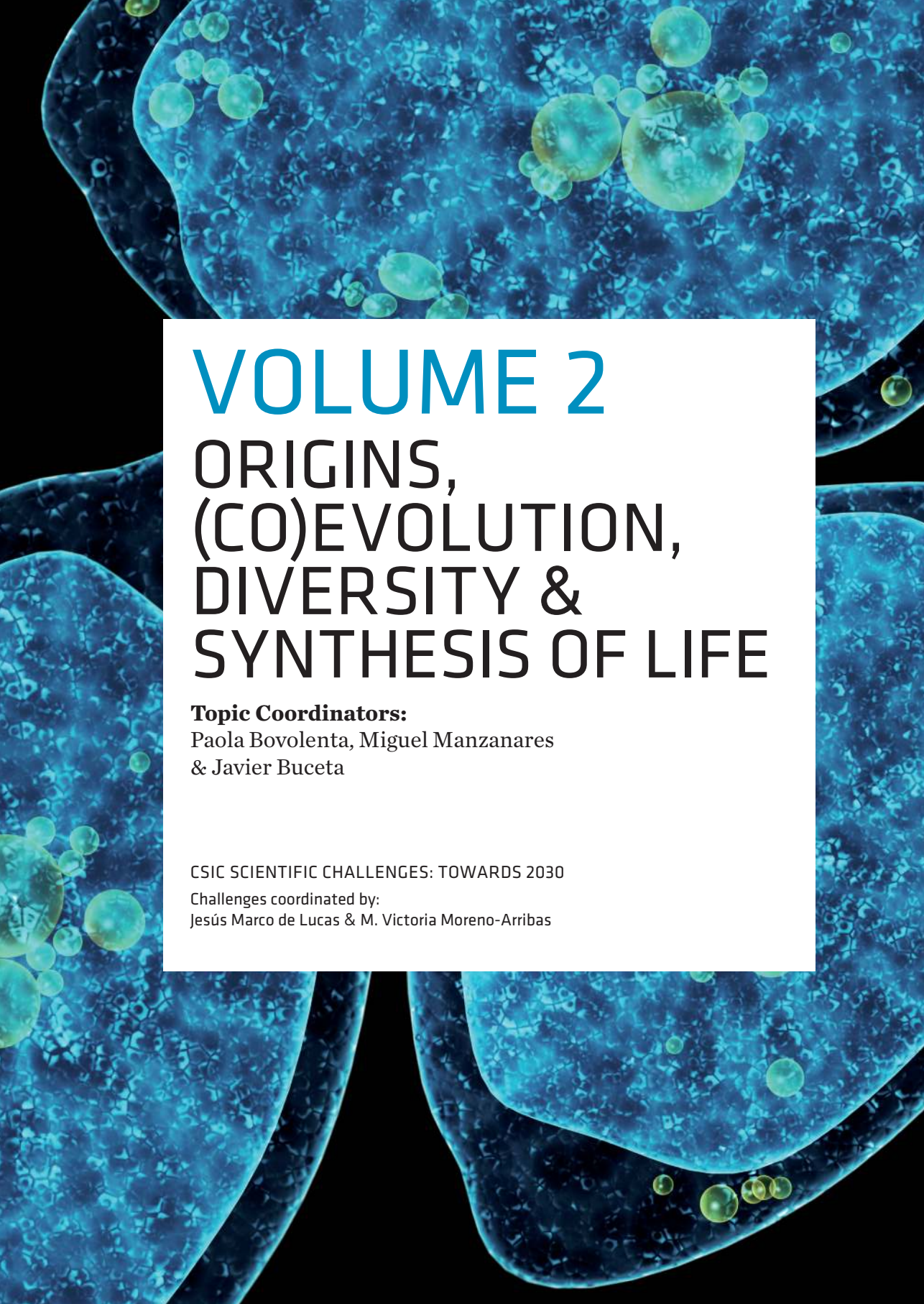


VOLUME 2

ORIGINS, (CO)EVOLUTION, DIVERSITY & SYNTHESIS OF LIFE

Topic Coordinators:
Paola Bovolenta, Miguel Manzanares
& Javier Buceta

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030

Challenges coordinated by:
Jesús Marco de Lucas & M. Victoria Moreno-Arribas

VOLUME 2

ORIGINS, (CO)EVOLUTION, DIVERSITY & SYNTHESIS OF LIFE

Catálogo de publicaciones de la Administración General del Estado:
https://cpage.mpr.gob.es

