide a rational basis for designing novel chemicals with lower impacts on human, animal, and plant health.

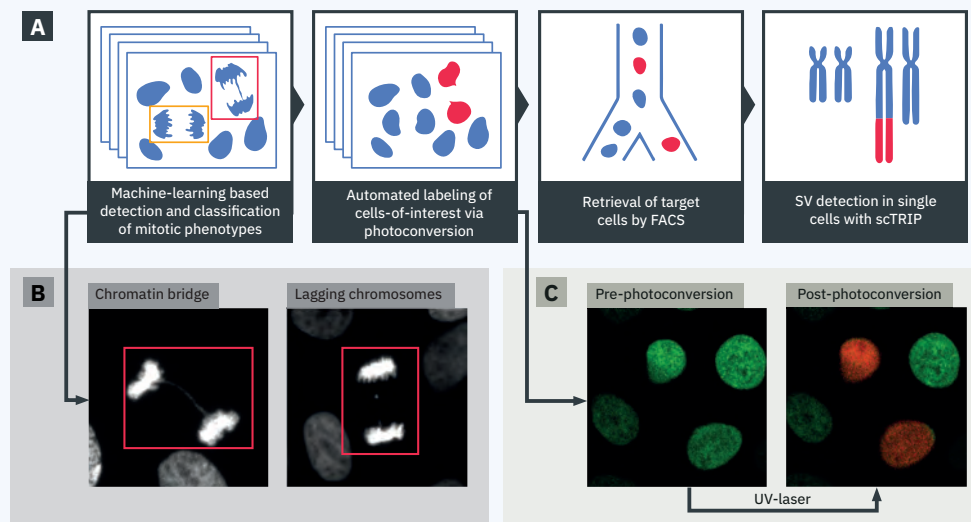
Another area where EMBL's expertise will be valuable is to understand the effects of drugs on genome variation, such as how chemotherapy induces genome instability. The genome of cancer cells is dominated by extensive somatic DNA rearrangements, also known as genomic structural variants (SVs), which include copy-number alterations as well as copy-neutral and highly complex SVs, contributing to tumorigenesis, metastasis, and therapy response. Chemotherapies given to cancer patients often lead to DNA damage, which can in turn cause the formation of SVs, thus fueling subsequent relapses (Tech Dev Box TD2_HE). Systematic perturbations (e.g. chemotherapeutics, CRISPR-Cas9 based gene knockouts) could yield novel insights into genomic instability - an enabling hallmark of cancer - especially with respect to variable chemotherapy exposure **(E)** and disease genotypes **(G)**.

Sophisticated systems are also being developed at EMBL to more closely recapitulate human organs and test for environmental factors. To model the interactions of the environment in a more physiologically relevant but still accessible scheme, researchers at EMBL have developed novel 3D vascularised *in vitro* tissues. An exemplar project is the development of the blood brain barrier to study the interaction of malaria infected-red blood cells with the brain vasculature, the immune and coagulation systems (Chapter 5: Infection Biology). EMBL Groups are developing strategies to develop 3D vascularised *in vitro* tissues, with a particular focus on cerebral and cardiac tissues. In these vascularised systems, different relevant environment perturbants, both toxins and drugs, can be introduced in defined experimental schemes. Furthermore, the genetic background of the organoid can be varied to explore the impact of genetic variation (either natural or engineered using CRISPR/Cas9 approaches for example) with these more physiologically relevant models of human tissues (Chapter 3: Cellular and Multicellular Dynamics).

Similar to cellular and tissue models, animal models allow the effect of variation in the host genotype to be assessed while exploring controlled variations in environmental factors. EMBL is developing new model systems that combine multimodal measurements and controlled environmental parameters to study the mechanisms underlying responsiveness to environmental cues (Chapter 3: Cellular and Multicellular Dynamics). 🧑🏫 A pilot project is in progress aimed to study the full complement of responses to environmental changes, exploiting the wild-derived inbred MIKK Kiyosu panel of medaka (Japanese rice-paddy) fish. This panel is formed from 80 inbred lines from a single population, inbred to near-homozygosity, and fully sequenced. EMBL researchers, together with collaborators, observed reproducible effects of selected chemicals on skeletal development, and found many cases of clear GxE effects, which will then be molecularly characterised using in depth phenotyping, CRISPR genetic tools, RNAseq, and *in vitro* assays.

Technology Development Box TD2_HE | SmartMS.

EMBL researchers are developing an innovative methodology (coined “SmartMS”) which bridges imaging and single-cell DNA sequencing to uncover the effect of mitotic errors on the formation of genomic structural variants in cancer. The researchers will employ SmartMS on longitudinally collected patient-derived leukemia samples, to allow investigating genomic instability in primary patient cells. By dissecting SV formation mechanisms at the single-cell level, SmartMS may reveal how cancer genome landscapes are largely shaped, and this could open up new avenues for personalised medicine.



This EMBL-developed technology integrates imaging with the scTRIP single-cell method, to systematically link microscopically detectable mitotic errors (noted with red box) with SV formation (A). Automated mitotic error detection will be pursued with machine learning, using a microscope equipped for adaptive feedback (B). Labelling of corresponding cells is possible thanks to a photoconvertible fluorescent marker (Dendra2-H2B; C), and this enables automated cell sorting via FACS. Target cells can then be subjected to single-cell sequencing. Same field of view shown before and after conversion (from green to red) by selective UV-laser illumination (C).

The Biological Environment

Studies of human ecosystems are complicated by the fact that, in many cases, the environment (E) is the product of genetic variation (G) itself; this occurs, for example, when the human gut interacts with the gut microbiome. The human microbiome and its host together form a discrete ecological unit called a holobiont. As outlined in Chapter 4: Microbial Ecosystems, EMBL has deep expertise in microbiome research, which is being leveraged to understand the involvement of the human microbiome in disease aetiology.

The role of the microbiome in human physiology and health – especially that of the gut microbiome, which is the most substantial microbial community in our bodies – is multifaceted and governed by complex interactions and wide-ranging effects. The gut microbiota has been linked to gastrointestinal diseases (gastrointestinal cancer, Crohn’s disease), metabolic disorders (obesity, metabolic syndrome), liver disease, cardiovascular disorders, and a number of neurodegenerative diseases (Alzheimer’s, Parkinson’s, amyotrophic lateral sclerosis), as well as neurological conditions such as autism. Given the rapidly growing evidence base indicating microbiome risk factors for human diseases, there is a need for interdisciplinary

research linking human cohort data, bioinformatics and statistical methods, microbial metagenomics, and high-throughput laboratory screening approaches – all of which are areas where EMBL has unique expertise. The approaches described in Chapter 4: Microbial Ecosystems can provide a mechanistic understanding of how the human microbiome impacts diseases, stimulating further human health investigations and enabling the development of specific therapeutic applications.

Tumours can also be seen as a cellular community within the human ecosystem, in which tumour cells cooperate with other tumour cells and with host cells in their microenvironment. The tumour ecosystem thus comprises cancer cells, non-cancerous red blood cells, endothelial cells, fat cells, stroma, and immune cells. As conditions change, this ecosystem can evolve and adapt to maintain the survival and growth of the cancer. Increasing molecular understanding of the intricate dynamics of this ecosystem has led to revolutionary treatments such as immunotherapies. Additionally, microbial cells may contribute to the cancer ecosystem, for example in gastrointestinal cancers. Thus, microbes could possibly be exploited as biomarkers to inform therapies, particularly immunotherapy, given their ability to modulate immune cells. Prior studies performed at EMBL identified key species (*Fusobacterium nucleatum*) in the gut microbiome, which have long been recognised as associated with colorectal cancer development (CRC) or progression. 🧑‍🔬 In a pilot project, EMBL researchers and clinical collaborators from the University of Heidelberg, are aiming to dissect the three-way interactions of cancer, immune cells and microbiota by integrating bulk-, gridded- and spatially-resolved analyses of CRC resections, genetic, transcriptomic, cellular heterogeneity and microbial colonisation within tumours, to build spatially resolved models of the whole CRC ecosystem (Figure HE2). A particular emphasis of the study will be on identifying differences between the cancer ecosystems of microsatellite-stable versus microsatellite-unstable (MSI) CRC (corresponding to 15% of all CRCs). MSI leads to higher levels of immune cell infiltration, and can render these cancers treatable with immunotherapies. This study could reveal insights into the effect of cancer-immune-microbiome interactions on therapy outcomes, where microbiota may act as additional stimuli (or suppressors), leading to differential response to immunotherapy treatment in MSI tumours.

This study leverages EMBL's expertise in cancer genomics and transcriptomics, imaging, technology development, and collaborative human cohort analysis and microbiome research - a combination of expertises only rarely found in an individual institution. Further mechanistic studies could be pursued using fluorescence microscopy, or even down to subcellular structures and possible infection mechanisms using EM, providing the opportunity to utilise RNA FISH against bacterial species infecting cancer cells (as has been shown for *F. nucleatum*), in order to identify cellular compartments affected. EMBL's imaging technology has unparalleled resolution scales from atomic to organismal scales, which is likely to facilitate these research angles. As human ecosystems research requires an in-depth understanding of the relationship between organisms, the spatial and temporal association of species and molecules is critical.

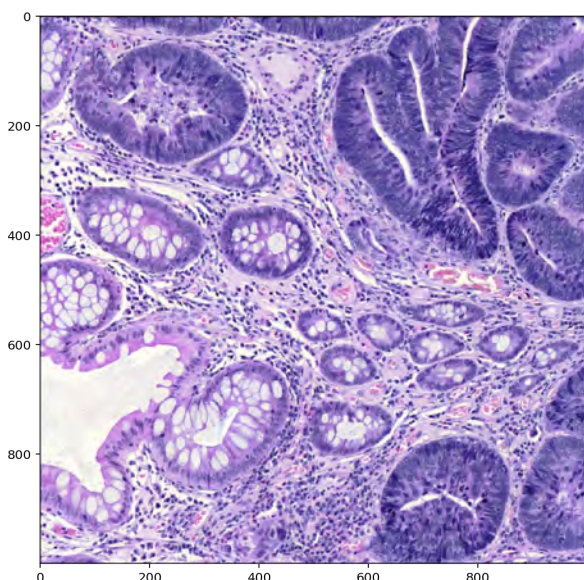


Figure HE2 | Colorectal cancer spatial omics study.

Microscopy image of a hematoxylin and Eosin (H&E) stained section of a colorectal adenocarcinoma. The section reveals normal and mutant colonic crypts (elongated purple structure; **top right**), stromal tissue (**middle**), fat cells (white bubbles; **left**) and immune cells (pepper-corn-like aggregates; **centre bottom**).

The Social Environment

Interactions among individuals are a central component of human ecosystems. The social environment is an important source of human well-being, but simultaneously also a major component in human morbidity, in particular with regard to mental health. As a result, understanding pathological mechanisms of the social environment is essential for ensuring healthy societies. The brain is the primary target organ for social interactions based on behaviour. The nervous system is unique in its ability to respond and adapt to the environment on a millisecond timescale, and to precisely represent the environment as electrical impulses. Neuroscientists have studied the cellular and molecular mechanisms of the brain's response to (and stored experience of) the environment for more than a century. However, this field is now poised to take advantage of the revolution in emerging brain and behavior-related human data and EMBL is well placed to leverage its expertise in big data analysis methods with its expertise in laboratory neuroscience in a manner that takes advantage of its close links to EMBL partner neuroscience institutions (e.g. the Nordic EMBL Partnership for Molecular Medicine, BRAINCITY, others through MPMU) as well as other collaborations and new recruits.

Social Genetic Effects

Understanding how our environment impacts our behavior is critical to understanding a wide range of key societal issues, including social conflict, substance abuse, artificial intelligence, and climate change. Just as is the case for biological environments, social environments are moderated by genetic variation in the social actor, allowing researchers to use genetics to probe this complex environment. This feature provides a powerful methodological access to social effects, in which genetic variation in one individual can be studied for its impact via social interactions on another individual, an approach called '**social genetic effects**' (Figure HE3). Social genetic effects were first described in farm animals where the systematic group housing of pigs, for example, allowed for the identification of genetic variants which imparted poor meat quality on all members of a pen.


Recently, EMBL researchers developed novel computational methods to identify social genetic effects, and applied these to comprehensively survey social genetic effects on biomedically relevant phenotypes in laboratory rodents. They examined both behavioral and non-behavioral phenotypes, and in a follow-up study developed methods to examine social genetic effects on a genome-wide scale – social-GWAS – for 170 phenotypes, including gene expression. Genetic variants underlying social genetic effects are expected to impact nervous system function, and referencing such variants to the cell-type specific transcriptome and chromatin accessibility maps emerging from the Human Cell Atlas, will allow for the identification of the brain cell-types and genes affected. In turn, this information will allow for the variants to be studied in model systems (cells, tissues, fish or mouse models) for the identification of intermediate phenotypes and molecular mechanisms.



Figure HE3 | Social genetic effects.

The behaviour of a person's partner can affect their own well-being. Social genetic effects describe how the genotype of a partner influences the other person's phenotype – for example, an inability to sleep well – and can lead scientists to understand the genetic variation that mediated these social effects. Studies by EMBL researchers have shown that social genetic effects can be substantial and can be driven by unexpected molecular pathways.

A critical step in evaluating social genetic effects will be the testing of variants in relevant cell-types in humanised mice (Chapter 10: Scientific Services) under controlled social conditions. Research on social genetic effects will be anchored to expertise at EMBL in the area of epigenetics, brain plasticity, and social behavior. Here, not only will new hypotheses be generated via the Centre for Human Brain Phenomics (below and Chapter 10: Scientific Services), but the careful genetic manipulation of neuronal circuits in mice and other organisms can also be used to specifically test and dissect key hypotheses. For example, EMBL researchers have carried out a series of studies to identify genes that moderate the effects of the social environment. Using an innovative reciprocal inter-cross breeding strategy, mice were exposed to either high or low levels of maternal care during early development. Later in adulthood these mice showed low or high levels of anxiety behavior, respectively. The introduction of controlled hypomorphic mutations in a series of candidate genes, was then used to identify significant gene by maternal care interactions (**GxE**). In at least one case, the researchers could then use histological, physiological, and gene expression profiling approaches to identify neural substrates which mediated the impact of maternal care on behaviour. Such substrates provide insight into how the social experiences alter behaviour and offer candidates for therapeutic intervention.

 To advance the methods required to detect, map, and functionally understand social genetic effects, researchers at EMBL are also undertaking a large genetic screen of a panel of 80 recombinant inbred lines of medaka fish (MIKK panel, see above). The robust statistical power of GWAS in medaka, combined with the high throughput behavioral screening possible in this species, will dramatically enhance the power to identify and map social genetic effects, and to develop new tools for their functional analysis that can be translated to human ecosystems research. Initial results show clear creation of social environments between medaka fish individuals; for example, timidity or boldness in exploring a novel environment transmitted to tank mates (Figure HE4). This social environment has clear genetic components which are amenable to genetic mapping.

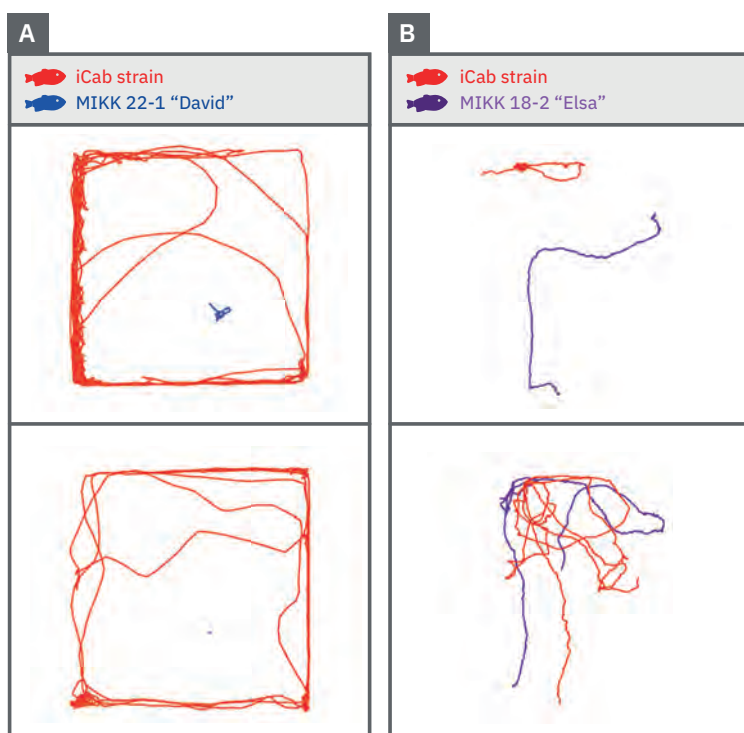


Figure HE4 | Social genetic effects in medaka fish.

Results of video tracking in four open field fish tanks, with the top and bottom rows showing replicate experiments. Each tank contains two fish – one from the reference iCab strain (**red**), and one from a MIKK panel strain (either **blue** or **purple**). In **panel A**, the iCab fish is paired with the MIKK 22-1 “**David**” strain (**blue**, almost stationary in the second replicate); in **panel B**, the iCab fish is paired with the MIKK 18-2 “**Elsa**” strain (**purple**). As shown in the red iCab traces, the behavior of the iCab fish is strongly influenced by the strain of their tank partner; iCabs paired with David show wide exploratory behavior despite David's near complete stillness, whereas when paired with Elsa, iCabs show a cautious, slow swimming behaviour, mirroring Elsa's behaviour, despite Elsa's higher overall movement relative to David.

Towards a molecular understanding of neurophysiology and behaviour

The dramatic expansion of available human phenotypic data related to brain function, including fMRI, EEG, MEG, and portable sensor-based behavioural and environmental exposure data, means that a systematic study of the human brain and its environment is now within reach. However, the systematic analysis of neurophysiological and behavioural phenotypes, or ‘phenomics’, needed to understand brain function, faces major obstacles. New statistical methods for data mining and data integration must be developed to handle imaging, portable biosensor, and social media datasets, for example, in order to extract meaningful hypotheses about the molecular mechanisms involved in human brain function.

To address these challenges EMBL proposes to establish a Centre for Human Brain Phenomics at its site in Rome. The centre will host data service activities (Chapter 10: Scientific Services), and conduct fundamental bioinformatics research, aimed at identifying genetic and environmental risk factors for human brain traits, and uncovering potential therapeutic targets for brain disorders. In particular, the Centre will focus on applying novel mathematical and statistical approaches to analyse human brain imaging and other complex, multi-modal phenotypic data to understand brain disease risk factors. Moreover, it will leverage its research expertise to develop innovative bioinformatics data mining and analysis tools and make these available to external researchers keen to exploit human brain datasets, but lacking the skills and expertise to do so, enabling them to test hypotheses relevant to their specific research questions.

Impact

Environmental factors are a leading source of disease risk, and much societal attention has focused on mitigating exposure to detrimental influences such as dietary factors, stress, pollutants, and infectious pathogens. The human ecosystems research programme aims to tackle numerous challenges and convert observations about environmental impacts on human phenotypes into mechanistic understanding of how these impacts occur. Mechanistic understanding greatly improves our ability to shape policy or undertake precision medicine interventions.

Ensuring the Effective and Ethical Use of Human Data for Mechanistic Understanding

One challenge is the extent to which the growing amount of human data emerging from medical practice will be shared and made available for research purposes. How will these be fairly and effectively used to promote human health? How does fundamental research, with its ultimately global scope, appropriately navigate the ethical and legal aspects of health data science? EMBL is well positioned to lead in this area, with its interest both in research and providing research infrastructure, and its leadership in national, European, and global initiatives. There are two key ways in which EMBL will fulfil this leadership role. First, EMBL will participate in the appropriate ethical and data governance components, both in international framework-setting forums (e.g. the GA4GH Regulatory and Ethics Work Stream) and as technical experts, providing options for national discussions (e.g. UK Biobank and GHGA) to enable data sharing. Second, EMBL will provide the technical delivery of complex engineering to enable responsible federated access, again at both a global international standards level (GA4GH), as the leading European institute in a variety of European contexts (Federated EGA, 1+ Million Genomes), and as a centre of technical expertise in national discussions. This technical expertise is described in more detail in the description of the Genomic Medicine Platform in Chapter 10: Scientific Services.

