

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.

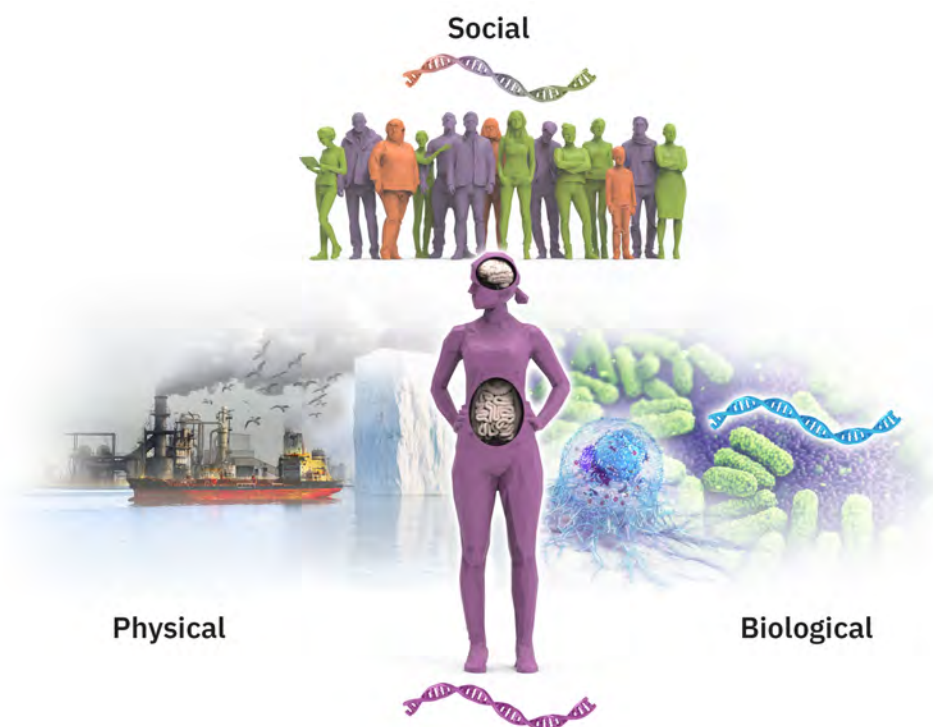


Figure HE1 | Human ecosystems.

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

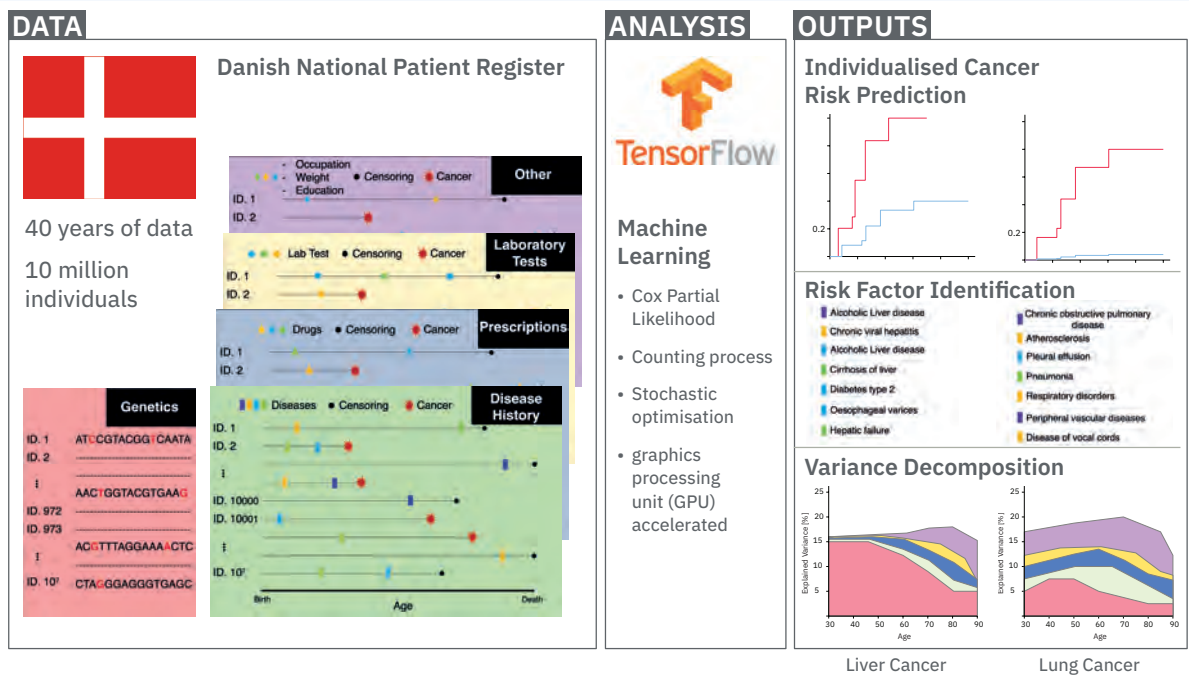
Novel statistical approaches can be applied to these data to enhance the quantitative understanding of diseases. This is critical in order to extract meaningful signals from big molecular datasets, such as genomics, epigenomics and transcriptomics, as well as large longitudinal records for (thousands 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.