Editorial CSIC:
http://editorial.csic.es (correo: *publ@csic.es*)

ISBN Vol. 2: 978-84-00-10737-6
ISBN O.C.: 978-84-00-10736-9
e-ISBN Vol. 2: 978-84-00-19735-2
e-ISBN O.C.: 978-84-00-10734-5
NIPO: 833-21-009-9
e-NIPO: 833-20-203-8
DL: M-2426-2021

Diseño y maquetación: gráfica futura

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030

VOLUME 2

ORIGINS, (CO)EVOLUTION, DIVERSITY & SYNTHESIS OF LIFE

Topic Coordinators
Paola Bovolenta, Miguel Manzanares
& Javier Buceta

CSIC SCIENTIFIC CHALLENGES: TOWARDS 2030

What are the major scientific challenges of the first half of the 21st century? Can we establish the priorities for the future? How should the scientific community tackle them?

This book presents the reflections of the Spanish National Research Council (CSIC) on 14 strategic themes established on the basis of their scientific impact and social importance.

Fundamental questions are addressed, including the origin of life, the exploration of the universe, artificial intelligence, the development of clean, safe and efficient energy or the understanding of brain function. The document identifies complex challenges in areas such as health and social sciences and the selected strategic themes cover both basic issues and potential applications of knowledge. Nearly 1,100 researchers from more than 100 CSIC centres and other institutions (public research organisations, universities, etc.) have participated in this analysis. All agree on the need for a multidisciplinary approach and the promotion of collaborative research to enable the implementation of ambitious projects focused on specific topics.

These 14 "White Papers", designed to serve as a frame of reference for the development of the institution's scientific strategy, will provide an insight into the research currently being accomplished at the CSIC, and at the same time, build a global vision of what will be the key scientific challenges over the next decade.

VOLUMES THAT MAKE UP THE WORK

1 *New Foundations for a Sustainable Global Society*
2 *Origins, (Co)Evolution, Diversity and Synthesis of Life*
3 *Genome & Epigenetics*
4 *Challenges in Biomedicine and Health*
5 *Brain, Mind & Behaviour*
6 *Sustainable Primary Production*
7 *Global Change Impacts*
8 *Clean, Safe and Efficient Energy*
9 *Understanding the Basic Components of the Universe, its Structure and Evolution*
10 *Digital and Complex Information*
11 *Artificial Intelligence, Robotics and Data Science*
12 *Our Future? Space, Colonization and Exploration*
13 *Ocean Science Challenges for 2030*
14 *Dynamic Earth: Probing the Past, Preparing for the Future*