Understanding and Managing Environmental Risk Factors for Policy Development

A mechanistic, molecular understanding of the biological pathways involved in responses to environmental exposures such as drugs, toxicants and other pollutants can inform both treatment and policy in environmental health (Chapter 15: Public Engagement, Communications, and Outreach). In most cases, however, there is considerable controversy around policy decisions concerning such risk factors, mainly because the appropriate scientific assessment is lacking, and adequate and compelling scientific evidence is needed to effect large-scale changes in practice. For example, the vastly varying national dietary recommendations reveal that there is little consensus about the impact of diet on health. Similarly, parents struggle to eliminate stress factors that might increase the severity of symptoms in autistic children, based principally on trial and error. Through multiscale approaches, EMBL will not only be able to show the specific effects of particular environments, but will also provide examples of how to bring a mechanistic understanding to the physical, biological, and social environmental factors that impact human health. EMBL in its unique transnational position can and must play a key role in this endeavour, through the proposed human ecosystems research and collaborations with European epidemiologists and clinician scientists. Partnering with member state institutes to aggregate expertise across Europe will allow policymakers to benefit from definitive scientific evidence in making key decisions. As Europe's only intergovernmental life sciences research organisation, EMBL can fulfil such a role as a neutral broker for research standards and open science, and can directly contribute to improved, science-driven environmental and health practices across the member states.

Advancing Precision Medicine

The capacity for precision medicine – in which treatments for individuals are selected based on rational, actionable biological information – is poised to expand exponentially, due to unprecedented amounts of molecular information that are now available, as well as the tools to monitor physiological states. However, precision medicine must be guided by solid, ideally mechanistic understanding of human biology. EMBL aims to lead in providing such mechanistic understanding. For example, EMBL plans to build on its track record in the area of cancer and microbiome–tumour interactions, to rapidly translate new human datasets into research actions as well as medical practice. An example of EMBL's impact in the realm of cancer is a recent international multicentre study that led to the identification of germline genetic variants that can predict treatment response in children with medulloblastoma, a common brain cancer. These findings have led to clinical guidelines in several countries, which recommend that all Sonic hedgehog-driven medulloblastoma patients should have their genome sequenced prior to receiving therapy, so that radiation therapy can be ruled out in children carrying *PTCH1* germline mutations, and secondary cancers avoided. In the new Programme, EMBL's human ecosystems research will push the limits of such precision medicine approaches. EMBL will collaborate with national clinicians and healthcare professionals to develop and disseminate innovative methods and tools, and to offer impactful examples of medical practice informed by basic research insight, which can then be more widely adopted in national healthcare systems.

7. Planetary Biology

Background

Planetary biology at EMBL is a new and ambitious initiative to understand life in its natural context. Organisms live in complex and dynamic ecosystems where biological communities – whether of microbes, animals, or plants – interact with each other, and with physical and chemical factors. All living systems are exposed to constantly changing, naturally occurring or anthropogenic environments. To decipher how life adapts, prospers, or declines in different contexts, the biotic and abiotic **environmental factors** that influence organisms must be identified. The **responses** that these factors trigger, as well as the underlying molecular mechanisms, must be elucidated to gain a true understanding of the basis of life. Untangling the complex relationships between organisms and their environments remains a longstanding scientific challenge. Exploring these relationships in molecular terms at the ecosystem scale represents an even greater challenge, yet this is the context in which life happens. Indeed, just as the cell is the basic functional unit of life, ecosystems are the basic unit of nature.

To better understand ecosystem function, new information at the molecular, cellular, and organismal levels will be key. For example, recent studies to unravel the mechanisms of coexistence between microbial communities and their hosts (holobionts) have required a real-time understanding of the gene regulatory networks and metabolic pathways that respond to nutritional variables. The emergence of new viruses and the spread of antibiotic resistance can only be understood by following microbial dynamics *in situ* in their naturally occurring habitats. The processes of rapid adaptation, pathogenicity, or mutualism between organisms require an understanding of community composition, genetic and epigenetic processes (such as horizontal gene transfer, transposable element activation, and environmentally sensitive gene expression), and defence pathways (such as RNA interference and immune responses).

In this highly **collaborative** endeavour, EMBL will leverage its expertise in **molecular biology to study ecosystems**, an area until now mainly studied by ecologists, epidemiologists, geobiologists, and environmental biologists. EMBL will employ a cross-disciplinary approach involving **experimental** and **theoretical** collaborators from various fields. Planetary biology research will be carried out in: (1) **natural contexts** (Figure PB1) using longitudinal, multidimensional sampling of microbes, animals, and plants, particularly at land–water interfaces; (2) **laboratories** at EMBL and at collaborating institutes around Europe for analysis and experimentation; and (3) **environmentally controlled ecosystems** (ecotrons, mesocosms). Carefully selected natural and experimental ecosystems will be studied to generate repertoires of biodiversity coupled with environmental variables to understand function and perform experiments to elucidate mechanisms.

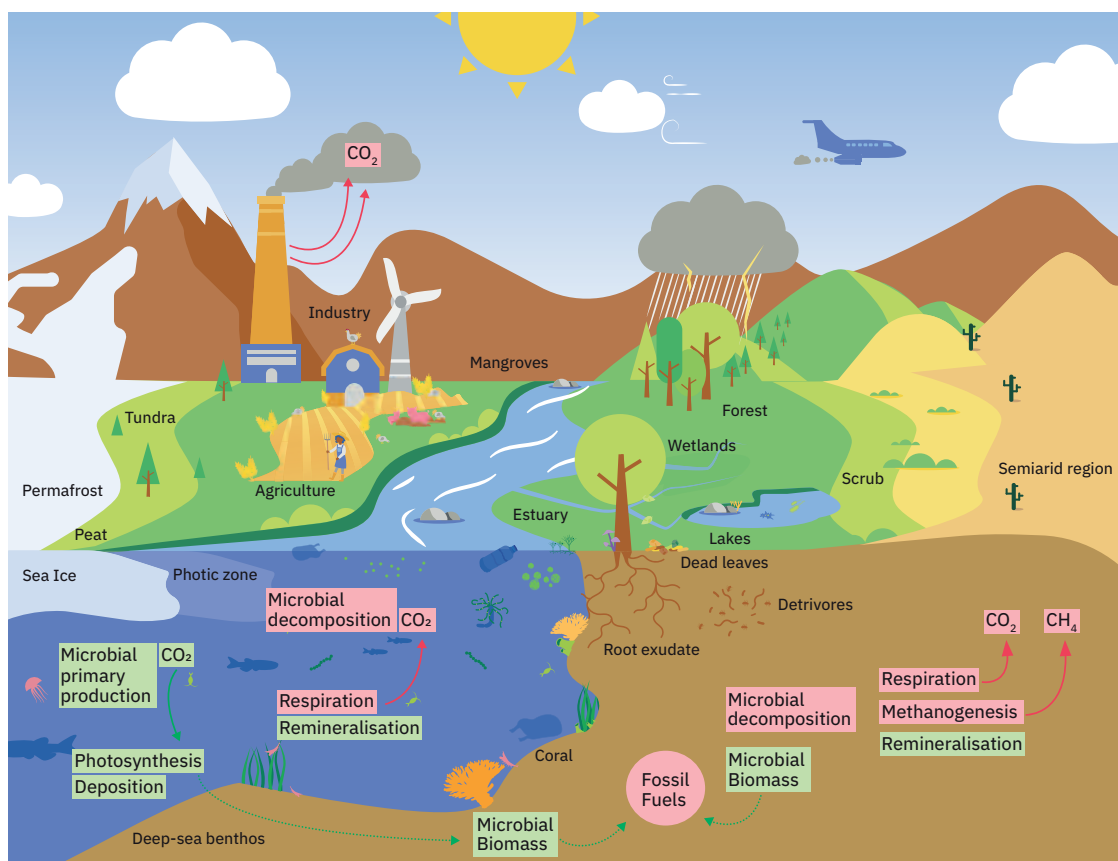


Figure PB1 | Example ecosystems, their biotic and abiotic components, and processes that are within the scope of EMBL’s planetary biology research projects.

One of the main challenges of characterising ecosystem function is identifying how the biotic and abiotic components interact to influence processes such as nutrient cycles and energy flows. Adapted from Cavicchioli R, Ripple WJ, Timmis KN, *et al.* Scientists’ warning to humanity: microorganisms and climate change. *Nat Rev Microbiol.* 2019;17(9):569-586. doi:10.1038/s41579-019-0222-5.

Planetary biology research should provide new molecular tools and expertise to track and understand biodiversity and ecosystem functions, including primary productivity and nutrient cycling, which are linked to environmental sustainability. The identification of new model organisms that are either keystone species (those that play a crucial role in ecosystem function) or dominant organisms (the most abundant species within an ecosystem by biomass) will also have far-reaching implications for scientific discovery and applications in human health and other areas, such as in fisheries or agriculture, where knowledge of ocean, freshwater, and soil health is critical. Aside from the fundamental importance of understanding life in its natural context, ecosystem function is crucial for human survival (for example, for food production, medicine, consumer goods, and materials). However, as biodiversity declines at alarming rates due to habitat destruction, deforestation, urbanisation, pollution, climate change, and other anthropogenic effects, natural ecosystems are changing on a massive scale. Some consequences include the emergence of new infectious diseases that affect humans and animals; the spread of antimicrobial resistance (AMR), which is predicted to cause 10 million deaths each year by 2050 due to drug-resistant infections; climate change and the erosion of soils, which are causing aridification and increased agriculture risks around the world; and the loss of forest and ocean life, which is impacting natural resources. To stop and reverse these effects, an understanding of the factors that destabilise ecosystems, and the impacts of these factors, is essential.

Just as scientists have developed molecular diagnostics and therapies for human health, **molecular approaches** can provide potential diagnostics and potential therapies for **planetary health**, which is intrinsically linked to human health. For example, obtaining fundamental molecular mechanistic insights into zoonotic diseases, horizontal transfer of AMR genes, or eutrophication would be major steps in the search

for methods to overcome or prevent these phenomena. EMBL will seek to build bridges not only between scientific disciplines but also with policymakers to ensure that evidence-based actions and decisions are taken (Chapter 15: Public Engagement, Communications, and Outreach).

Through its initiatives in planetary biology, EMBL will seek to drive a new era for the life sciences in Europe, combining data-driven and hypothesis-driven research in partnership with ecologists and other experts. **As Europe’s only intergovernmental life sciences research institution, EMBL regards this as a unique and compelling opportunity to form new interdisciplinary collaborations with scientists across its member states, fulfilling its pan-European mandate.** The knowledge gained from EMBL’s planetary biology research will enable scientists to answer fascinating fundamental questions about the impact of the environment on biological processes, and will pave the way for solutions to pressing environmental and societal challenges, including global warming, environmental pollution, harmful zoonotic diseases, and AMR.

The Opportunity

The remarkable technological advances of the past half century, as well as the recent unprecedented capacity to collect, store, and analyse data, mean that molecular biologists can now effectively study life in its natural context. These advances coincide with urgent environmental and societal challenges relating to the rapid deterioration of ecosystems, which require solutions based on scientific discovery and evidence. With its planetary biology theme, EMBL will build upon its strengths in fundamental molecular biology research to capture and explore environmental information and the responses of organisms to environmental variations. Although the goal is ambitious, EMBL has a long track record of **successful technology development, multidisciplinary collaborations, and advanced data integration and provision.**

Planetary biology research will build on EMBL’s ongoing research and past experience. EMBL has a successful track record in **global ecosystem exploration**, thanks to highly collaborative projects with the Tara Ocean Foundation and recent soil microbiome studies (Figure PB2). The Tara Oceans expedition was devised by EMBL scientist Eric Karsenti and took place from 2009–2013. This pioneering, cross-continent scientific survey of planktonic and coral ecosystems in the context of climate change produced a tremendous amount of data, which has enabled major discoveries and secondary research around the world. The publicly available resources represent the largest sequencing effort in marine science and the most expansive description of the world’s largest cohesive ecosystem – the ocean. The success of Tara Oceans was in part due to standardised sampling and analysis protocols, deep environmental surveillance, and open data policies. The Tara Oceans data led to a multitude of discoveries about the diversity of viral, prokaryotic, and eukaryotic species, their genetics, and their interactions, including novel symbiotic relationships. Major ecological insights included the discovery of latitudinal gradients of biodiversity in the ocean, how diatom evolution relates to changing climates, and how plankton are dispersed by ocean currents. Recent groundbreaking contributions from the Tara Oceans data include estimates of the distribution of the world’s biomass, new discoveries about the origins of mitochondria, identification of heterotrophic nitrogen-fixing bacteria in the ocean, the discovery of a new class of microbial rhodopsins with potential applications in optogenetics, and the discovery of ‘giant’ marine phages with new CRISPR–Cas systems.

EMBL has also participated in major projects that involved sampling global soil microbiomes. Soil microbiomes are among the most diverse on the planet and carry out critical processes including support of plant growth, nutrient cycling, and carbon storage. Sampling studies at EMBL have provided remarkable insights into fungal and bacterial communities with diverse responses to precipitation or pH, and evidence for bacterial–fungal antagonism (inferred from antibiotic production), indicating the importance of biotic interactions in shaping microbial communities. These examples illustrate the remarkable contributions and opportunities that systematic large-scale ecosystem exploration can provide.

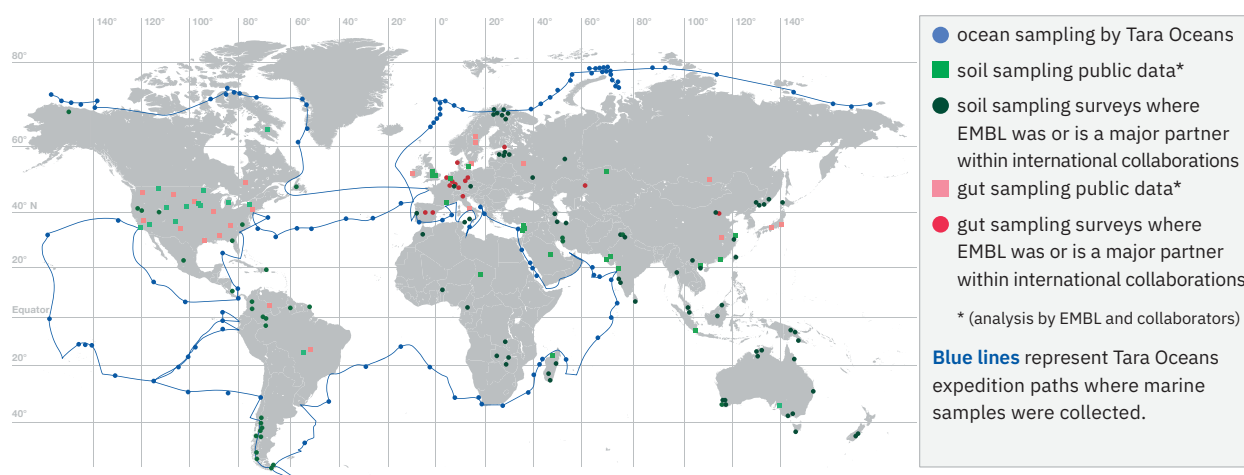


Figure PB2 | Sampling and analysis of global microbial habitats.

Publicly available survey data (represented by lighter dots) can be combined with surveys where EMBL was or is a major partner within international collaborations (represented by darker dots). Most global microbial surveys have been specific to a single habitat such as the ocean (blue; special issue in *Science* 2015; Gregory *et al.* *Cell* 2019; Salazar *et al.* *Cell* 2019), soil (green; Bahram, Hildebrand *et al.* *Nature* 2018), or human gut (red; Wirbel *et al.* *Nature Medicine* 2019; Almeida *et al.* *Nature* 2019).

EMBL has a strong tradition of research in studying **model organisms**, both unicellular and multicellular. The study of how these organisms adapt to environmental changes will be an important pillar of EMBL's planetary biology research. Current research includes investigating developmental dynamics and phenotypic plasticity, as well as responses to environmental signals in diverse organisms such as the marine annelid *Platynereis*, the sea anemone *Nematostella*, and the ascidian *Phallusia* (Chapter 3: Cellular and Multicellular Dynamics). These studies have been made available to the scientific community through multimodal atlases that combine whole-organism, cellular, and subcellular imaging with single-cell transcriptomics, epigenomics, and metabolomics. EMBL also has a distinguished history of carrying out studies on genetics and developmental biology in fruit flies (*Drosophila*) and mice (*Mus*). These foundations are complemented by more recent metabolic analyses and experimental setups to study phenotypic diversity due to genetic variation and environmental factors, including microbiota. EMBL's expertise in using such models, combined with the array of technologies now available for establishing and studying new model organisms, will be crucial to EMBL's planetary biology research.

EMBL's aim is not to explore all ecosystems, but to focus on dedicated collaborative projects that will serve as **paradigms for future endeavours to study ecosystems at the molecular level**. The intention is not to become an ecological research organisation, but rather to collaborate and build bridges with communities of scientists to build a new era of European life sciences together.

Research Aims

A new era of molecular and cellular approaches to ecosystems

Most natural variation studied to date has been limited to observable characteristics and genomic information. Today, technologies exist to explore molecular and subcellular information that varies in response to the environment, including at the levels of stress and metabolic responses. Spatial phenomics, which combines single-cell metabolomics, gene expression, and epigenomics with advanced imaging at the molecular, subcellular, whole-organism, and community levels, has opened up a new era of investigation.

The main objectives of the planetary biology theme will be to **recognise and understand phenotypic changes that are induced in the natural environment**, using the array of tools available for molecular, structural, genomic, cellular, and developmental biology, and the powerful technologies that enable visualisation and perturbation of processes. Genotype–phenotype relationships and the contributions of genetic variation (both between and within species), and environmental factors in influencing phenotypes will be explored in their natural context on the molecular and cellular scales and at the ecosystems level. By measuring the tempo of genetic, phenotypic, and metabolic variation of organisms in their communities; their interactions with each other (e.g. infection, symbiosis, mutualism, competition); and responses to natural variables and environmental stressors, the complex relationships between organisms and their environment can be untangled.

To gain molecular mechanistic insights into organisms within ecosystems, EMBL will first determine ‘who is there?’ (Figure PB3) via sampling, measurement, and analysis. This will be followed by exploration of mechanisms – ‘what do they do?’ and ‘how do they do it?’ – through isolation, cultivation, and perturbation studies in the lab and in controlled settings. Advanced data analyses (including AI) linked to EMBL’s Data Sciences programme (Chapter 8: Data Sciences), as well as theoretical approaches (Chapter 9: Theory at EMBL), will be crucial to fully identify and understand significant correlations and to create testable predictions. The discovery of indicator species which reflect environmental states, and new biomarkers and biosensors to measure the states of ecosystems in nature, can then inform further experiments, enabling new measurements and further perturbation methods.