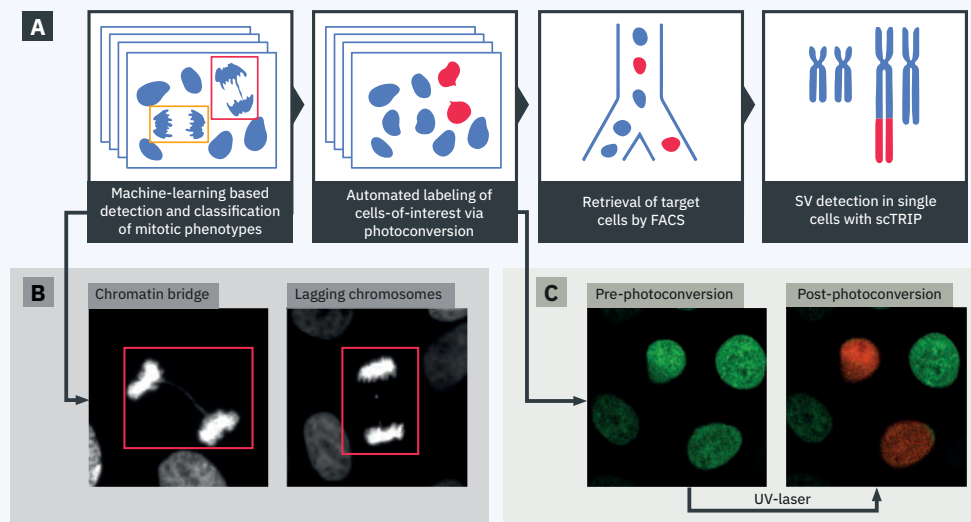
Another area where EMBL's expertise will be valuable is to understand the effects of drugs on genome variation, such as how chemotherapy induces genome instability. The genome of cancer cells is dominated by extensive somatic DNA rearrangements, also known as genomic structural variants (SVs), which 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.

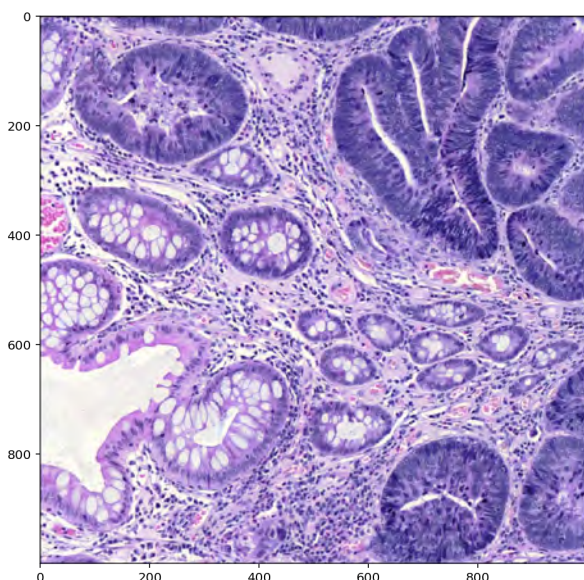


Figure HE2 | Colorectal cancer spatial omics study.

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.