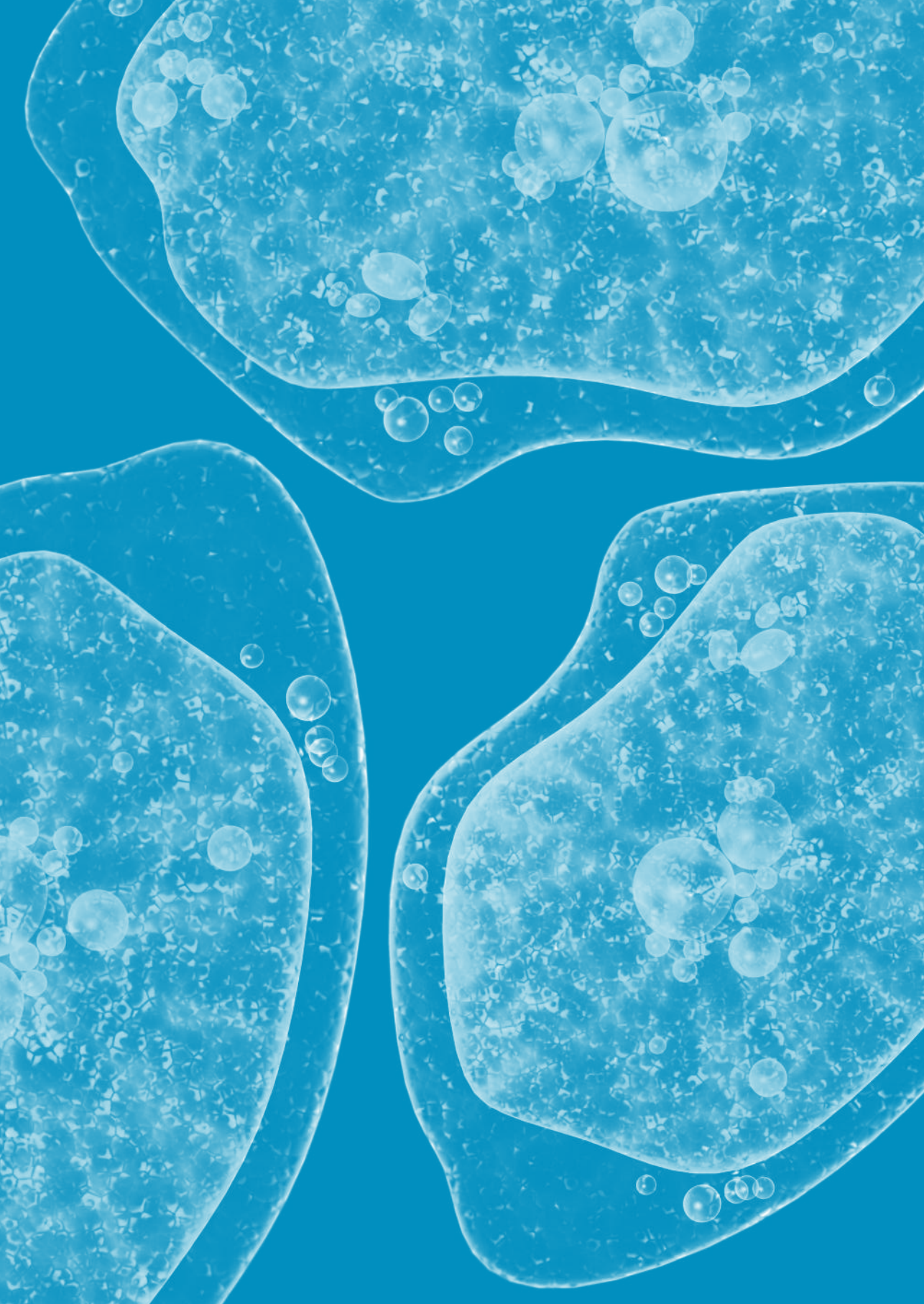
CSIC scientific challenges: towards 2030
Challenges coordinated by:
Jesús Marco de Lucas & M. Victoria Moreno-Arribas

Volume 2
Origins, (Co)Evolution, Diversity and Synthesis of Life

Topic Coordinators
Paola Bovolenta(CBM-SO), Miguel Manzanares (CBM-SO) and Javier Buceta (I2SysBio CSIC-UV)

Challenges Coordinators
Carlos Briones (CAB, CSIC/INTA); Izaskun Jiménez-Serra (CAB, CSIC/INTA); José M. Valpuesta (CNB); Santiago Ramón-Maiques (CBMSO); Rafael Zardoya (MNCN-CSIC); Ana Riesgo (MNCN-CSIC); Fernando Casares (CABD, CSIC-UPO); Ignacio Maeso (CABD, CSIC-UPO); Sergi Valverde (IBE); Saúl Ares (CNB); Carles Lalueza-Fox (IBE); Ignacio de la Torre (IH-CCHS); Santiago F. Elena (I2SysBio); Iñaki Comas (IBV); Germán Rivas (CIB-MS) and Eva García (ICP)

Participant Centers
Centro de Astrobiología (CAB, CSIC/INTA)
Centro Andaluz de Biología del Desarrollo (CABD-CSIC)
Centro de Biotecnología y Genómica de las Plantas (CBGP)
Centro de Biología Molecular Severo Ochoa (CBM-SO)
Centro de Investigaciones Biológicas Margarita Salas (CIB-MS)
Centro de Investigación Cardiovascular (CIC)
Centro de Investigaciones sobre Desertificación (CIDE-CSIC)
Centro Nacional de Biotecnología (CNB)
Estación Biológica de Doñana (EBD-CSIC)
Estación Experimental de Zonas Áridas (EEZA-CSIC)
Instituto de Biología Integrativa de Sistemas (I2SysBio CSIC-UV)
Instituto Andaluz de Ciencias de la Tierra (IACT, CSIC-UG)
Instituto Botánico de Barcelona (IBB-CSIC)
Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC CSIC-UNICAN)
Instituto de Biología Evolutiva (IBE)
Instituto Biofisika (IBF, CSIC-UPV-EHU)
Instituto de Ciencias del Mar (ICM-CSIC)
Instituto de Ciencia de Materiales de Madrid (ICMM)
Instituto de Filosofía (IFS-CCHS)
Instituto de Historia (IH-CCHS)
Instituto de Investigaciones Marinas (IIM)
Instituto Mediterráneo de Estudios Avanzados (IMEDEA-CSIC)
Instituto de Neurociencias (IN, CSIC-UMH)
Instituto de Nanociencia de Aragón (INA)
Instituto de Parasitología y Biomedicina López-Neyra (IPBLN)
Instituto de Productos Naturales y Agrobiología (IPNA-CSIC)
Instituto de Química Física Rocasolano (IQFR-CSIC)
Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS-CSIC)
Museo Nacional de Ciencias Naturales (MNCN-CSIC)
Real Jardín Botánico (RJB-CSIC)
Universidad Autónoma de Madrid (UAM)
Universidad de Alcalá de Henares (UAH)
Universidad de Extremadura (UEX)