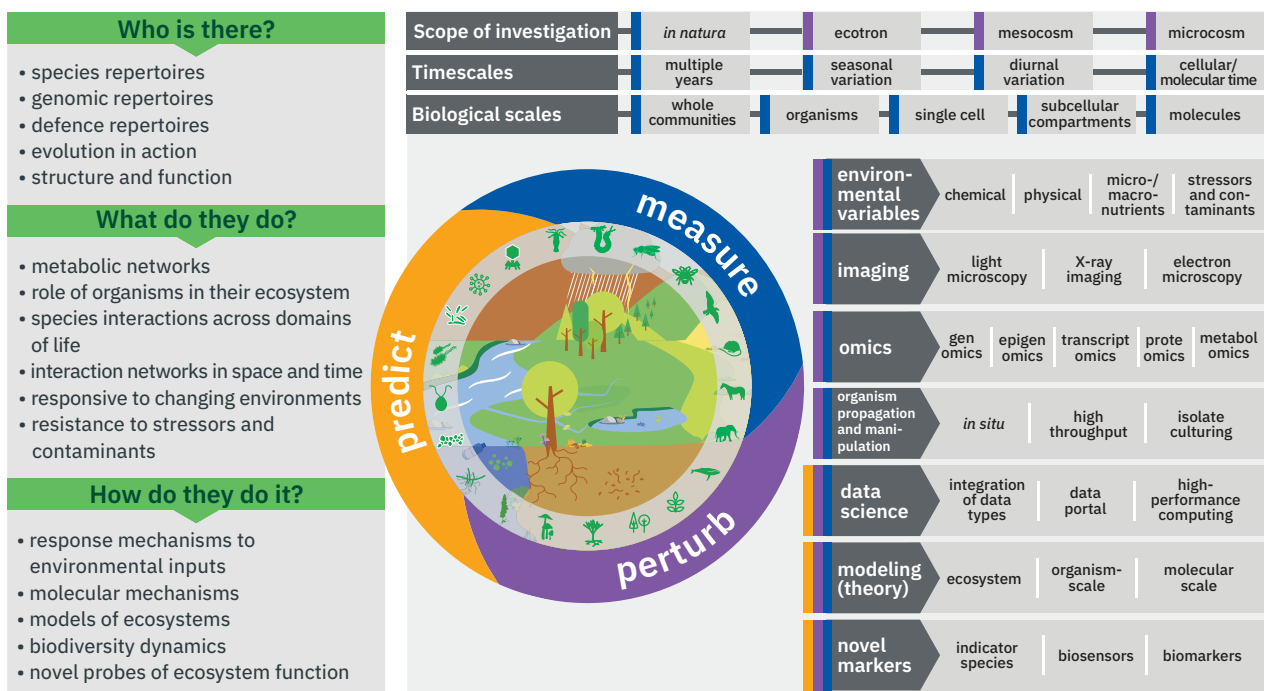


Figure PB3 | Planetary biology processes and questions.

EMBL’s planetary biology processes will enable questions (left) to be asked using the experimental, computational, and theoretical approaches applied in natural and laboratory contexts (right). The processes by which scientists measure (blue), perturb (purple), and predict (orange) can encompass various combinations of relevant tools, techniques, or parameters.

Today, molecular ecology – the application of molecular techniques to ecosystems research – focuses mainly on the use of metagenomic information (sequencing of DNA in environmental samples to obtain a (meta) genomic blueprint), metatranscriptomics (sequencing of RNA in environmental samples to observe gene expression states), and barcoding (16S rRNA for prokaryotes, 18S rRNA for eukaryotes, and the internal

transcribed spacer (ITS) region for fungi) to identify species. Although metagenomics provides repertoires of genes and species within ecosystems, it cannot provide multidimensional spatio-temporal information about how organisms coexist (in communities, as symbionts, etc.) or how they change their molecular and cellular phenotypes in different environmental conditions, be they nutritional variations, natural seasonal changes, or environmental insults such as abnormal temperature or the presence of pathogens or chemical toxins.

To meet this need for **multidimensional measurements**, planetary biology research at EMBL will combine **advanced omics technologies** (including single-cell transcriptomics and metabolomics), with **multiscale imaging** (e.g. high-resolution fluorescence microscopy and correlative light and electron microscopy). Quantitative and dynamic measurements of these multimodal parameters need to be generated in living systems (Chapter 3: Cellular and Multicellular Dynamics). Importantly, these multiscale phenomics technologies at EMBL will have to incorporate and monitor **chemical and physical variables** at the same time. Parameters such as temperature, weather conditions, salinity, pH, oxygen levels, water, abundance of macro- and micronutrients, and biomass of organisms will be measured in conjunction with levels of sulfides, nitrous oxide, nitrate, nitrite, hydrogen, and pollutants from human activities such as micro- and nanoplastics, antibiotics, herbicides, pesticides (e.g. endocrine disruptors), and fertilisers.

High-throughput molecular data, structural biology studies, and high-resolution imaging on a range of scales will lead to many hypotheses, which will need to be tested. Measuring and predicting the response to stressors requires the controlled introduction of **perturbations**. The tools already available at EMBL to manipulate genomes, RNA, or proteins under different conditions and to visualise and measure the impact of perturbations both *in vitro* and *in vivo* will be used (Chapter 2: Molecular Building Blocks in Context). New tools that enable combinations of approaches (e.g. optogenetics, dCas9, and single-cell omics) to perturb and measure molecules will also be developed.

To test **predictions** and gain mechanistic insights into how organisms exist in communities and respond to environmental factors, a range of approaches will be applied. Customised, untargeted metabolomics and chemical screens with a large library of known biotic and xenobiotic substances will be performed using EMBL's Chemical Biology Core Facility, in collaboration with other institutes and networks in Europe. Combinations of substances will be tested, together with physical parameters. Novel assays will be developed using organisms (e.g. insect and plant models), mammalian cells and plant cells in culture, marine organisms in aquaria, and bacterial communities in bioreactors (see Chapter 4: Microbial Ecosystems).

Specific research aims

Some general questions include how responsive or resistant organisms are to environmental changes in natural contexts; for example, how and why biodiversity changes within an ecosystem, how stable ecosystems are, or how reversible the impact of environmental changes can be. Such questions are being asked by ecologists, evolutionary biologists, and epidemiologists. EMBL aims to bring new perspectives and methods to help answer these questions in three broad areas of study. **EMBL aims to:**

- I. **Explore how organisms in communities are affected by natural and anthropogenic environmental variables.** EMBL aims to undertake longitudinal biotic and abiotic sampling with advanced technologies in mobile labs. This will form the basis of multiple projects focusing on microbial ecosystems, as well as specific animals and plants in soil, sediment, aquatic, or arid land contexts. A major area of interest will be capturing biodiversity and accompanying natural and anthropogenic environmental variables at interfaces such as coastal regions, lakes, and rivers, where environmental gradients exist.

Specific studies will include surveying plankton communities at various temporal and spatial scales, across various environmental gradients, to unveil their diversity and phenotypic plasticity; identifying key taxa in microbial communities using molecular methods, to understand how nutrients moving from land to water lead to eutrophication in seas, lakes, and rivers; studying soil microbes to understand the productivity of soils and the impact of synthetic products such as pesticides; understanding the spread of AMR by following microbes in changing soil and water ecosystems to see how antibiotic resistance genes are disseminated; and exploring microbial biodiversity shifts in regions undergoing rapid aridification. To understand gradual versus acute changes induced by the external environment, the variables that affect potential adaptations must be monitored over time, followed by investigations of changes in communities in controlled laboratory settings (see Aim III below) to derive mechanistic insights from the molecular to the systems levels.

- II. **Understand responsiveness to natural and anthropogenic environmental changes using model organisms.** Model organisms studied at EMBL can be used to investigate the impact of specific environmental variations. The aim is to identify the cellular mechanisms that underpin responsiveness and the environmental factors that drive organism adaptations and ultimately their evolution. For example, the marine annelid *Platynereis*, with its highly stereotypic development, can be used to investigate the impact of multiple environmental variables, including light, temperature, acidity, nutrition or the relationship between associated microbiomes and cellular morphology, physiology, metabolism, and behaviour. The sea anemone *Nematostella*, with its remarkable phenotypic plasticity, can be used to understand the impact of food availability on tentacle development and other features (Chapter 3: Cellular and Multicellular Dynamics). Medaka fish (Chapter 6: Human Ecosystems) can be used to examine the impact of environmental toxins on multiple phenotypes, including phenotypes relating to reproduction and behaviour. To understand how animals can adapt to environmental effects, such as temperature, nutrition, or toxins, the classic *Drosophila* model can be used for controlled evolution experiments. Other important models for the study of environmental factors that are not currently used at EMBL – such as plants – will be explored through collaborations and by potential new faculty recruits. More advanced goals will be to identify key organisms or communities that can be used as **new models** to answer particular questions related to the environment. These organisms can be made genetically tractable and amenable to exploration through modern isolation and culturing techniques, and by the development of chemical and genetic perturbation approaches, to enable **new mechanistic insights** and the development of tools for further exploration. These new model organisms will be studied in labs, *in natura*, and in the context of environmentally controlled settings. Within ecosystems of interest, the identification of keystone species and dominant organisms that can be rendered genetically tractable, would provide a crucial starting point for studying interactions within and between species, and for modeling ecosystems in the lab, to understand how they work.
- III. **Study specific communities under controlled environmental conditions.** By culturing communities and organisms under controlled variations of environmental conditions, the aim is to explore ecosystems in natural contexts, or to create simple synthetic ecosystems in the lab, to analyse metabolic repertoires, interaction networks, and response or resistance to changing environments. This is already a reality at EMBL in studies of selected small ecosystems, such as the microbial subcommunities in the human gut, which are being carried out in defined, closed settings (anaerobic chambers). In these studies, chemical perturbations are employed to dissect microbial interactions and to measure the impact of environmental factors, including the effects of therapeutic drugs with direct clinical relevance (Chapter 4: Microbial Ecosystems, and

Chapter 5: Infection Biology). EMBL will seek to advance this framework and apply it to other ecosystems.

Select ecosystems will be analysed in bioreactors, mesocosms, or ecotrons to understand specific environmental effects of nutrients or hydrological, morphological, thermal, or xenobiotic factors. EMBL's experience in building bioreactors will be leveraged to usher in a new era of ecosystem investigation. One example study is an investigation of the biotic and abiotic variables in arid soil ecosystems that are in danger of losing their capability to support productive ecosystems. A specific goal would be to identify the variables that could be used to detect vegetation decline and the onset of the soil disruption phase, to better predict systemic collapse.

By defining biotic and abiotic parameters and how they vary over time in natural ecosystems, ecotron research infrastructures can be used to study multiple questions about when, where, how, and why mutualistic, pathogenic, or symbiotic relationships occur, particularly in the context of environmental changes such as temperature, nutritional variation, toxin exposure, or stress. Example communities include bacteria and fungi in soil, plants and their microbial communities, or oceanic microalgal symbionts in acantharian hosts. These studies can lead to significant progress in explaining what underlies certain types of phenotypic plasticity in nature, how zoonotic diseases and AMR spread under various conditions, and the impact of the environment on early development, which can lead to phenotypic changes in later life and intergenerational effects in some organisms.

Planetary biology research principles:

- **Collaborative research.** To maximise scientific discovery and societal impact, EMBL will work with scientists from many disciplines through a scientific advisory board and networks of collaborators, including molecular biologists, population geneticists, ecologists, geobiologists, engineers, systems biologists, and data scientists. New expertise will be brought in via collaborations and new recruitments (Chapter 1: Introduction, Figure IN2). By deploying mobile lab services, EMBL will join forces with scientists from its member states in new collaborations, and will provide training via postdoctoral schemes and specialised sabbaticals (Chapter 11: Training), equipping a new generation of scientists to carry out planetary biology research. To maximise cooperation, EMBL will publish calls for joint ecosystem research projects and will organise stakeholder meetings to select and develop projects that would be timely, would foster excellent research, and would ensure the highest scientific impact.
- **International scope and pan-European mandate.** Research on ecosystems is often limited by national borders. Planetary biology at EMBL provides an opportunity to link national environmental science projects that might otherwise be geographically restricted. EMBL also aims to link to international networks and infrastructures such as European marine biology stations, via the European Marine Biological Resource Centre (EMBRC-ERIC); ecotrons, via the Analysis and Experimentation on Ecosystems (AnaEE) infrastructure; and the UK's Darwin Tree of Life project and the international Earth BioGenome Project. As an intergovernmental organisation, EMBL is ideally situated to collaborate with national initiatives to pursue and further such societally relevant research in an integrative manner.
- **Technology development.** Profiling ecosystems at the molecular and organismal levels will require technical innovations such as automated environmental sampling, advanced omics technologies (nanopore sequencing or metabolomics), single-cell phenomics, improved genetic and chemical screening, and customised modern microscopy techniques (from cryo-EM

to super-resolution microscopy and live-cell imaging with chemical probes). Equipment for sample preparation and some *in situ* analyses will also be needed when field studies are performed. This will enable data to be simultaneously captured at multiple levels – genetic, transcriptomic, epigenetic, metabolic, and morphological – and to be analysed in an integrated way.

- **Advanced data integration and provision.** EMBL is a world-renowned hub for the integration and provision of various types of molecular data, with expertise and critical mass in data storage, processing, handling, harmonising, and analysis. This will be essential for capturing and leveraging the enormous and ever-growing quantities of data obtained from global high-resolution surveys of ecosystems. EMBL's commitment to providing open access data and analysis tools will guide these activities.


EMBL's Approach

I. Explore how natural communities are affected by natural and anthropogenic environmental variables

EMBL will undertake **longitudinal sampling** projects, applying advanced molecular and cellular approaches to describe ecosystems exposed to natural and anthropogenic variables. The focus will initially be on specific soil and aquatic ecosystems at **land–water interfaces**. These interfaces are important for biodiversity and ecosystem functioning, and several recent studies have shown that coastline areas have suffered rapid declines in biodiversity and higher rates of antimicrobial resistance spread, with gradient effects going from land to sea. Synchronised land and aquatic sampling will allow evaluation of the impact of natural and anthropogenic factors on soil, rivers, and possibly air, as well as the impact of wind and floods on land. The focus will be on understanding the effects of these factors on terrestrial and aquatic microbial biodiversity and ecosystem functions.

To realise this ambitious endeavour of studying ecosystems at scale, EMBL proposes to establish mobile labs, with state-of-the-art technologies, to conduct carefully planned sampling expeditions in collaboration with scientists from EMBL's member states, developing standardised protocols and applying unified data science concepts. **Mobile labs** will combine **research** and **services**, together with **training** and **outreach** activities. The container-based mobile services (Chapter 10: Scientific Services) will bring advanced equipment to the participating and collaborating labs throughout each expedition. Science and service will be combined, with numerous training and outreach events in the member states (Chapter 11: Training, and Chapter 15: Public Engagement, Communications, and Outreach).

Standardised protocols will be developed to measure environmental variables, capturing a maximum of information *in situ*. The concentrations and fine-scale chemical gradients of oxygen, sulfide, nitrous oxide, nitrate, nitrite, and hydrogen in the natural environment can all be monitored using **microsensors**, enabling high-quality microscale measurement of analytes. Measurements will be used to resolve this environmental information in heterogeneous settings, such as the root surfaces of plants, animal burrows in marine sediments, microbial communities, porous soils, and wastewater locations. Alongside this, the consumption rate of oxygen over time, or the oxygen exchange rate across the water–sediment interface, can be investigated in different geographical locations. Other environmental variables of relevance include temperature, weather conditions, salinity, and levels of macro- and micronutrients. As the relationships between organisms and the environment continue to be studied, technological developments to create new **biosensors** will be enabled by collaborations between molecular biologists, engineers, and chemists at EMBL and within its member states. These biosensors could potentially be used as novel diagnostic tools to help identify ecosystems under threat (see also Tools for Planetary Health below).

 An ocean–land sampling pilot project was carried out by EMBL in collaboration with the Stazione Zoologica Anton Dohrn in Naples in October 2019, to test some of these approaches, including the appropriate tempo of sampling. A longer-term planetary biology project is being developed. This example flagship project will develop planetary biology approaches and protocols, and will pioneer the use of mobile labs. Selected communities and model systems will be further analysed in the lab under controlled conditions (see sections II and III).

TREC Flagship Project: Based on EMBL’s experience in global ecosystem surveys of oceans and soils (Figure PB2), EMBL aims to initiate a European coastline expedition called TREC (Traversing European Coastlines), to study microbial communities and populations of selected macroscopic model organisms. Simultaneously serving as a proof of principle and as a flagship project, TREC will aim to characterise marine and terrestrial ecosystems along the European coastline, extending also to freshwater–land interfaces such as rivers and lakes. In cooperation with multiple partners (Figure PB4), the project aims to address the impact of various environmental factors – including pollutants – on microbial communities, as well as on key organisms about which EMBL has expertise. The strategy couples cross-sectional molecular profiling and imaging with in-depth analyses of selected organisms, communities, and contextual environmental data. Numerous locations are being considered, where scientific interest exists and where a wide variety of organisms, including bacteria, fungi, viruses, micro-eukaryotes, plants, and animals, could be collected from land–water interfaces (Figure PB5).

The TREC expedition will rely on **core resources, infrastructure, and expertise** provided by the core partners, and there are initial plans for three projects (described below). This core support will also be used to facilitate numerous plug-in projects with member state scientists who wish to undertake cross-sectional or longitudinal studies. These plug-in projects will be selected based on existing consortia or open calls. EMBL will partner with scientific institutions and link up to national projects or complementary consortia.

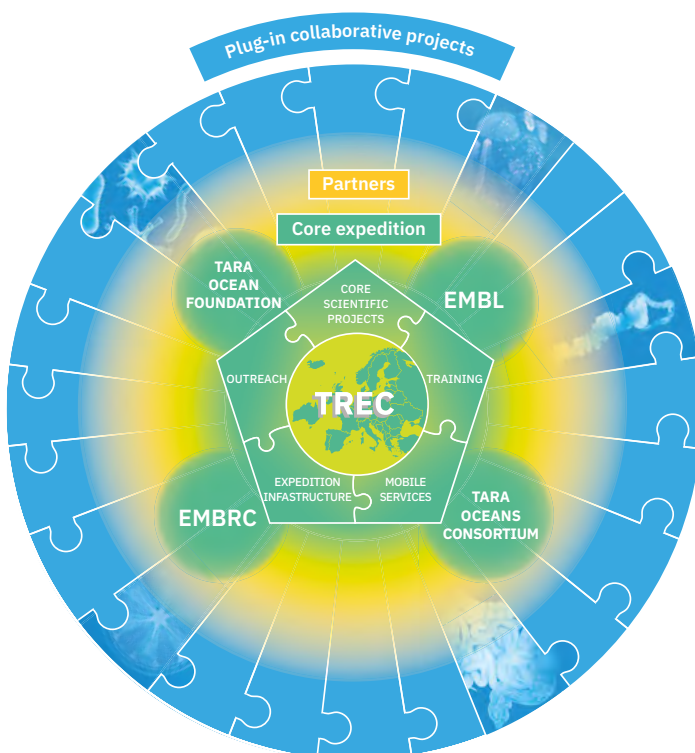


Figure PB4 | Organisational structure of the TREC expedition.

The central ring (green) represents the core expedition, comprising core infrastructure and core activities. Current core partners are the Tara Oceans Consortium, the Tara Ocean Foundation, and the European Marine Biological Research Centre (EMBRC-ERIC). Additional partners (yellow) will support and leverage TREC infrastructure, technologies, and expertise. Plug-in collaborative projects (blue) from member state scientists and others can be integrated to maximise the scientific and societal impact of the expedition. Core partners will support the expedition with resources, infrastructure, and expertise. EMBL will provide much of the land-based infrastructure (mobile labs, vehicles, and equipment). The Tara Ocean Foundation and the Tara Oceans Consortium will provide the Tara research vessel and associated equipment and sampling expertise. EMBL will also collaborate with marine stations with EMBRC-ERIC infrastructure and with local scientific communities.

Three examples of **TREC core scientific projects** that could be carried out in parallel are described below (two in this section and the third in Section II of EMBL’s Approach). In each case, additional partners, institutions, or consortia could join to expand and leverage the proposals described below.

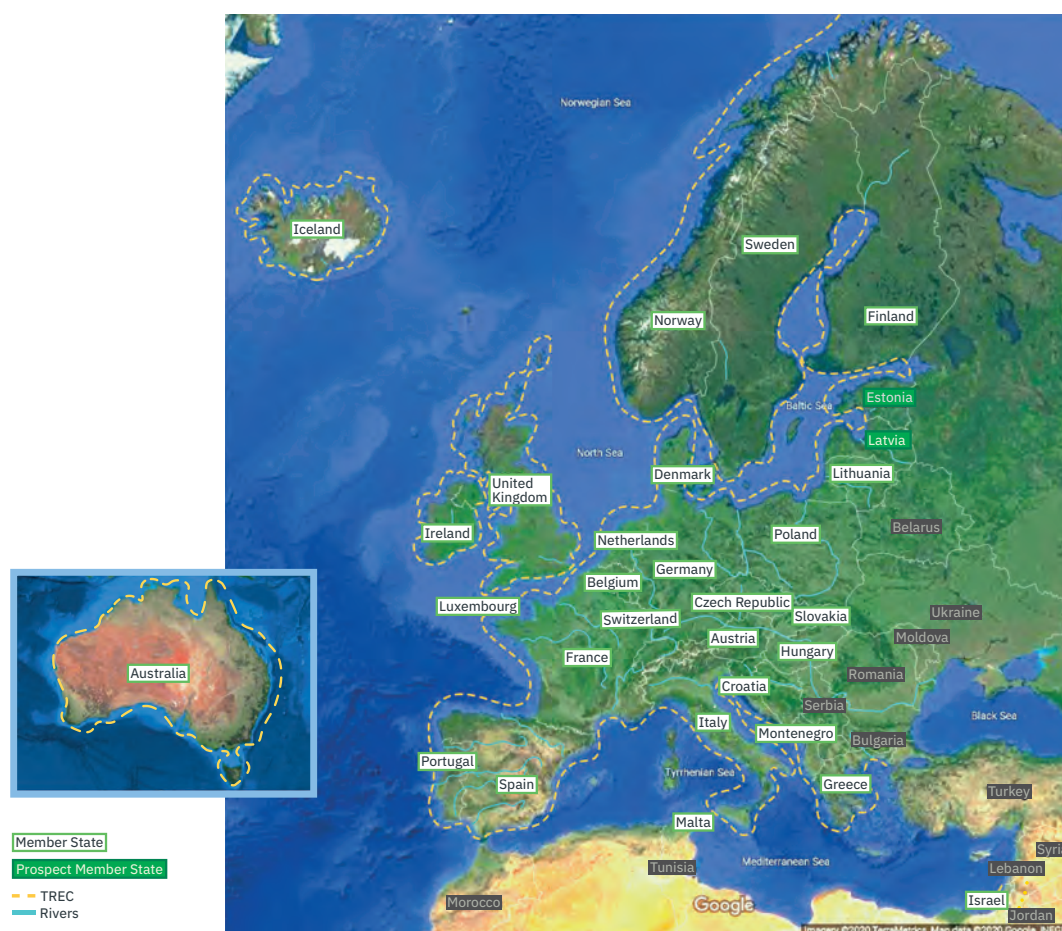


Figure PB5 | TREC expedition potential sampling sites.

This map highlights the coastline of the EMBL member states. EMBL scientists and collaborators will explore marine and terrestrial ecosystems along the European coastline, extending also to freshwater–land interfaces such as rivers (indicated in blue) and lakes. EMBL will continue to work with member state scientists, collaborators, and representatives across multiple disciplines to select sampling sites along the coast and inland.

TREC Core Project 1: Sampling Coastal Plankton Communities

Plankton communities, comprising for example diatoms, dinoflagellates, or other microalgae (phytoplankton), play a key role in Earth’s biosphere as photosynthetic producers in marine waters, yet are still poorly understood. Microalgae have fascinating endosymbiotic states that vary with climatic and nutritional contexts. Given their high protein and oil content, they can also be cultured easily for use as biofuels or animal feeds. The TREC project will sample coastal plankton communities at various scales, across various environmental gradients, to unveil the diversity and phenotypic plasticity of microalgae and zooplankton, including their architecture and metabolism at subcellular levels. By integrating organismal abundance data and data on functional traits, it should be possible to deduce environment-dependent interactions between species. These interactions range from tight symbiotic or parasitic relationships to indirect associations, reflecting nutrient dependencies or contributions to interspecies molecular pathways.

Sampling sites will be selected to provide a balance between relatively undisturbed habitats and those affected by human activity. The former will require extensive systematic screening for pollutants such as microplastics. Sampling and processing protocols will be developed and standardised as was previously done for the Tara Oceans expedition. Protocols will also include measurements of temperature, pH, oxygen saturation, and salinity, as well as quantities of trace elements. Methodologies combining measurements of environmental parameters with high-resolution imaging and single-cell transcriptomic and metabolomic analyses will enable a direct correlation of imaged phenotypes with molecular biotic and abiotic information.

TREC Core Project 2: Sampling Coastal Soil and Sediment Microbiomes

Microbial soil communities along coastlines and in sediments represent dynamic ecosystems that can provide insights into the natural biodiversity at land–water interfaces, the productivity of soils, and the impact of synthetic products such as microplastics, antibiotics, herbicides (e.g. glyphosate), pesticides, and endocrine disruptors. Soil ecosystems contain viruses, bacteria, archaea, fungi, protozoa, and nematodes. Topsoil microbiomes are critical for the support of plant growth and the cycling of carbon, nitrogen, and other nutrients. To understand soil functioning, it is necessary to model the distribution patterns and functional gene repertoires of soil microorganisms, as well as the biotic and environmental associations between the diversity and structure of soil communities.

Recent studies have revealed that both competition and environmental filtering affect the abundance, composition, and encoded gene functions of bacterial and fungal communities, indicating that the relative contributions of these microorganisms to nutrient cycling varies spatially. These studies have also revealed the existence of bacterial–fungal antagonism in topsoil and ocean habitats, inferred from antibiotic resistance genes (ARGs). To outcompete bacteria, many fungal taxa secrete substantial amounts of antimicrobial compounds, which may select for antibiotic-resistant bacteria and effectively increase the relative abundance of ARGs. These genes can provide indications about correlations between antibiotic producers and associated species, but their abundance profiles can also reveal patterns of antibiotic biogeography and dispersal. In particular, it has been proposed that chemical pollution in urban and industrial areas, as well as the extensive use of antibiotics in farming, have largely contributed to the spread of AMR and the emergence of multidrug-resistant bacteria (Chapter 5: Infection Biology).

The goal of this sampling programme will be to gain a deeper understanding of diverse soil microbial communities, their adaptations to different environmental factors – including pollution – and their exchanges with adjacent marine microbiomes. Studies will also explore microbial communities for antibiotic production and associated resistance. EMBL will collaborate with local partners, including marine stations and agricultural institutes, to join forces with local centres of ecological and geobiological expertise. The study of sediment is also important, since it can be key to understanding the reproductive habitat of many organisms and the formation of beaches, sandbars, and estuaries. Importantly, sediment provides unique insights not only into living ecosystems at land–water interfaces, but also into historical biodiversity, since ancient DNA can now be collected and analysed to obtain insights into human impacts during industrialisation, environmental impacts during the Medieval Warm Period or Little Ice Age, and deeper in time. The resulting knowledge could be used to develop molecular biomarkers for pollution states and to explore bioremediation strategies.

II. Understand responsiveness to natural and anthropogenic environmental changes using model organisms

To conduct in-depth analyses and establish a baseline understanding, EMBL will investigate classic model systems and establish new model systems. These will range from individual species with close symbiotic relationships to selected microbial communities, which will be characterised in their environmental context *in natura* and under controlled lab conditions. Model species, model systems, and selected ecosystems will be characterised using systemic approaches that include perturbations and comprehensive profiling of molecular, cellular, and phenotypic features. A molecular and mechanistic understanding underlying interactions within an ecosystem is a prerequisite to predict biotic and abiotic interactions in a context-dependent manner and abstract those interactions into more comprehensive models which, in turn, must be validated. Studies of experimental evolution can also be performed to observe evolutionary processes in real time and under various conditions, to study adaptation and estimate evolutionary parameters.

EMBL also aims to identify key organisms or communities that can be used as new models to answer questions relating to the environment, and which can be made genetically tractable. Collaborations with other large national or international initiatives to characterise biodiversity on Earth will also support the identification of ecologically important keystone species and dominant organisms. Initial collaborative efforts include the Earth BioGenome Project (a global effort to sequence all 1.5 million known species of animals, plants, protozoa, and fungi) and the Darwin Tree of Life project (which aims to sequence the genomes of all 60,000 species of eukaryotic organisms in the UK and Ireland). One aim of EMBL's planetary biology theme will be to dissect ecosystem function, either by using single organisms or simple communities exposed to different biotic and abiotic variables, including those defined in longitudinal sample initiatives such as TREC (for example, photosymbiosis between eukaryotic microalgae and unicellular hosts along environmental gradients; interactions between fungi and soil bacterial communities in the context of antibiotic exposure and antibiotic resistance). Some illustrative examples are described below.

TREC Core Project 3: Sampling Selected Animal and Plant Species Along Environmental Gradients

EMBL will use its knowledge and expertise to select model species and sample and analyse them in more depth along European coasts. For example, the marine annelid *Platynereis dumerilii* has been extensively studied in evolutionary developmental biology and neurobiology, and is responsive to environmental signals such as light and photoperiod, and chemical cues including alcohols, esters, amino acids, and sugars. *P. dumerilii* usually inhabits bright, shallow marine environments but it can be found in less favourable environments, such as thermal vents or polluted or acidic areas. Subpopulations of *P. dumerilii* larvae have been found in environments with various pH values, indicating that *Platynereis* seems to adapt to a wide range of environmental conditions. Widespread along European coasts, *P. dumerilii* shows multiple haplotypes adapted to varying conditions. It is now possible to map allelic variants, differential gene expression, metabolites, and morphological variation over the entire body of a *Platynereis* larva with cellular resolution (Figure PB6). This also includes measurements of distinct microbiomes to analyse the variable composition and spatial distribution of microbiota contributing to the *Platynereis* holobiont. *Platynereis* can be grown and manipulated in the lab, and is one of several powerful models at EMBL for exploring the impact of environmental variables on development, cellular morphology, physiology, and metabolism, and for studying the links between genetic variants and environmentally induced phenotypic changes. Correlating multiple modalities to sample sites (and thus to environmental conditions) will make it possible to establish links between differing ecological parameters and cellular assimilation and adaptation *in natura*, and will generate hypotheses that can then be functionally tested in the laboratory. In addition to *Platynereis*, several other marine invertebrate model species could be sampled and investigated.

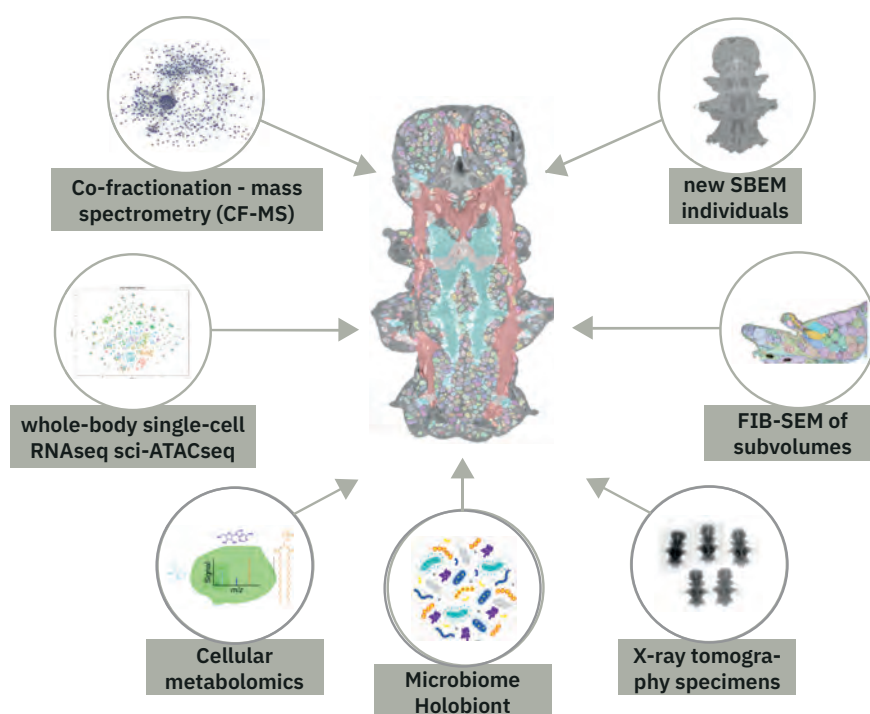



Figure PB6 | The *Platynereis* cell atlas.

This is a unique and collaboratively driven resource that builds on the synchronous and stereotypical development of *Platynereis*, and combines gene expression and ultrastructural information. Cells and nuclei have been segmented and mapped to a cell-type atlas so that segmented cells and nuclei can be correlated with full gene expression information for each cell type in the body. This resolution will be applied to several selected environments to study phenotypic plasticity and adaptation to the environment at cellular resolution. Cutting-edge imaging techniques, including automated electron and light microscopy and X-ray tomography, will be integrated to link morphometry of phenotypes to genetic variation (see Chapter 10: Scientific Services, Tech Dev Box TD3_SS).

Terrestrial organisms that will be investigated include **plants, nematodes, and insects**. Plants in particular are of major interest, as they are key biosphere producers by fixing CO₂ and producing oxygen for the planet. They represent the majority of biomass on Earth, and due to their sessile nature they are exquisite biosensors of environmental change. They have also evolved with a plethora of microorganisms that play important roles in their growth – both above and below ground (phyllosphere and rhizosphere, respectively). A considerable amount of information is now available on the structure and dynamics of plant microbiota, and on the functional capacities of isolated community members. However, much innovation is needed to connect the functional potential of plant microbiomes with plant physiology in natural contexts. The longitudinal sampling of plants and their microbiota along coastlines as part of TREC represents a unique opportunity. Although no plant scientists currently work at EMBL, new group leaders in this area may be recruited in the future, and plug-in projects will lead to collaborations with numerous experts around Europe.

Free-living nematodes including the classic lab model ***Caenorhabditis elegans*** are important components of terrestrial ecosystems. They live in complex habitats (e.g. soil, rotting fruit) in which they routinely experience large fluctuations in temperature and nutrient availability, and are exposed to diverse pathogens including viruses, bacteria, and fungi. *C. elegans* has been an invaluable model organism in research fields including evolutionary and developmental biology and more recently in ecology. Due to its high sensitivity to different pathogens and contaminants, its ease of use for experimentation, and its important functional roles in ecosystems, it is an ideal model for exploring organismal adaptations to changing environments. In a biomedical context, the implication of pesticides in the aetiology of several neurodegenerative diseases has been extensively investigated using *C. elegans* as a primary model organism, with *C. elegans* likely to be extensively integrated into environmental risk assessment procedures. New nematode models are now being

established in many laboratories in Europe due to their interesting biology, including variations in pathogen responses and defence strategies. They can also be used to tackle environmental challenges: for example, the root-knot nematode *Meloidogyne incognita* causes considerable damage to global agriculture and could serve as a model of plant parasitism and pathogenicity.

Insects are the largest and most diverse group of organisms on Earth. A classic model used over the past century by scientists, including those at EMBL is the **fruit fly *Drosophila melanogaster***. Fruit flies can now be used to measure the impact of environmental fluctuations and invasive species *in natura*, in part due to the immense power of genetic, developmental, and phenotypic approaches that exist to analyse *Drosophila* species and their microbiota. *Drosophila* can also serve as an easy and rapid indicator for monitoring local species diversity across ecosystems. These indicators are based on standardised sampling procedures focusing on the identification of key morphologies. During TREC, *Drosophila* species will be collected at each site and analysed for genetic, microbial, and metabolic diversity, each at high resolution. Phenotypic changes due to environmental variables, including insecticide exposure and nutritional changes, can be taken back to the lab and used in a predictive context (see Chapter 3: Cellular and Multicellular Dynamics). The environmental context will be studied by molecular dissection of gene regulatory networks, metabolism, and proteomics at the population and ecosystem levels.  In a pilot project, EMBL scientists are using MALDI imaging mass spectrometry for high-throughput metabolic screening of thousands of *Drosophila* embryos to characterise how their metabolism is reprogrammed upon genetic and environmental perturbations. The aim is to understand metabolic profiles and changes, and link these directly to the uptake of exogenous environmentally critical molecules such as pollutants, herbicides, and unwanted drugs. Experimental evolution in the presence of these toxins will enable the systematic analysis of metabolic evolution in a complex animal.

III. Explore specific communities under controlled environmental conditions

Quantifying the effects of changing environmental conditions at scale and over time by molecular profiling will generate new knowledge about the contributions of genetic variation (both between and within species) and environmental factors in influencing phenotypes. This will, in turn, generate hypotheses that need to be tested and validated under complex but carefully controlled conditions. To fully model natural and anthropogenic environmental changes, ecosystems of interest from planetary biology research will need to be studied in large-scale specialised research infrastructures. Such infrastructures are needed to replicate natural environments and enable scientists to effectively control, measure, and perturb the effects on organism function of specific environmental variables such as levels of nutrients or hydrological, thermal, or xenobiotic factors. The results can be used to create models of complex interactions, ideally moving from perturbation to prediction. This will enable scientists to test for conditions that can change ecosystem properties, and allow desired properties to be introduced or resilience to detrimental factors to be increased. Experiments can start in simple communities, such as carbon fixation in certain algae that form symbiotic relationships with coral, but can gradually build up to more complex ones, such as modelling ecological tipping points in soil and plant microbiomes in arid regions. Depending on the type and complexity of the ecosystem under study, multifaceted microcosms, mesocosms, or ecotrons will be used for ecosystem exploration.

Ecosystem Research Infrastructures

Microcosms are small, artificial, simplified ecosystems that are used to simulate and predict the behaviour of natural ecosystems under carefully controlled laboratory conditions. Associations with environmental factors observed by longitudinal sampling can be followed up in on-site microcosms, with the goal of

performing tests for functional traits or organismal interactions. EMBL has expertise in building and maintaining microcosms and has developed bioreactors to optimise yields for particular microbial products or to stably and reproducibly grow small, defined bacterial communities (Chapter 4: Microbial Ecosystems). The development of bioreactors was a prerequisite for comparing defined perturbations of individual species with synthetic and *ex vivo* gut microbial communities, with the plan to also incorporate host interactions (Chapter 4: Microbial Ecosystems). While robotic systems to automate perturbations are already in place, technology is being developed to increase throughput and to incorporate integrated molecular profiling and imaging technologies of different resolutions to span spatial scales, from viruses to bacterial communities.

Mesocosms are medium- to large-sized, artificially constructed model ecosystems, which provide a link between field surveys and highly controlled laboratory experiments. Mesocosm studies are normally conducted outdoors to incorporate natural variation. Mesocosms provide control over multiple parameters that often confound experiments in natural settings. They also allow controlled testing of toxins that cannot be released into nature, and nutrients that can profoundly disturb ecosystems. Mesocosms are, for example, suited for studying soils, as could be done with samples obtained from the TREC expedition.

Ecotrons are large infrastructures designed to reproduce naturally occurring ecosystems and elements in a simplified way. Advanced equipment and technology make it possible to combine different environmental variables according to a defined time cycle, while avoiding undesired variability. On a technical and operational level, an ecotron consists of large chambers, offering the possibility of studying agro-ecosystems. The variables that can be regulated include light (spectrum, intensity, and photoperiod), air (temperature and humidity), rainfall, wind, CO₂ and ozone concentrations, and boundary conditions. Ecotrons can also be designed to incorporate genetically modified organisms.