CONTENIDO

ABSTRACT

How life appeared on Earth and how then it diversified into the different and currently existing forms of life are the unanswered questions that will be discussed this volume. These questions delve into the deep past of our planet, where biology intermingles with geology and chemistry, to explore the origin of life and understand its evolution, since "nothing makes sense in biology except in the light of evolution" (Dobzhansky, 1964). The eight challenges that compose this volume summarize our current knowledge and future research directions touching different aspects of the study of evolution, which can be considered a fundamental discipline of Life Science. The volume discusses recent theories on how the first molecules arouse, became organized and acquired their structure, enabling the first forms of life. It also attempts to explain how this life has changed over time, giving rise, from very similar molecular bases, to an immense biological diversity, and to understand what is the phylogenetic relationship among all the different life forms. The volume further analyzes human evolution, its relationship with the environment and its implications on human health and society. Closing the circle, the volume discusses the possibility of designing new biological machines, thus creating a cell prototype from its components and whether this knowledge can be applied to improve our ecosystem. With an effective coordination among its three main areas of knowledge, the CSIC can become an international benchmark for research in this field.

KEYWORDS

tree of life | archeology of human evolution | astrochemistry | bioengineering | developmental biology | structural biology | systems biology | evolutionary systems biology | synthetic biology | minimal cell | host-pathogen coevolution | evolution phenotype | phylogenomics | evolutionary genomics | genotype-phenotype map | darwinian medicine | minimal metabolism | RNA world | prebiotic chemistry | paleoanthropology | paleogenomics | protocols | complex systems

ORIGINS, (CO)EVOLUTION, DIVERSITY AND SYNTHESIS OF LIFE

Topic Coordinators
Paola Bovolenta, Miguel Manzanares and Javier Buceta

EXECUTIVE SUMMARY

Some of the mayor known unknowns of modern science deal with how life appeared on Earth and how from there it diversified into the different life forms present today. These questions delve into the deep past of our planet, where biology intermingles with geology and chemistry, to explore the origin of life. In addition, by learning how biological systems change over time, we can design novel biological machines based on this knowledge to fulfil unmet tasks. The main overarching theme addressed in this topic is evolution, given that "Nothing makes sense in biology except in the light of evolution" (Theodosius Dobzhansky, 1964). Understanding evolution will provide us with clues about the origin of life, and on the precise molecular mechanisms that operate in living beings and how they change in time. We will then be ready to attempt putting together these mechanisms in novel ways, paving the way for synthetic biology. Evolution is the single most overarching and one of the few, if not only, general principles in Biology.

In this volume, we explore and discuss various aspects related to the study of evolution, considered as a fundamental and central discipline in the Life Sciences, which should permeate the different areas of research in the coming years. The eight challenges that compose this volume summarize our current knowledge and future research directions in different but related aspects of this central theme.

The first challenge analyzes the origin and early evolution of life with the aim of explaining the synthesis of biochemical components and their polymerization to originate functional and information-bearing biopolymers responsible for the biochemistry of life, as well as to their coupling in autonomous systems with evolutionary capacities. Understanding the foundations of life and its evolutionary diversity also requires deciphering the three-dimensional structure and dynamic nature of all macromolecules, how they assemble and function in a coordinated, timely and precise manner. The second challenge addresses this problem and discusses how this knowledge could foster our understanding and treatment of diseases, and enable us to exploit biological processes for biotechnological purposes and for designing new biological entities.