EMBL will create such structures when needed, but will preferentially use existing facilities through partnerships. To gain expertise in these infrastructures and explore potential projects and collaborations, EMBL will engage with existing infrastructures in Europe and beyond (for example, member state ecotrons including AnaEE in France, or AQUACOSM, which brings together 19 aquatic inland and marine ecology experimental platforms from 12 European countries). Small model laboratories are sufficient for chemotrophic and anaerobic microbial communities, but other, more complex communities require experimental ecosystems that are closer to natural settings. In this context, plans are under way to establish microcosms and even a small mesocosm in a new building, the Molecular Biology Centre for Human and Planetary Health, at EMBL's Heidelberg site.

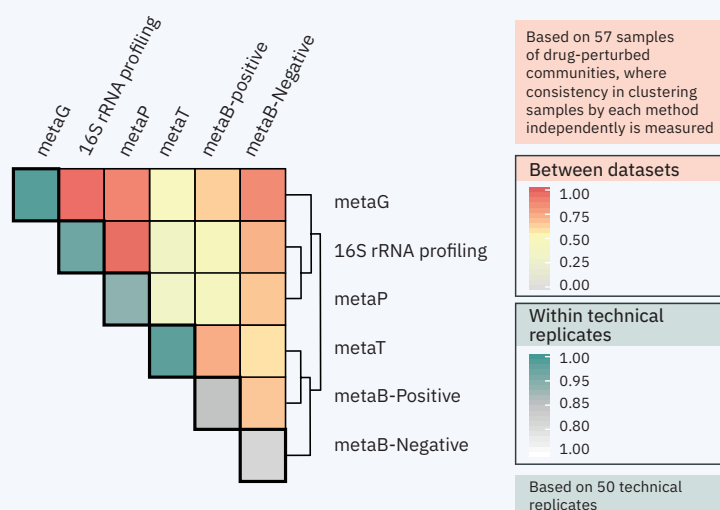
An example of a controlled environment study in collaboration with researchers in Spain relates to **arid ecosystems** (defined by an average annual precipitation of <250 mm), which make up a large proportion of the Earth's surface and are in danger of being pushed over an ecological tipping point, beyond which they collapse into desert and are no longer capable of supporting productive ecosystems. As with many ecosystems, there is currently a limited understanding of the molecular processes involved, or the repercussions of this ecosystem collapse. To obtain mechanistic insights and develop molecular biomarkers, *in situ* systems-level analyses will be complemented with studies in controlled experimental settings. EMBL wants to partner with ecologists and geobiologists to take the first steps in exploring **ecological engineering**, or terraformation, of arid soil ecosystems. The plan is to use mesocosms to bring these intact soil ecosystems into partially controlled laboratory conditions. Here, temperature and CO₂ levels can be increased, and humidity decreased, to model changes in the soil system in response to induced global warming conditions. Longitudinal metagenomic sampling will allow microbe biodiversity to be surveyed, while metabolomics will be used to explore whether there are metabolic signatures that could be used as biomarkers to warn of ecological collapse. The long-term goal is to increase the resilience of ecosystems to climate change by modelling and predicting the functional traits required by microbes in these ecosystems, which can then be introduced by the microbes carrying them.

Data Collection, Processing, Integration, Analysis, and Dissemination

Longitudinal sampling projects and the characterisation of model species and selected ecosystems under controlled lab conditions will generate data that must be analysed and translated into knowledge. Genomic data must be linked with data from other increasingly utilised molecular profiling techniques such as metatranscriptomics, metabolomics, and metaproteomics (Tech Dev Box TD1_PB), or targeted and untargeted chemical profiling approaches. EMBL also intends to develop technologies to integrate molecular profiling data with molecular, cellular, and environmental imaging (such as X-ray tomography of soil), which requires novel bioinformatics approaches. EMBL is developing data management systems to unify and harmonise workflows, and to integrate and disseminate organismal and environmental data (Chapter 8: Data Sciences). Open science will allow data access and reuse, and will stimulate further research in various fields. Data will be publicly released via **specialised integrative data portals** (Chapter 10: Scientific Services) to facilitate data analyses by scientific communities.

Technology Development Box TD1_PB | Towards integration of multiple molecular profiling technologies by bioinformatics.

Several omics technologies will be employed to profile model systems *in vitro* and *in natura*. Metagenomics, transcriptomics, proteomics, and metabolomics data from marine and soil samples must be integrated using bioinformatics methods, accounting for the strengths and weaknesses of different molecular profiling technologies. For example, metagenomics provides an unprecedented genetic catalogue of biodiversity and functional potential, but metatranscriptomics is better suited for discriminating between ancient DNA and that in living cells, and for quantifying functionality. In turn, both of these techniques give only indirect clues about chemical pollution, which can be more directly measured by targeted and untargeted metabolomics. As each profiling method has its own biases, significant efforts must be made to integrate data from these different approaches. Some measures are more reproducible than others (as shown by the ‘Within technical replicates’ key in the figure), as illustrated in a mock experiment with



Benchmarking as prerequisite for integration of molecular profiling: (meta) -genomics, -transcriptomics, -proteomics and metabolomics as well as 16S RNA profiling

a synthetic microbial community of gut bacteria. The analysis of this experiment shows the agreement between different methods in terms of the similarity across 80 samples after different perturbations of this community (as shown by the ‘Between datasets’ key in the figure). Considerable efforts are ongoing to normalise these data to add value to their interpretation. Proper quantification is a stepping stone towards modelling interactions between species, and their interactions with the environment.

Tools for Planetary Health

Detailed knowledge of ecosystem functioning not only serves as a basis for unravelling the crucial molecular mechanisms underlying macroscopic changes, but can also support the development of novel biomarkers for the diagnosis of ecosystem states. As well as the development of novel biosensors (see Section I in EMBL's Approach), planetary biology research has the potential to contribute greatly to biomonitoring methods. These methods can be used to determine the presence of chemicals, and – more significantly – the way organisms respond to exposure to such environmental chemicals, for example from plastic degradation or pesticides. Molecular biomarkers that can be used to measure the effects of environmental pollutants on organisms and communities at the molecular, cellular, organ, and organism level, and can be integrated with chemical monitoring or population studies to determine the toxic effects of pollutants even when they are present at low, sublethal concentrations. Molecular biomarkers may therefore provide a sensitive early warning of adverse effects that could occur later within populations.

Examples of existing biomarkers of exposure include parameters that reflect exposure to a specific class of pollutants in bivalves or fish as sentinel organisms. Such parameters could be tissue levels of metallothioneins, inhibition of cholinesterase activity, peroxisomal proliferation, or mixed function oxidation. However, these biomarkers may not identify the toxic agent responsible, or may not make a distinction between the exposure to or detrimental effect of a pollutant. With the application of molecular biology technologies and planetary biology approaches, it may now be possible to develop more sophisticated molecular pollutant biomarkers. EMBL's expertise in genomics, transcriptomics, proteomics, metabolomics, metagenomics, and high-throughput imaging could contribute to the discovery and further development of improved, sensitive molecular pollution biomarkers. In parallel, complex yet controlled laboratory environments such as microcosms and mesocosms could be used to validate and test results from the field. New molecular biomarkers could be specifically associated with a toxicant's mechanism of action, and could be sufficiently well characterised to relate the degree of biomarker modification to the degree of the adverse effect. In this way, EMBL's expertise in fundamental research and technology development could contribute to solving societally relevant challenges that are currently being tackled by EU environmental risk assessments.

Impact

A fundamental molecular and mechanistic understanding of biological processes in complex ecosystems will have a far-reaching impact, from both a scientific and societal perspective. New scientific knowledge gained through EMBL's collaborative networks and research strategies will usher in a new era of life science research. This knowledge will help to provide solutions for some of the major global challenges of our time, such as climate change, pollution, and increased AMR.

Biotic and abiotic environmental perturbations in model organisms, and modelling and modulation in microcosms, mesocosms, and ecotrons, will be key to moving from repertoires and descriptions to molecular mechanisms. Data on global biodiversity and functional traits from planetary biology research projects will be made openly available, and will advance the collective understanding of biology and evolution. EMBL will sample and analyse communities to study the effects of natural and anthropogenic environmental variables, gaining an experimentally derived mechanistic understanding at the molecular level. This will have a wide impact on a number of scientific fields including ecology, the environmental sciences, zoology, and geobiology. The resulting data will help to foster secondary research to bolster conservation efforts, protect and restore biodiversity, and consequently create new benefits for society and human welfare.

Molecular surveys of ecosystems, together with measurements of relevant environmental factors, are also likely to reveal organisms and functional traits of biotechnological relevance, such as those that can increase

carbon fixation. Some highly efficient microalgae have been extensively studied in this respect, but even more efficient groups may exist. These could be revealed using such approaches, as might entirely new functionalities that make the core process of carbon fixation more effective. EMBL's aim to understand the mechanisms by which antibiotic resistance spreads in the environment may facilitate the development of new strategies to combat the emergence of multidrug-resistant pathogens.

Molecular biomarkers may also be identified using field experiments, and validated and further investigated in microcosm and mesocosm settings. These biomarkers could provide an improved understanding of biotic and abiotic interactions within ecosystems, and could enable research to quantify the states of ecosystems. Such discoveries could also have practical applications and be used for assessments of planetary health (Chapter 12: Innovation and Translation). These practical applications would help scientists to understand the effects of pollution levels on an ecosystem and provide an early warning system for potentially irreversible ecological changes in specific at-risk ecosystems. In turn, this could inform policy initiatives across Europe and around the world. The development of diagnostic capabilities and the identification of solutions for specific problems can then be powerfully leveraged by EMBL's member states.

Through this novel scientific endeavour, EMBL in close collaboration with its member states will help to address fundamental and pressing scientific questions about the impact of the environment on biological processes, while also addressing societal questions about planetary health. Planetary biology research represents an extraordinary fusion between intrinsically fascinating biological questions and modern environmental challenges, which EMBL can and must help to overcome.

8. Data Sciences

Terms Used Throughout this Chapter

Application programming interface (API) is an interface between different parts of a computer programme to simplify the implementation and maintenance of software.

Artificial intelligence (AI) is the theory and development of computer systems able to perform tasks normally requiring human intelligence (such as visual perception or decision-making).

Bioinformatics is the science of collecting, analysing, and understanding complex biological data.

Cloud computing is the practice of using a network of remote servers hosted on the internet or in local data centres providing on-demand access to flexibly scalable resources to store, manage, and process data, rather than a local server or a personal computer.

Computational biology is the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological systems.

Containers are standard units of software that package code and its dependencies. This enables the container to be shared between different computing environments to enable quick, reliable, and standardised analysis of data.

Data management is the development of architectures, policies, practices, and procedures to manage the data life cycle.

Data science is the science of using research methods, processes, algorithms, and systems to extract knowledge and insights from (typically large) structured and unstructured data.

Data science centre is used here as an umbrella term that includes data itself, its management, its bioinformatic processing and analysis (including but not limited to statistics, AI, and imaging) by scientists, the provision of code, tool and data services to internal users and the public, and links to (potential) local/national partners.

Data stewardship is a functional role in data management and governance, with responsibility for ensuring that data policies and standards turn into practice within the steward's domain. Data stewards support an organisation to leverage its data assets to full capacity.

FAIR Data Principles are a set of guiding principles for scientific data management and stewardship which aim to make data findable, accessible, interoperable and reusable.

Graphics processing unit (GPU) is a specialised electronic circuit designed to optimise the output of images to a display device. The highly parallel structure of GPUs makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel.

High performance computing (HPC) is the practice of densely aggregating computing power, memory and storage at large scale linked via ultra-fast network links to deliver much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business. It is used to solve computational problems that either are too large for standard computers or would take too long.

Machine learning is the study of algorithms that perform a specific task without explicit instructions. These algorithms automatically improve through experience, by relying on patterns and inferences in the sample data they are given.

Theory is a branch of science that uses mathematical models and abstractions of objects and systems to rationalise, explain, and predict natural phenomena. This is in contrast to experimental science, which uses experimental tools to probe these phenomena.

Background

The life sciences are being transformed by the presence of cheap and scalable technologies that can generate immense volumes of data, coupled with computational and infrastructural advancement that have radically changed the way scientists currently perform bioinformatics. Most biological subfields have recently experienced exponential increases in data, with biodata doubling in volume every 18 months. In addition to the increased use of nucleotide sequencing, biology is at the brink of an additional data revolution involving high throughput bioimaging technologies, such as those provisioned by the new EMBL Imaging Centre. Hence, the computational sciences in biology, including bioinformatics, modelling and computational biology, have become a necessary methodology which is relevant across many biological domains.

In this context, data science - the science of using large-scale data generation techniques along with tools and systems to extract insights from large structured and unstructured datasets - will be a key driver of the future of biology. Data science is also a major driver of startups and business innovations. There is the well-known adage that “data is the new oil”, with one of the implications being the sheer size and complexity of the IT infrastructure required. Novel analysis approaches, including those based on artificial intelligence (AI) and dealing with data of increased scale and complexity, will drive many biological innovations of the future.

Data science approaches in biology depend on coherent data structures, storage, and management, so that large datasets can be combined and utilised to fully exploit their potential for research. Consequently, the ability to manage, analyse, and make growing amounts of biodata accessible is of outstanding strategic importance for EMBL’s internal activities, as well as for external services and training provided to its member states. EMBL is uniquely positioned to lead in biological data science in Europe, since it combines capabilities and service facilities for generating large volumes of high-fidelity biodata, leading research activities in molecular biology and bioinformatics, and the hosting of widely used data repositories - all in a single institution.

The Opportunity

Molecular biology experiments lead to growing amounts of raw data that need storing with the appropriate metadata. This leads to various processing and analysis steps, sometimes in the context of complex international collaborations, and finally the need to make parts, or all, of the data available to the public, ideally in dedicated repositories. Similar to life sciences institutions everywhere, the various data flow scenarios at EMBL are complex, since data can have different origins and users, for example from within EMBL or from external researchers.

At EMBL, data ranges from: (i) Raw data generated by EMBL researchers (in the order of >8 Petabytes (PB) or >8 million Gigabytes annually); (ii) Raw data produced by external users of EMBL facilities and experimental services; (iii) Raw and processed data brought to EMBL via collaborators, consortia, or citizen science projects, as well as public and controlled-access data submitted by the scientific community to be hosted at the EMBL-EBI (now ingesting >10 PB of data annually); (iv) Several layers of processed data derived from (i)-(iii).

The large-scale data generation at EMBL is supported by several core facilities and structural biology, imaging, and genomics experimental services (Chapter 10: Scientific Services).

The growing volume and heterogeneity of the data have recently resulted in a massive growth of data science activities across EMBL’s sites. These have evolved rapidly, but in a rather independent manner with different solutions often being implemented across the organisation. The need to maintain efficiency and scientific

excellence in the context of growing dataset sizes; to maximise opportunities for collaboration on large-scale biodata sets within EMBL; to foster EMBL's future engagement in cross-disciplinary large-scale research efforts, especially in the new directions in this Programme; and the need to enable optimal interactions with member state researchers and research infrastructures requires **putting biodata at the centre of our research strategy**.

As EMBL undertakes new scientific directions under the Molecular to Ecosystems Programme, a highly coordinated approach to biodata will be essential to keep EMBL at the forefront of molecular biology; to allow it to efficiently and economically deal with large-scale data; to attract and develop young talents and foster careers; to make EMBL ready to embed the most modern analytical approaches in big data and AI; and to create new bridges between disciplines. As such, EMBL can act as a role model to empower the development of biological data sciences across Europe.

Aims

Throughout the next EMBL Programme, EMBL plans to establish and grow a **data science centre** which will coordinate biodata activities across EMBL, allowing for flexible and adaptable management. The name data science centre is used as an umbrella term that includes data itself, its management and bioinformatic processing by dedicated staff, the provision of code, tools and data services to internal users and the public, and links to (potential) local or national partners, as well as the promotion of open science principles.

EMBL's data science centre will be highly coordinated, since needs at each site are heterogeneous and complex. Support staff and teams relevant to EMBL's data science centre can therefore be spread across the EMBL sites. Overall it is expected that these factors will lead to higher quality, efficient, and robust internal and external services, crucial to EMBL.

EMBL's data science centre will also leverage EMBL's data science capabilities to foster and integrate data-driven biology and data science research in Europe through mutually beneficial collaborations. It will also allow EMBL to maximise participation in internationally and nationally funded activities relevant to the biological data sciences aligned with EMBL's missions in science and infrastructures.

EMBL's data science centre may take on some or all of the following roles:

1. **Provide internal research support for biologists:** Data science research support teams will provide research support not only in classical bioinformatics domains, but also in omics analysis, statistics, machine learning, visualisation, bioimage analysis, computer-simulation based experimental design and modelling. They will additionally be associated with internal experts (group and team leaders) in the respective areas relevant to data science.
2. **Facilitate research advances in data science:** While the main role of the data science centre is to enable, there will also be an important place for performing research to develop and improve methods and tools, and for further discovery science. Thus the data science centre can have embedded group and team leaders, and attract increased capacity for innovative technological solutions via recruitment of skilled data science professionals, postdocs, and PhD students.

3. **Develop common representation of data, conventions, workflows:** The data science centre will enable researchers to capture increasing amounts of metadata, in order to maximise the utility of large-scale biodata sets. This will also enable EMBL-EBI best practice to be propagated for simpler and more coherent data management, to transfer to national and global depositories, and to make data findable, accessible, interoperable, and reusable (FAIR Data Principles; <https://www.go-fair.org/fair-principles>).
4. **Provide training and career development:** The data science centre will further deliver training to develop the careers of EMBL's researchers as well as member state researchers. EMBL is in a fortunate situation to have many in-house specialists who can deliver, in a joint effort, a wide spectrum of data science training, from the basics of data analysis, statistics and bioinformatics to advanced and newly arising topics including AI and theoretical biology.
5. **Offer critical public data resources:** EMBL can provide services such as public data resources and tools, for the growing volumes of biodata and the diversity of biological applications, to empower the scientific community to work with big datasets. EMBL will also implement and promote open biodata practices, by making visible and enabling FAIR data and tools.
6. **Act as bridges to member state activities:** The data science centre at all EMBL sites will have the ability to get involved in national infrastructures and member state activities. The data science centre will interact with national ELIXIR nodes, and engage in collaborative fundraising (public and industry funds) for external facing bioinformatics services and training beneficial to both the EMBL site and the member states.

EMBL's data science centre will be the primary arena to fulfil EMBL's commitment to **open science** principles and practices. Open science includes open data and open methods, with one aspect being the FAIR guiding principles for scientific data management and stewardship, which intend to improve the findability, accessibility, interoperability, and reusability of data. In spite of its clear advantages, open science in biology is still in its relatively early days. Currently, only positive and fully interpreted results tend to be published by biological communities. By comparison, raw data as well as protocols and software used to process these data are often not published, which affects reusability and reproducibility of science. Thus the practical application of open science principles will require both policy development and implementation planning. EMBL aims to act as a leader by using and promoting open science practices. These practices encompass the management of research software and data products, the data that streams from community use of the research facilities at EMBL, and the wider, global contribution to data management via the public data resources hosted by EMBL on behalf of the research community.

The data science centre will also implement considerations around data protection, including EMBL's implementation of the EU's General Data Protection Regulation (GDPR). Research activities with human data will significantly increase in the future, including controlled access data types. Thus, data protection and all of its implications will need more attention. Here, EMBL will continue to develop appropriate policies and processes which can facilitate robust collaborations between EMBL and member state researchers in a shared legal and regulatory framework.