Once created, "life" has evolved giving rise to an immense biological diversity, the proportion of which we do not really know. The third challenge discusses how the widespread application of high-throughput genomic techniques will allow the reconstruction of the Tree of Life and the identification of genomic targets of natural selection, providing fundamental keys to understanding the genesis of this diversity. However, are all genomic changes reflected in phenotypic changes? The fourth challenge aims to answer this question by analyzing how a phenotype is generated. This is a fundamental question in biology, with practical implications for human health, food production or climate change, that also affects several areas of engineering and the social sciences. This same problem can be analyzed with a systemic approach. Thus, the fifth challenge aims at generating a mechanistic and evolutionary understanding of genotype-phenotype maps at multiple scales using a combination of mathematical, molecular and cellular approaches.

Within the evolutionary scenario, Human Evolution has always attracted particular attention. The sixth challenge seeks to understand the processes of social and biological adaptation that took place throughout human evolutionary history. These processes have molecular, genetic, behavioral, social and anatomical morphological dimensions, the understanding of which can only be resolved with new multidisciplinary and technological approaches. Closely linked to our curiosity about human evolution is our need to understand why we get sick and more generally why living beings get sick. Diseases are the result of homeostatic alterations, caused by endogenous (for example, hereditary diseases or cancer) or exogenous (for example, infections or poisoning) disturbances. The seventh challenge examines diseases as the result of

co-evolutionary processes, providing a perspective that provides a better predictive capacity, for example, in the evolution of pandemics, and may allow fight diseases more efficiently.

Closing the loop and linking to the first, the eighth challenge examines the possibility of assembling a minimal vital unit with programmable functionality. Being able to build a synthetic cell from its essential components should contribute to our understanding of the basic principles of life, providing tools for novel solutions for environmental and biomedical problems.

The CSIC is in a unique position to face the challenge of understanding evolution with all its implications and ramifications, and through an effective coordination among its three major areas of knowledge, it can become an international benchmark for research in this field.

INTRODUCTION

Some of the mayor known unknowns of modern science deal with how life appeared on Earth and how from there it diversified into the different life forms present today. These questions delve into the deep past of our planet, where biology intermingles with geology and chemistry, to explore the origin of life. In addition, by learning how biological systems change over time, we can design novel biological machines based on this knowledge to fulfil unmet tasks.

The main overarching theme addressed in this topic is evolution. Not because of much repeated is Theodosius Dobzhansky 1964's statement less true: *Nothing makes sense in biology except in the light of evolution*. Understanding evolution will provide us with clues about the origin of life, and on the precise molecular mechanisms that operate in living beings and how they change in time. We will then be ready to attempt putting together these mechanisms in novel ways, paving the way for synthetic biology. Evolution is the single most overarching and one of the few, if not only, general principles in Biology.

In this topic, we explore and discuss several aspects related to the study of evolution, as a fundamental and core discipline in Life Sciences, which must permeate different research areas in the coming years.

Origins, trees and the genesis of the phenotype. The origin and early evolution of life is one of the most challenging scientific topics, as it aims at explaining the synthesis of biochemical building blocks and their polymerization to give

rise to functional and informative biopolymers responsible for life's biochemistry. This was a complex process that allowed the transition from chemistry to biology on Earth, perhaps even beyond our planet. Multiple mechanisms could have contributed to biogenesis, including the synthetic reactions in astrophysical environments, the delivery of organics by meteorites and cometary nuclei, as well as the geophysical and geochemical processes that took place on early Earth, about four billion of years ago. The discovery of over two hundred small molecules in interstellar space and extra-terrestrial bodies suggests that basic prebiotic processes are ubiquitous in the Universe.

The combination of compounds in different environments on early Earth gave rise to a growing number of bio-monomers, which could eventually couple into autonomous systems. These would represent the first self-reproducing and evolvable organisms should have been able to keep their molecular components together and distinguish themselves from their environment; stay away from thermodynamic equilibrium by capturing energy and material resources from the environment; and transmit heritable information to their progeny. Membrane compartments, metabolic machineries and replication mechanisms should have thus originated and combined in the transition from complex (though still thermodynamically driven) chemical systems into proto-biological ones and, eventually, into (kinetically and spatially controlled) living organisms.