EMBL's Approach

The implementation of the data science centre requires careful and coordinated planning. EMBL has identified five priority areas that will be the focus for data sciences throughout the next EMBL Programme. These priority areas will enable EMBL to maximise the discovery of research data, to effectively manage the growing volumes of biomolecular data, and to engage with the diversity of biological applications and the varied needs of a growing and global scientific community (Figure DS1).

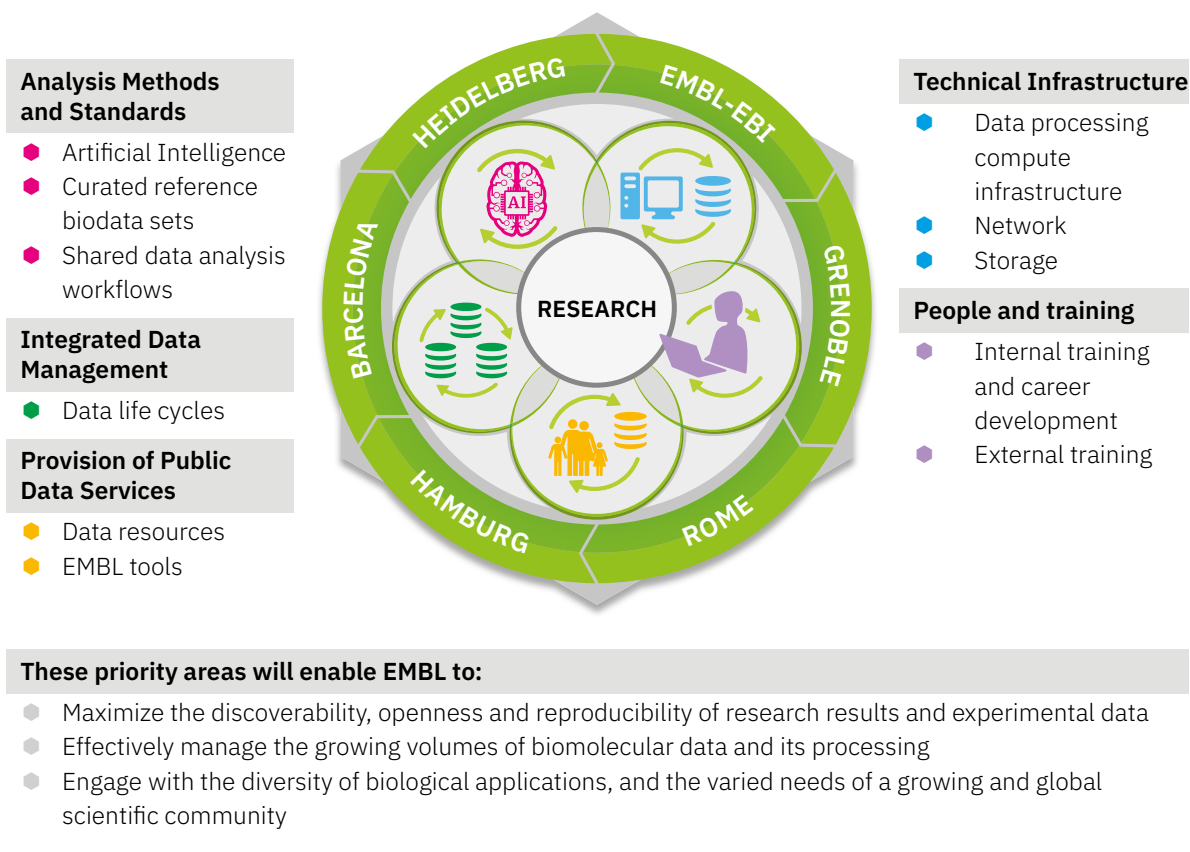


Figure DS1 | Data Sciences Programme, connecting data science at all EMBL sites.

EMBL's new Data Sciences Programme aims to connect data science focused on five priority areas at all EMBL sites. This activity will be highly connected across sites to leverage synergies across all of EMBL and to share expertise and experience while integrating into the local research and service landscape to ensure bespoke solutions where needed.

Analysis Methods and Standards

Artificial Intelligence (AI)

AI technologies will be key drivers of innovation in data science in the coming years. EMBL's research and service teams are already active in various aspects of AI, both as users and developers of novel AI approaches. This includes bioimage analysis; investigating human genetic variation in non-coding regulatory DNA; and integrating somatic mutation and pathology image data for cancer classification. AI is an area of expected future growth, since AI approaches such as deep learning have ample potential to facilitate the analysis and integration of structured and unstructured data, especially in the context of large-scale datasets. EMBL is well-positioned to become a leading player at the interface of AI research and molecular biology, where

EMBL's strengths include:

- The ability to manage and integrate large amounts of biodata data types (e.g. omics, bioimaging);
- Expertise in curating data (e.g. to label biodata with the “ground-truth”);
- Data-producing service facilities and involvement in pan-European consortium activities creating large-scale reference datasets which are highly relevant for training AI software;
- *In silico* AI research to build and leverage AI methods in the life sciences.

EMBL will continue to monitor progress in the general machine learning and AI research communities and adapt relevant approaches to biological data or develop new ones where appropriate. EMBL's aims in this are threefold: **driving biological discovery science, contributing to machine learning/AI research, and improving EMBL's services**. One example of current research at EMBL is an approach to deep transfer learning connecting histopathology and patterns of mutation and gene expression in cancer (Chapter 6: Human Ecosystems), which illustrates EMBL's engagement in leading-edge AI developments and applications in a research setting.

As part of the new EMBL programme, EMBL will engage in AI research in several areas particularly relevant to biology, including:

- **Un-/semi-/self-supervised learning to tackle lack of training data.** This aspect will be especially relevant to cell type classification and disease classification problems, where ground truth data are often lacking.
- **Explainability and uncertainty quantification.** AI applications in the life sciences can be empowered by obtaining insights into the relevant AI models. Deriving causal and/or mechanistic insights, and quantifying model uncertainty, will thus represent a focus area.
- **Interpretable low-dimensional representations and metric learning.** Complex multimodal high-dimensional data (such as single-cell multi omics data) require methods for dimensionality reduction and analysis of interrelations, which will represent another focus of EMBL's AI research.
- **Privacy-aware and federated learning.** In the context of human-derived data, federated algorithms and infrastructures that can preserve privacy while allowing aggregate information across different compute centers and jurisdictions will be developed.

The conceptual underpinnings of AI methodology in the data sciences theme will be tightly linked to the new theory theme, which addresses questions such as model conceptualisation, interpretation, and critique (Chapter 9: Theory at EMBL).

To advance machine learning in the life sciences, EMBL will focus on **selected areas** where EMBL is generating data or managing unique data resources. Strategically, EMBL sees great potential added value of AI to biological research and services in the annotation and interpretation of sequences (genomes, transcripts, and proteins) and sequence variants, bioimages (segmentation, navigation, connectomics, and phenotyping), and biomolecular structures. To achieve these aims, the service aspects around data science will facilitate the annotation, navigation, cross-modal integration, and interpretation of huge amounts of structural, bio-image, and omics data using modern algorithms and AI. This includes the ability to train and compute directly on archived data, such as the BioImage Archive recently established by EMBL (Chapter 10: Scientific Services). It also includes the need to embed advanced AI algorithms directly into research and service infrastructures that generate these data, thus facilitating their use by EMBL researchers and users of EMBL's external services.

Key requirements for assessing progress in AI are benchmark datasets and open challenges, where tools are objectively compared on representative datasets. Critically, there is a scarcity of such reference datasets for major biodata domains, including microscopy, where few reference datasets exist. EMBL researchers will take a leading role in **establishing, collecting, and disseminating curated reference biodata sets** and problems, for which AI-based methods are particularly suitable. EMBL researchers will also develop baseline solutions and the required infrastructure components, which will allow biologists and AI researchers to synergise and address open questions in molecular biology.

To bolster AI efforts, EMBL aims to engage with **AI research communities** at multiple levels to foster developments inside and outside of EMBL. AI and machine learning research at EMBL will be connected to European networks and AI initiatives, including the European Lab for Learning & Intelligent Systems (ELLIS; <https://ellis.eu>). In this context, EMBL's successful proposal to create ELLIS Life, a cross-institutional ELLIS unit between EMBL, the German Cancer Research Center (DKFZ), and the University of Heidelberg, will foster this interaction.

A key advantage of AI methods is that they can be reused for a wide range of recurring tasks, where biologists only need to provide input or training data to parametrise an existing algorithm. Alongside EMBL's role within ELLIS, EMBL has leadership roles in several AI communities which distribute pre-trained models for such recurring tasks via repositories of code and parameters, so-called model zoos, such as the model zoo for genomics (<https://kipoi.org>) and the model zoo for bioimaging models which is expected to be launched in late 2020. While these repositories are important in the core AI domain to ensure reproducibility and full exploitation of available training data, they become even more crucial in domains where the models are deployed by non-experts (e.g. wet lab biologists) as they provide a level of trust in the choice of algorithms.

EMBL also aims to strategically **hire more AI talent** to further strengthen research at the interface of AI and molecular biology. This includes recruiting researchers and service staff already trained in the mathematical, statistical, and computer science foundations of machine learning, and training them in biology as well as training biologists in the advanced use and adaptation of machine learning/AI algorithms. It is foreseen that innovations in AI are likely to strengthen and help drive EMBL's future biological discovery research, as well as innovations enabling technology transfer, and service provision. Strengthening AI research in Europe could counteract the brain drain of AI talent to North America and ultimately raise the competitiveness in "biodata AI" in EMBL's member states. The tools and resources developed by EMBL, reference datasets archived at the EMBL-EBI, and cloud resources with access to graphics processing units (GPUs) and data, will facilitate wider adoption of AI in biology in Europe.

Shared Data Analysis Workflows

The growing utility of data across all biological domains implies a need for more efficient data science practices and economic use of advanced computing capabilities. EMBL's past success has partly been grounded in its fast-moving, "bottom-up", decentralised planning of research studies. However, this brings with it the risk of duplication and redundancy in the development of data analytic workflows. Large-scale consortia studies such as Pan-Cancer Analysis of Whole Genomes (PCAWG) or the Human Cell Atlas (HCA) have highlighted key advantages of common (shared) computational workflows for recurring tasks (Figure DS2). For example, shared workflows can save productive time, increase reproducibility and scalability, and thus can free resources to tackle new questions.

To support EMBL's research including the new research themes, EMBL aims to extend the use of such 'standardised' workflows for recurrent activities, such as what is commonplace at the Broad Institute in the United States of America. This will be first achieved within EMBL, to subsequently enable and empower

external research communities across EMBL's member states. These harmonisation efforts will embrace the relevant community standards like those developed by the Global Alliance for Genomics and Health (GA4GH) or the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI) to maximise international comparability and reproducibility.

Software packaging, deployment using classical high performing computing (HPCs) and clouds, and life cycle management services (e.g. Bioconductor) and containerisation (e.g. Docker, Singularity) will facilitate workflow dissemination. The development of such workflows will also provide additional opportunities for EMBL to lead in coordinating the development of standards within relevant biological communities.



Figure DS2 | Shareable workflows to enable standardised analysis of protected data.

Shareable analysis tools (containers) can be sent to locations where large data sets reside, such as a particular EMBL site, or, in the case of human data research protected data hubs, as previously achieved in the Pan Cancer Project led by EMBL (www.nature.com/collections/PCAWG), to enable the standardised analysis of biodata irrespective of where the data are stored.

Integrated Research Data Management

EMBL's approach to coordinated data sciences aims for seamless data flow throughout the whole life cycle of data, from sample to production of the raw data to data analysis and finally to open access publication of the findings and deposition of the data in the appropriate databases (Figure DS3). This will not only lead to all EMBL data being made FAIR, but higher-quality data will be more efficiently produced and analysed, as well as allow for EMBL research outputs to be more rapidly made available for scientists in the member states and beyond.

Data life cycles are often complex involving individual and highly customised workflows according to the needs of the specific discipline. But general principles are shared across disciplines, including workflow stages such as data processing and analysis, quality assurance, integration with other data, dealing with sensitive data where access needs to be controlled, sharing within or across international consortia, data archiving, and publishing. Currently, data intensive research projects typically involve a lot of data movement, and thus, copying and duplication across IT infrastructures, different data storage devices, clouds, and data archiving systems can be made much more efficient.

Implementation of optimal data management policies, processes and tools will require expertise in data management planning and data stewardship to integrate new efforts with those that already exist at EMBL such as the data management applications available for the microscopy core facilities. Researchers, facilities managers, IT staff, and public data resource managers will all be involved in implementing a software framework spanning from wet lab groups all the way to the open databases at EMBL-EBI. A new internal data deposition brokering network, helping the process from locally generated data towards public deposition and representation, will be hosted within the data science centre. Dedicated data scientists will also support wet lab researchers through providing services in data analysis, integration, and FAIR data management and sharing. These data scientists will facilitate good practice in scientific computing and enable community building across EMBL sites.

EMBL's data management tools will in the future:

- Permanently link individual datasets through stable identifiers such as DOIs or accession numbers (Figure DS3, dashed blue arrow), allowing data to be consistently referred to within electronic lab notebooks and, at the same, avoiding data duplication;
- Aid collaborative efforts across EMBL by giving a comprehensive overview of where data are stored and organised according to a project, and on their current state of processing (eg. quality assurance, analysis, publication, and archiving);
- Manage access to and open the sharing of data to facilitate data integration across data sources, and to maximise the quantity, quality, and reusability of research outputs;
- Support key functions needed for physical data movement including long-term archiving with the inclusion of relevant metadata;
- Help EMBL's data-generating core and service facilities to adapt their IT workflows, increasing their efficiency and quality.

Pilot implementation initiatives are in progress and focus on multi-omics data, X-ray diffraction data (Figure DS3), and imaging data. These pilot initiatives will help find answers to big data questions and establish use cases for shared use of IT resources and data analysis and management across EMBL. Member state users of these EMBL facilities and services will likewise benefit from enhanced data management, data representation, and data flow capabilities.

The planned initiatives around research data management could have a significant impact beyond EMBL. Projects like Euro-BioImaging and Instruct-ERIC (Chapter 13: Integrating European Life Science) may be starting points for scaling beyond EMBL as the problem is common to all research institutions and particularly those supporting life sciences facilities and services. Efforts could also be linked to the ELIXIR-led EOSC-Life initiative aiming to provide generic services accessible across Europe to support open science within the context of the EOSC (Chapter 13: Integrating European Life Science). Through these activities EMBL could help develop open science best practice that will be at the disposal of the wider scientific community to adopt and use, or directly provide data science services and infrastructures to member states.

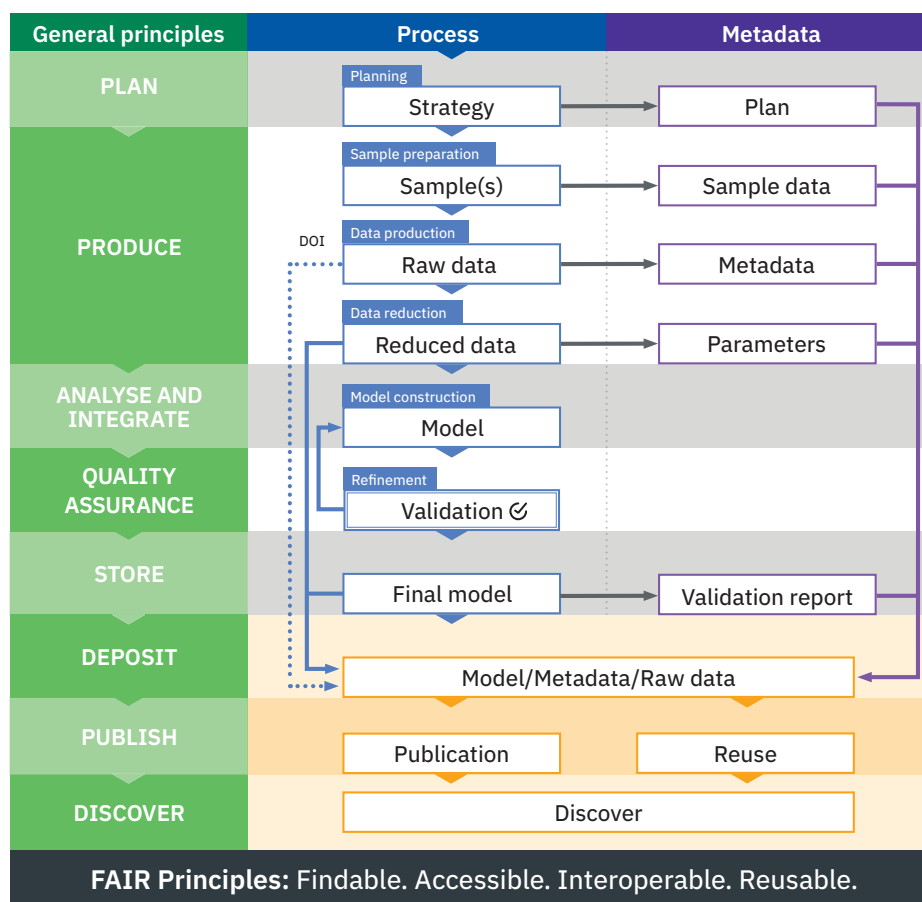


Figure DS3 | EMBL data management workflow.

Supported by FAIR principles, the general principles of the EMBL data management workflow guide the flow of data and metadata for bespoke research examples.

Provision of Public Data Services

Molecular Data Resources

Through EMBL-EBI, EMBL has a community mandate to develop and maintain a comprehensive array of open access and up-to-date **molecular data resources** as a key part of its services mission (Chapter 10: Scientific Services). These data resources cover a huge range of molecular biology subdisciplines including nucleotide sequence data, bioimages, protein sequences and families, chemical biology, structural biology, systems, pathways, ontologies, and scientific literature. In short, EMBL-EBI's data resources collate, integrate, curate, and make the world's scientific biodata freely available to all.

The EMBL-EBI data resources include deposition databases (archives) that store primary experimental data generated by scientists worldwide, as well as knowledge bases that integrate and add value to experimental data, making it easier to use and understand, with many having both functions. Through EMBL's engagement with the research community - whether that be the development of data resources to meet continuously changing research needs or via training and industry engagement programmes - EMBL strongly supports and propagates the concept of FAIR biodata.

EMBL-EBI data services are managed as an institutional commitment on behalf of the research community, rather than as the sole responsibility of individual group or team leaders. These open and searchable resources provide all researchers with direct access to the scientific record, enable access to and reuse

of experimental data to verify original results, and, by combining multiple data records, provide analytical insights. Deposition databases also provide reference data for the research community. Through the use of search tools, researchers can rapidly compare their own unpublished data with open access datasets.

EMBL-EBI monitors the usage of resources (for example, web hits, data downloads, citations), which feeds into processes for managing the data resources from start-up, to full operation, to, in some cases, retirement. The management of these resources is driven by scientific need. Therefore, in the same way that new resources may be launched to meet emerging needs (such as the BioImage Archive), others may be downgraded as needs for particular technologies or data types change. The data stewardship expertise at the institutional level at EMBL-EBI is of fundamental importance in managing these processes.

Research Software

Research software produced at EMBL is of two main types: (i) data analysis workflows or “scripts” associated with a specific dataset, scientific question and, often, publication; and (ii) more generic methods, algorithms, or tools intended to be useful across a range of applications.

Many computational tools developed by EMBL researchers are highly used by the scientific community. Some of the highly used tools in 2019 were InterProScan (accessed 55 million times to annotate sequences), DESeq2 (downloaded 286,000 times for differential gene expression analysis), Ensembl Variant Effect Predictor (used 177,000 times to analyse over 180 million variants), ilastik (downloaded 60,000 times for image classification and segmentation), webPRANK (a phylogeny aware multiple-sequence aligner used 10,000 times), and limix (downloaded 8,000 times for genetic analysis of multiple traits).

Currently, research tools developed by EMBL researchers are made publicly available according to the needs of individual groups. This is dependent on the availability of funding, resources, and time, rather than being coordinated and managed institutionally. The majority of public data resources and tools are currently at EMBL-EBI, and a growing number of such data services have recently emerged from other EMBL sites. In addition, the large-scale provision of services to produce bioimage data by EMBL researchers and external users within the new EMBL Imaging Centre will create immediate needs in image data analysis in the user community from member states. EMBL aims to extend its expertise in bioimage data analysis to develop new image data analysis methods as well as provide these tools in a user-friendly manner to the internal and external user community via robust services.

In the future, EMBL will strategically align the provision and life cycle management of external facing data resources and tools declared to be a service at all EMBL sites. EMBL will develop a more cohesive framework for deciding which research data and tools should be supported institutionally and how those resources should be developed and supported both in terms of skills and expertise and technical infrastructure.

The institutional approach to public data services provision aims to:

- Develop processes and support to identify, develop and maintain computational research tools and data resources that can be declared to be of enough strategic value or impact to be supported by EMBL as a service;
- Have clear rules for open vs. closed source, defined in view of the competition of algorithms and commercialisation of software, with a strong preference towards open software;
- Train people involved in services in software engineering to achieve robust, production-level tools and software life cycle management;

- Ensure that the tools and data resources deemed to be of value as a service will from then on share a common brand or certification, which EMBL expects to result in the increased visibility of tools or data resources, both within and outside of EMBL;
- Ensure that tools and resources have a path to be officially retired, for example if the relevant technology becomes obsolete.

As EMBL embarks on new scientific directions, it is envisioned that EMBL can pioneer the provision of public data services (both data resources and tools) for these new initiatives, as well as provide data platforms for new communities of users such as ecologists, marine biologists, clinicians, and epidemiologists. This can be similar to how EMBL currently provides bioinformatics resources to clinicians and pharmaceutical researchers for genomic medicine diagnosis and drug discovery.

Technical Infrastructure

Compute requirements, including for classical high performance, high throughput and cloud computing will grow with EMBL's needs for data generation and analysis. These call for broader sharing of access to IT resources, including improved distributed access to HPC, cloud services, and long-term data storage at all EMBL sites.

Cloud technology has the vast potential to simplify connecting and federating distributed IT infrastructures, as well as enable cross-site research studies, where data may not always be able to move between EMBL sites at the generation and analysis stages (Figure DS2). Currently EMBL-EBI operates the Embassy Cloud which, when using OpenStack, enables fast and scalable provisioning of virtual machines in the manner of an infrastructure-as-a-service (IaaS) cloud solution. Thus, this allows for dissemination of tools and computation on big data for local scientists and remote collaborators alike. Similar benefits from reuse of pre-packaged or “containerised” workflows and tools are provided by the container hosting platforms operated both at EMBL-EBI and EMBL Heidelberg. The concept of an EMBL Science Cloud has recently been launched to promote the use of clouds, initially building on existing infrastructures such as the EMBL-EBI Embassy Cloud and the 3D Cloud in Heidelberg (Chapter 14: People, Processes, and Places).

For the next EMBL Programme, EMBL will increase the accessibility of cloud resources (e.g. OpenStack-based) and the use of cloud application programming interface (APIs), augmented by software for orchestration (e.g. Kubernetes), to create a distributed computing environment enabling data processing with similar tools, irrespective of where the data are. This is driven by the fact that more and more data cannot easily flow to, from, or between EMBL sites owing to their size or legal or regulatory constraints. This could include datasets generated from EMBL's new directions such as Planetary Biology and Human Ecosystems. EMBL internal clouds also need to be able to interact with commercial clouds, in a hybrid cloud model, wherever this is a better solution, whether for financial, legal, or regulatory reasons. The increased usage of AI at EMBL will also result in future GPU needs that EMBL aims to satisfy.

With regard to data storage, all stages in the life cycle of a particular dataset will be appropriately tracked as described under **Integrated Research Data Management**, including the context of the experiment or project, the physical location, copies, and ownership. **Modern data storage solutions**, such as object storage to complement file system storage, will be explored to support this in a cost-efficient manner. The FIRE service established at Hinxton provides a scalable resilient data storage infrastructure that can ingest over 1PB of data a month and has scaled already to over 30 petabytes.

People and Training

Data science is a multi-disciplinary area which includes bioinformaticians, statisticians, mathematicians, and theoreticians. Personnel involved in EMBL's data science activities are based at all of EMBL sites and can fall into a number of different categories (Figure DS4).

In silico researchers in purely dry groups make up roughly 18% of all research groups across EMBL and are located at all EMBL sites, with the majority residing at EMBL-EBI and EMBL Heidelberg. Research staff undertaking data sciences in predominantly wet labs and core facilities (33%) can range from dedicated bioinformaticians with bioinformatics and/or data analysis expertise in a specific domain to experimental scientists working at the wet-dry boundary. A growing number of wet lab scientists are acquiring data science skills to become “hybrid” (wet-dry) researchers to analyse individually generated large-scale data. A few dedicated individuals (~10) provide primarily internal bioinformatics and data analysis core services, as well as specialist expertise and advice for EMBL wet lab scientists. They also provide some internal bioinformatics and data science training, often supported by data scientists within the research groups. Staff responsible for EMBL's public bioinformatics service activities make up roughly 400 of the 1,800 EMBL personnel. Although now mostly at EMBL-EBI, as the number of outward facing services increases at the other EMBL sites, these capacities will expand during the new Programme.

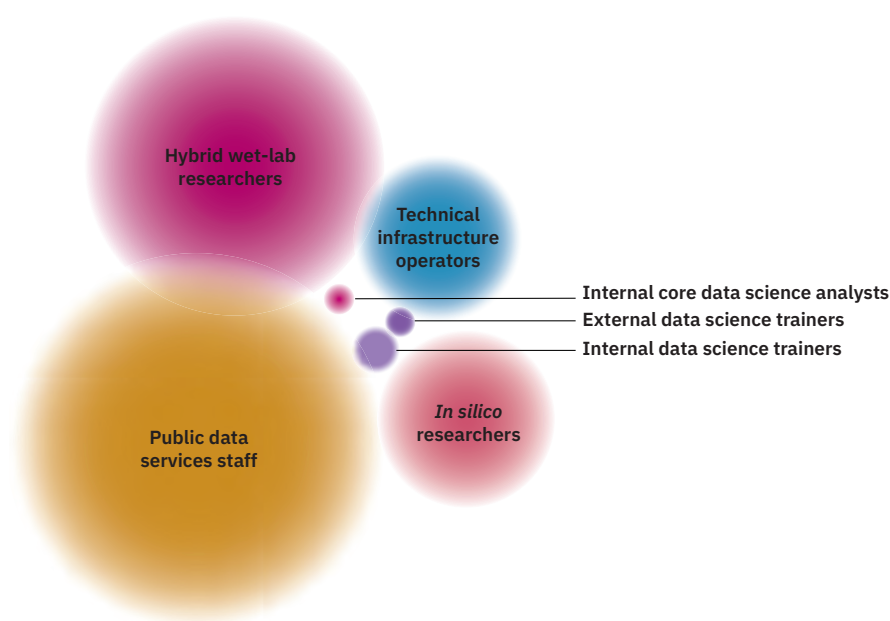


Figure DS4 | Proportional representation of the number of data scientists by their main activity.

Data scientists at EMBL can fall into multiple categories spanning *in silico* researchers, a growing proportion of hybrid researchers, provision of internal and external bioinformatics training, as well as the provision of infrastructure and technical services.

The data science centre needs to ensure that staff in all of the categories listed are adequately and centrally supported to carry out their respective functions in this competitive and fast-evolving area. As such, EMBL needs to ensure it is positioned to identify and attract the **best talent** in strategically important research areas; ensure career data scientists are **recognised** for their contributions; provide mechanisms for **networking** with bioinformatics peers across EMBL and externally; and ensure appropriate **workload management**, especially for those who provide core and public data and analysis services.

The need for central support for **internal training and career development** (Chapter 14: People, Processes, and Places; Career Development) for the new generation of data scientists at the frontier of wet and dry biology is expected to grow. Future training will enable EMBL staff at all levels to stay current in their knowledge and use of data science methods and tools (including for new approaches such as AI) and prepare them for their future careers. Access to data management training will allow EMBL researchers to efficiently use data management resources on offer as well as comply with data management strategies and policies.

As the data science centre develops, EMBL anticipates that staff will become **key providers of external training**. EMBL-EBI's Train Online platform, which already incorporates some courses on basic data management principles, can be used to support advanced, face-to-face external training at EMBL and in the member states (Chapter 11: Training). Ongoing participation in ELIXIR's training platform plus the creation of pop-up training rooms, most likely using commercial cloud-based compute, will be developed to cost-effectively manage the 'bursty' need for compute to support training in and beyond the member states. Empowering scientists to construct their own solutions will be one pillar of these new directions, which will ensure that the very best of EMBL data science is shared with the member states.

Impact

EMBL envisions that a new EMBL-wide approach to data sciences could have significant impact externally including in EMBL's member states and the global scientific community, as the problem of growing datasets is common to all research institutions in the life sciences. Specifically, the EMBL-wide approach to data sciences will lead to:

Research support – Provide enhanced research support for member state users of EMBL's core and service facilities in managing, analysing, sharing, and archiving their data. Member state researchers will not only benefit from a more streamlined data flow, but will also be able to reuse shared standardised data processing workflows developed at EMBL to accelerate their research and improve the reproducibility of their analyses.

Training and career advancement – Help develop the careers of wet lab biologists and data scientists within Europe's increasingly interdisciplinary life sciences research landscape. In addition to offering external courses at the six EMBL sites, EMBL will work towards developing a network of partners in other member states to deliver data science training locally.

Promoting research collaborations with member states – Provision EMBL's enabling data science resources and tools for data-intensive collaborations involving member state scientists, such as the European Open Science Cloud (EOSC), the Human Cell Atlas (HCA), as well as the initiative 1+ Million Genomes which works towards access to at least 1 million genomes in the EU by 2022.

Formation and strengthening of local hubs with national ELIXIR nodes – Engagement of individual EMBL sites with member states.

Support in open science – Support member state scientists in producing FAIR data, to facilitate their adherence to grant requirements and (inter)national standards.

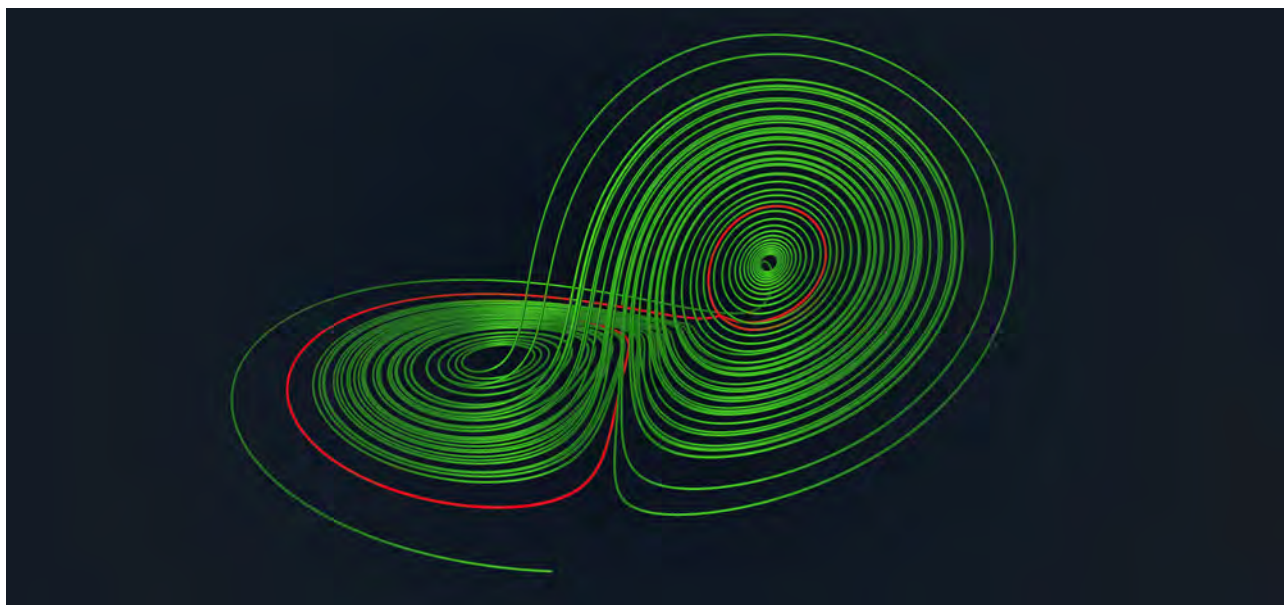
Through these efforts EMBL could serve as a role model for life sciences institutions that need to cope with growing volumes of data, or even directly provide infrastructures to member states. This would particularly empower those member state users that in their home institutions lack the facilities and infrastructures to deal with storing and analysing large-scale datasets.

9. Theory at EMBL

Background

As EMBL moves into exploring life in context, there is now a great opportunity and need for a revitalised inclusion of theoretical approaches in the life sciences. Interest in the use of theory in biology has been repeatedly expressed in the past. However, limitations of the experimental possibilities to query biological systems have restricted the use - and impact - of theory to selected, simplified systems. This situation has fundamentally changed in recent years. Researchers now have unprecedented capabilities to study living systems in their holistic context.

It has become possible to perturb living systems in more dynamic and versatile ways than ever before and to extract precise quantitative data using a host of different assays from omics to imaging. Life scientists are now able to experimentally address the dynamics and complexity of living matter across many scales, at high-resolution, and at systems-wide coverage. The multiple scales of living systems (e.g. molecules, organelles, cells, tissues, organs, organisms, ecosystems) are a fundamental aspect of their nature. They are full of emergent properties: properties that only manifest themselves at the integrated systems-level view and that are not directly evident from the properties of the system's components. Because of that, the working principles of living systems can often not be understood through purely intuitive approaches, but deeper understanding can be gained through formal mathematical reasoning and modelling. A theory- and modelling-guided approach is thus needed and timely, and can now be integrated with improved mathematical techniques, better computational power, and new experimental possibilities in order to **reveal the essential features and general principles of living systems in their natural environment**.



The Opportunity

EMBL aims to build a new conceptual theory programme which will complement the data-driven computational approaches that are already, and will remain, one of the strengths of EMBL (Chapter 8: Data Sciences). Modern biology, with its recently acquired capabilities to produce quantitative data across all scales and to generate unprecedented, dynamic perturbations, is now well-positioned to fully benefit from and partner with

theory-guided approaches. Conversely, these developments enable for the first time concrete theoretical advances towards treating the unique complex properties of living systems, with major implications across scientific disciplines. The interplay between theory and experiments aims to uncover general principles in living systems on Earth. An integrated approach builds on EMBL's previous successes of identifying molecular players and interactions and will be essential to **turn data into understanding**.

In one previous example of the successful use of theory at EMBL, researchers in cell biology used a theoretical biophysical model to study the properties determining self-organisation of microtubule and motor proteins. The researchers were able to experimentally test and verify model predictions because of the development of a very simplified and controllable *in vitro* system. Today, the limitations of using theory in only a small number of experimental settings have fundamentally changed because of breakthroughs in experimental and data-generating capacities. Theoretical approaches can now be applied to more problems and in models of higher complexity. For instance, following the successful path of using theory in conjunction with experiments in neuroscience research, researchers at EMBL are building theoretical models of the retina, not only to recapitulate its visual response patterns but also to derive predictions on the neuronal circuit structure and function. These predictions are then experimentally tested to better understand the neuronal principles and mechanisms of visual information processing (Figure TH1).

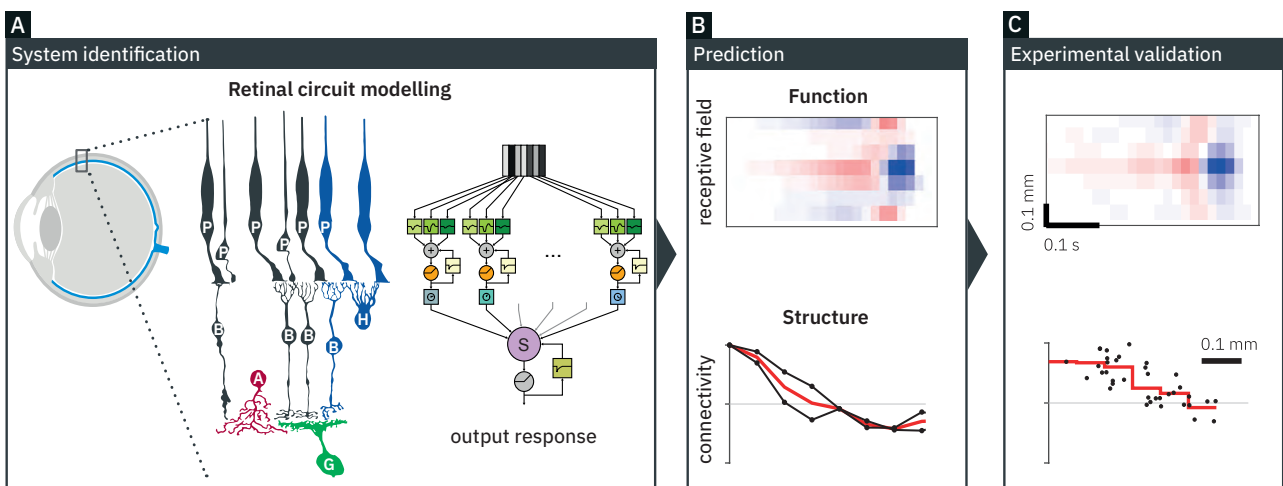


Figure TH1 | Theory-driven approach for neural circuit characterisation, followed by experimental validation.

Neuroscience research faces a need to link big data on brain anatomy and physiology, as high-throughput measurements of these become increasingly feasible. Neural circuit models (A) (e.g. of the retina) can provide predictions of such links (B), which can be subsequently tested by experiments for validation (C).

EMBL researchers also use a theory-driven approach to study complex systems; a theoretical framework is used to investigate the origin and function of oscillatory gene expression dynamics in embryonic cell ensembles. Building on concepts from **synchronisation theory**, an entrainment strategy was established and has successfully provided, for the first time, predictive and precise control over gene expression rhythms and oscillation in mouse embryos. This theory-guided approach provides the researchers with an abstracted phase-oscillator model representation of a network that is in reality much more complicated. It offers entirely new insights and reveals a functional role of oscillation rhythm in pattern formation (Figure TH2).

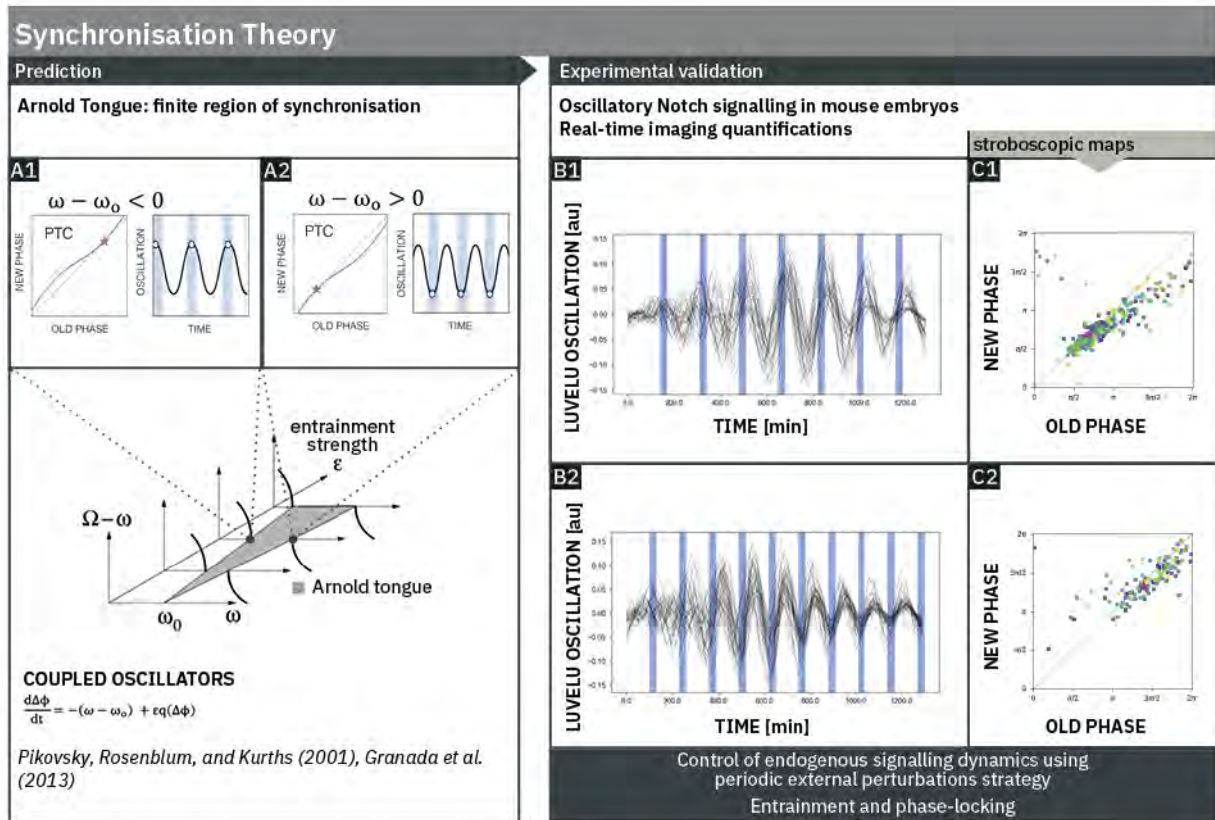


Figure TH2 | Synchronisation theory guides experimental strategy to predict and control segmentation clock oscillations in mouse embryos.

Theory (Arnold tongues) makes universal predictions about the dynamic adjustment of coupled oscillators during entrainment (**A**). In particular, the mismatch in frequencies ($\omega - \omega_0$) directly impacts on the phase-locking after synchronisation i.e. compare **A1** with **A2**. Experimental validation/reveals that indeed mouse embryonic oscillators can be controlled (entrained) to different frequencies/periods (**B**, **C**: **B2** = 170 min, **B2** = 130 min). In **B**, individual time series correspond to real-time quantifications of Notch-signalling oscillations in mouse embryonic cells. Importantly, as predicted by theory, phase-locking changes as a function of frequency mismatch (compare stroboscopic map in **C1** vs. **C2**).

Research Aims

The goal of EMBL's new conceptual theory theme is to build up first-principle approaches to gain a novel understanding into the complexity of living systems. It will build on EMBL's strong culture of interdisciplinarity and collaboration, and aims to bring together **the entire community of experimentalists and theoreticians**. This linking aspect is key to reach the goals and to assemble a research programme that integrates theoretical approaches established in other fields, such as physics, mathematics, and information theory, to solve challenging biological problems which, in turn, are expected to inspire new theory.

Research in the theory theme aims to explain biological phenomena using mathematical formalism and models and will cover the whole spectrum of EMBL's research. Theory is **more than collections of models**. It aims to uncover the **essential and general principles that transcend details of a particular system**. Theory aims to **make experimentally testable predictions** and can hence guide, and not just follow, experimentation. Through iterative cycles of theory and experimentation, the goal is to establish a theory-guided pattern to discovery that lifts the quality and efficiency of the scientific process in life sciences at EMBL and beyond.

EMBL aims to promote theoretical research in the life sciences, and as a pan-European organisation it is ideally placed to lead this effort. The goal is to ensure theory is embedded in active, vibrant experimental research and that both approaches are complementary and relevant to each other. One mechanism to achieve this is via the EMBL Interdisciplinary Postdocs (EIPOD) programme which fosters collaborations between interdisciplinary groups at EMBL (Figure TR2 in Chapter 11: Training). For theoreticians from outside the life sciences community, EMBL is an attractive place to immerse themselves and benefit from its wealth of both experimental and computational biology. As an intergovernmental organisation, EMBL is well-positioned to serve as a European exchange node, helping to promote the landscape of ongoing theoretical research in biology across member state institutions, as well as strengthening and growing the field as a whole. Finally, EMBL's wide-ranging and ambitious future research vision outlined in the Molecules to Ecosystems Programme is suitably connected and interlinked by theory and modelling. A theoretical research programme is of strong strategic importance to all aspects of EMBL research, including planetary biology, human ecosystems, infection biology, microbial ecosystems, cellular and multicellular dynamics of life, and molecular building blocks in context, with several examples highlighted below.

EMBL's Approach

EMBL aims to establish a new, dedicated theory theme that will be tightly embedded within the Molecules to Ecosystems Programme. Below are a number of examples which highlight how theoretical and experimental approaches will be integrated to tackle the fundamental questions outlined in this EMBL Programme.

Planetary Biology: The aim of Planetary Biology research is to study how organisms - and populations - respond to and integrate environmental cues and function in ecosystems. Addressing these complex questions requires an integrated theoretical and experimental approach, as highlighted also by the examples described above. For instance, the development of a new theoretical, conceptual framework was instrumental to move forward the biodiversity in ecosystem functioning (BEF) field and allowed it to extract mechanisms from experimental data, revealing a key role for species-complementarity in ecosystem function. With the advancement in the acquisition - and complexity - of data across spatial and temporal scales, as laid out by the Planetary Biology theme, there is an even greater need to complement experimental approaches with advanced **data integration and the development of new theoretical frameworks**, to gain an understanding of underlying mechanisms. **Multiscale modeling** will also be essential to guide **ecological engineering** efforts to develop ecosystems in controlled laboratory conditions, such as mesocosms, and to reveal principles and critical motifs underlying terraformation and, more generally, ecosystem function (Chapter 7: Planetary Biology).

Human Ecosystems: The central question of this new research area at EMBL is how environmental factors can precipitate disease and, more generally, how genotype and environment influence human phenotypes, and how living systems maintain, or fail to maintain, homeostasis. Such questions are being addressed by, for example, **control theory and systems engineering**, which have been developed for human-made devices to make sure systems work robustly at desired operating points. Despite the far-reaching implications that control theory could have for understanding human ecosystems, work in this area is still scarce. Moreover, the study of living systems holds fundamental advances for the field of control theory itself. Theoretical advances in control theory open novel experimental and data-analytical scenarios that would provide deep insights into the universal principles through which homeostasis is achieved, or lost. Importantly, a better understanding of these principles has the potential to open up entirely new strategies, guided by control theory, to steer living systems away from a disease state back to a stable, healthy equilibrium. The theory of **critical fluctuations** and phase transitions might also provide insight into underlying principles for observations in different diseases where large fluctuations precede the onset of the disease, with possible uses for early detection and prevention.

Infection Biology: The potential impact and importance of theoretical models became apparent during the SARS-CoV-2 pandemic of 2020, when decision makers around the world, who had to make some of the most far-reaching political and economic decisions for generations, looked to mathematical models for guidance. These models are highly complex, depend on large numbers of parameters, and integrate many diverse fields of expertise. In this area, theory researchers will be able to contribute components and building blocks for various models e.g. on pathogen diversity and evolution, transmission mechanisms, and the spread of antimicrobial resistance. There is also a great need for better quantitative and predictive understanding of the immune system, including at the individual level, using data from modern single-cell omics assays. Such models could, for instance, help explain, and even predict, the high variance of disease severity after a virus infection.

Microbial Ecosystems: Theory and modelling build our understanding of microbial ecosystems at multiple levels. First, these approaches can help tackle the vast genetic variation of individual microbial species by combining current knowledge on gene function and effects of sequence variation to predict strain traits, such as antibiotic resistance and virulence, based on genome sequence information. Second, coarse-grained molecular or ecological models and/or organism-scale metabolic models can be used to understand species interactions: chemical warfare, co-dependencies, and metabolic cross-feeding. This will open the path for the design of artificial communities, the understanding of emerging behaviours, and the rational modulation of complex real communities, like the gut microbiota of an animal, to acquire specific traits or members.

Cellular and Multicellular Dynamics of Life: Life is multi-scale and dynamic in nature; it emerges from the interplay of countless molecular building blocks and their dynamic interactions, integrating biochemical, physical, metabolic, and environmental cues, and bridging many temporal and spatial scales. As such, living systems differ fundamentally from the problems that classical theoretical fields have focused on previously. The treatment of out-of-equilibrium systems - a defining property of biological systems - is a central challenge of modern theoretical physics. Researchers at EMBL and beyond now have unprecedented experimental possibilities to obtain quantitative, dynamic, large-scale datasets across scales, for instance in developing embryos. Theory is central to extract generalisable meaning from such data (Figure MD4 in Chapter 3: Cellular and Multicellular Dynamics). For instance, **model reduction methods** need to be further developed to identify effective degrees of freedom, state variables, and the important parameters. The goal here is to extract the essence of complex systems in simplified, yet meaningful and predictive models.

Morphogenesis of multicellular systems poses challenges due to pattern and organ formation in space and time. The revolution in microscopy imaging technologies provides an unprecedented wealth of information in four dimensions, posing new challenges to carry out meaningful comparisons and quantify spatially- and temporally-structured information. **Geometrical modelling** studies the formal description of shapes and can provide novel strategies to describe morphological changes across scales and experimental microscopy modalities (Figure TH3). Thanks to theoretical abstraction, the exploration of morphological differences and similarities is reformulated as characterising the geometry of the manifold of shape. It becomes possible to model and quantify morphogenetic processes from different sources of image data and gain insight into mechanisms that allow living systems to develop and adjust their shape over time and under changing conditions.

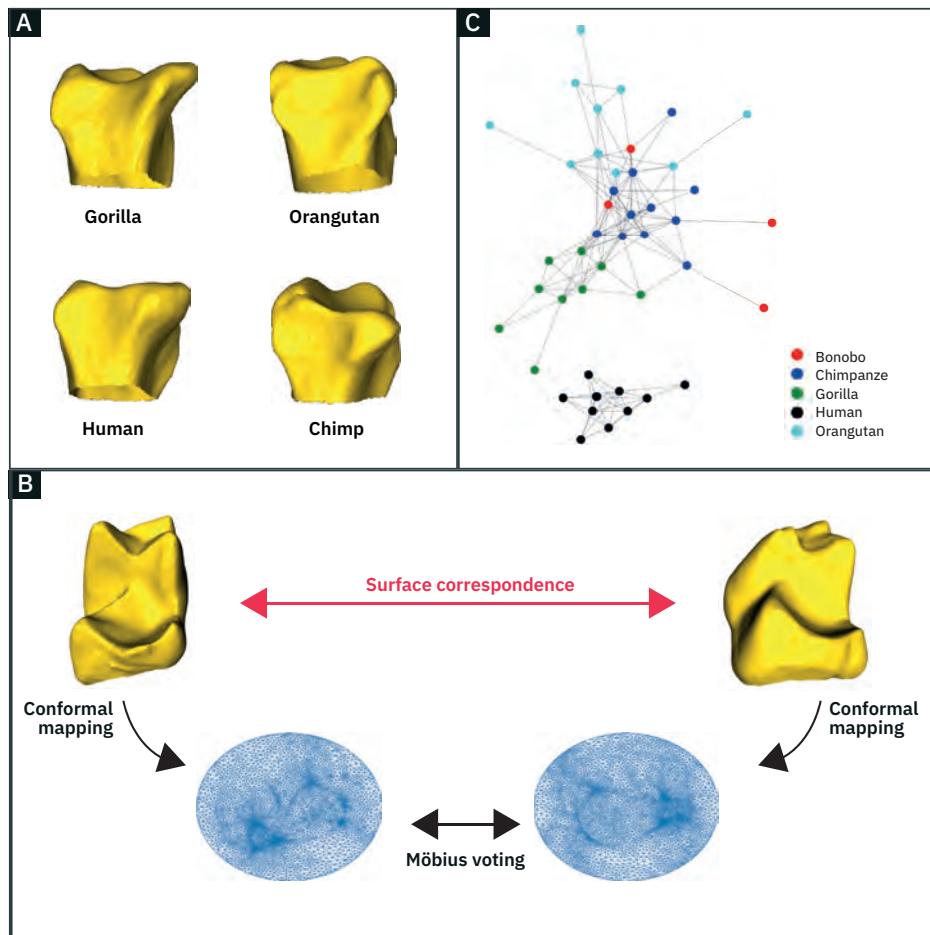


Figure TH3 | Geometrical modelling.

This type of theory is used to quantitatively explore morphological differences, here in tooth development across different ape species and humans. Meshes are extracted from 3D images of samples from each considered species (A). Correspondence is established between meshes with diffeomorphic mapping (B). Inter- and intra-species variations in tooth morphology can be investigated and modelled from these maps (C).

Another important characteristic of all biological systems is time. The rich variety of frameworks for dynamical systems in physics and applied mathematics enables researchers to conceptualise biological phenomena such as cellular decision-making and development in terms of symmetry and symmetry-breaking, bifurcations, effective dimensions, manifolds, attractors, limit cycles, stability, determinism versus stochasticity, perturbations, or self-organisation. Excitingly, experimental approaches are now available to link to **dynamical systems theory**. These theories can be instrumental, when combined with experimentation, in revealing the fundamental and universal principles that guide dynamic cellular, multicellular and tissue function (Figure TH2).

Molecular Building Blocks in Context: Molecular building blocks self-organise and exhibit collective properties that can serve to counter the variability seen in living systems. Examples of such collective processes are the functioning of the transcriptional machinery of eukaryotic cells involving tens or hundreds of protein subunits, or the build up and maintenance of cytoskeleton components out of large numbers of monomers. **Statistical mechanics** shows that the joint operation of such interacting elements leads to collective effects including phase separation and phase transitions. **Phase separation** has recently been identified as a fundamental organisation process in cell biology. This provides scientists with an opportunity to use theoretical concepts to address questions such as scaling (e.g. how does the size of membraneless organelles depend on the global physiological state of cells) and the derivation of phase diagrams. In a different context, **phase transitions** can be related to cellular differentiation events. Again, tools of statistical

physics are very important to study those processes. In particular, continuous phase-transition points are characterised by critical states in which fluctuations of all scales coexist. Critical behaviour has been associated with optimal information transfer in different physical and biological systems, and it appears to be a fundamental organising principle in biology, although this is still an open question. Additionally, critical phenomena are known to be a signature of universal properties and provide a basis for the **coarse-graining** of collective phenomena. Therefore, identifying critical states within cells and tissues in a univocal manner will allow researchers to establish universal unifying principles of living processes and to determine their correct scale of description.

Spatial modelling and computational simulations are already used to determine structures of biological macromolecules and analyse their dynamics. Better **theoretical understanding of the experimental processes and associated noise models** that generate the data used for structure determination, for example in cryo-electron microscopy, would facilitate building more accurate structures and assessing their uncertainty. Establishing expertise in **molecular dynamics** and other **simulation approaches** would enable understanding the dynamics of molecular systems at multiple scales, from conformational changes in macromolecules to the spatial dynamical behaviour of subcellular systems.

The **core elements** of EMBL's new theory theme:

- I. A new **Theory Transversal Theme** (Chapter 14: People, Processes, and Places) will be established at EMBL. This pan-EMBL structure will allow the recruitment of group leaders in the theory area and be an organising structure for visiting scientists (see below). Establishing theory as a research theme at EMBL is essential to make further efforts viable to recruit theoretically-inclined staff across the organisation, from PhD students to group leaders. An environment composed of a critical mass of people following a coherent paradigm is crucial for such research to flourish and acquire visibility. Within this new Programme, EMBL aims to build local alliances between EMBL and European institutes.



A pilot project is being established at EMBL Barcelona: the EMBL-CRG Collaborative Environment for Data-Driven Predictive Modelling. This is an exciting new initiative where visiting and local theoreticians will work together to develop computational modelling from the molecular to the tissue scale, initially focusing on the multicellular dynamics of developing organs, tumour growth, and synthetic tissues.

- II. A new **Theory Visitor Programme** will be implemented at EMBL to help bring seemingly disparate scientific disciplines together. Visiting theory researchers will be able to apply for financial support to cover visits at any EMBL site for a period of weeks to months, in order to initiate or deepen collaborations with EMBL groups. It will complement the newly established Theory research theme at its beginnings and it will help strengthen it through time.



A pilot project was established in February 2020 with the aim that three-week to six-month fellowships will help promote theory-based approaches across EMBL through the formation of new contacts and collaborations.

(<https://www.embl.org/about/info/scientific-visitor-programme/theoryembl/>)

- III. EMBL will offer new **training opportunities** in theory-driven approaches for the next generation of life scientists both at EMBL and in the member states. A common language and reciprocal understanding (and appreciation) are crucial to enable the new generation of life scientists to lead a constructive dialogue with theoreticians. Interdisciplinary training will also make it easier for theoreticians to enter into biology and vice versa.

- IV. EMBL will hold **conferences and workshops**, including smaller, highly intensive scientific meetings as well as broader European strategy meetings. Some meetings have already been planned, such as the ‘Perspectives of Theory in Biology’, and the EMBO | EMBL Symposium ‘Theory and Concepts in Biology 2022’.

Theory at EMBL will be treated as a fundamental addition not present in the current configuration of EMBL science. It raises the need for, and at the same time will catalyse, even deeper interdisciplinarity. It will be firmly linked to EMBL’s research and training missions, as theory will be supported by these cornerstone missions and will drive them further.

Impact

The proposed new emphasis on a firm interplay between theory and experimentation will be an integral requirement for achieving EMBL’s overall goals outlined in this EMBL Programme. EMBL’s new research directions, from molecules to ecosystems, will only be possible by establishing a firm link between theory and experimentation to uncover general principles and gain predictive understanding. Theory is needed to achieve a sound understanding of the multi-scale hierarchical organisation of biological systems and to design experiments and data acquisition at each relevant scale and across scales. These efforts conversely hold the potential to unlock substantial research advances in theoretical fields ranging from applied mathematics to statistical physics.

Formulating general principles enables researchers to make testable and quantitative predictions, which in turn guide future biological experiments. Predictive models enable systems engineering, including biotechnologies (e.g. synthetic biology for energy, food, or chemicals production), medicine (targeted, curative therapies), geoengineering (understanding and mitigating greenhouse gases emissions and the effects of climate change), and planetary biology (modelling biodiversity loss and resistance spread).

Theory helps to create interdisciplinary contacts and can serve as an accelerator for scientific collaborations, granting theoreticians greater access to data and experiments. A new theoretical research track will also strengthen and complement existing, data-driven, computational, and service activities across EMBL. Embedding a dedicated theory research programme within EMBL will generate outstanding opportunities for fundamental research as well as for the life sciences community. To collect and incorporate input from the community, EMBL researchers invited 16 international theory researchers with a strong link to life sciences to participate in a strategic discussion in which the participants will outline the benefits of bringing more theory to the life sciences and map out the best ways to do so. The participants will share their views on the future potential of theory in biology and will discuss the implementation of theory in the context of EMBL. The outcomes and recommendations that result from this workshop will constitute additional input that will be integrated at a later stage.

With the explosion in data-generating and measurement technologies over the past three decades, biology risks becoming mostly descriptive, with vast amounts of data that are hard to navigate, further increasing the risk of subdividing into evermore specialised subfields that are siloed. Such directions will be contrary to the primary purpose of scientific research. Strengthening theory, coupled with experimentation, will provide a feasible mechanism to support high-quality, resource-efficient, and conceptually advanced science.