Understanding how these components come together and change is also a fundamental aspect of modern evolutionary studies. This requires deciphering the three-dimensional structure and dynamic nature of all macromolecules underlying living processes, and how they ensemble and function in a coordinated, timely and precise manner. Addressing this challenge will allow us to understand and treat diseases, harness biological processes for biotechnological purposes and synthetically design new biological entities. Macromolecular machines, formed mostly by proteins and nucleic acids, have been perfected through evolution, becoming more complex, sophisticatedly regulated and integrated into dedicated operative pathways. In the case of proteins, the major working molecules of life, their functions are largely dependent on their three-dimensional shape and dynamics. It is therefore necessary to learn these processes at atomic level to understand the molecular mechanism of action in depth, including its dynamics, which would allow us to find solutions for their malfunctioning and ultimately create new activities for our benefit in biomedicine and biotechnology. In the case of the nucleic acids, the

structural landscape is much more diverse than previously thought. Non-regular DNAs are emerging as key structures in a variety of biological processes, such as genome transcription, repair or telomere maintenance, and RNA transcripts, including small and long non-coding RNAs appear to regulate almost every step of gene expression and have broad impacts on development and disease.

Paradoxically, the function of an increasingly number of proteins, protein regions and also RNAs appears to reside in their ability to remain unstructured. These intrinsically disordered macromolecules are involved, among other things, in promoting changes in the physical state of the cell milieu (liquid-liquid phase separation), allowing the formation of membrane-less cell compartments with multiple purposes such as transient storage or stress-response functions.

A further level of complexity is the analysis of how, where and when these macromolecular machines assemble and act in concert. The cell can be considered as a factory with multiple and sometimes short-lived compartments where the macromolecular content depends on cell needs and must be carefully controlled. This subcellular arrangement of macromolecules and their corresponding associated functions – termed "molecular sociology" of the cell – has a purpose that needs to be recognised. Of particular interest is the role of membranes not only as barriers to separate cellular components, but also as clustering regions for specific functions in which membrane proteins have key roles. A thorough comparative structural analysis of proteins, RNA and macromolecular complexes working in similar processes in different organisms will provide essential information to reconstruct the history and evolution of life. Ultimately, this knowledge will allow us to design new biological objects and entities and harness synthetic biology.

A fundamental aspect in evolutionary studies is to unravel the history behind diversification of life and being able to organize living beings in a structured and meaningful tree of life, where the relationships between species is established. The advent of high-throughput sequencing permits assembling chromosome-level genomes, characterizing single-cell transcriptomes, and determining epigenomic modifications. Once widely applied to the diversity of living organisms, the reconstruction of the Tree of Life and the identification of the genomic targets of natural selection will be achieved. Main efforts will be centred on obtaining samples from biotic frontiers, dealing with giant genomes and important proportions of repetitive elements, identifying

homology types and ploidy, detecting genomic hallmarks of selection, inferring candidate gene functions, and on gathering and incorporating long term natural history, geological, ecological, and environmental associated metadata under a phylogenetic framework. In the long run, we should be able to catalogue biodiversity, unveil the mechanisms underlying evolutionary adaptation, and to direct our conservation efforts based on evidence.

Knowing the basic cellular constituents and how they appeared during evolution, the next open question is understanding how these basic molecular and cellular systems achieve a higher-level organization as individuals. Here, we have to comprehend the development, maintenance and decline of living beings, i.e. the genesis, of the phenotype. A multidisciplinary approach based on the framework of evolution needs to address how biological information is stored and how this information is maintained, inherited, and changed, as well as how this encoded information drives the emergence of the phenotype, i.e. of all perceivable traits and processes of living systems. We also must understand how predictable is the phenotype, in order to assess our capacity for engineering living systems. Finally, it is paramount to address how new phenotypes evolve and what are the temporal and spatial scales of evolutionary phenotypic changes.

Social, environmental and health implications of human evolution. Questions posed when studying evolution acquire a distinct interest when applied to the human being. Not only can evolution help us understand the origin of our species and the historical patterns of migration and colonization of various habitats around the world, but also provide a fresh look at another seemingly more distant problems such as health and disease.

One of the greatest challenges in the study of human evolution is to understand the social and biological adaptive processes that took place along our evolutionary history. In a deeper timeframe, the continuous sourcing of full-genomes from multiple species that afford information about the phylogeny of our lineage allows for detailed studies on when and how some human-defining traits appeared along evolution. The application of -omics techniques to the study of the past now enable researchers to tackle a range of questions that previously were targeted almost exclusively by disciplines in the Humanities such as History and Archaeology, when investigating migrations; or in Medicine, such Medical Genomics, when investigating the genetic architecture of complex traits and diseases. In a deeper timeframe, the continuous sourcing of full-genomes from multiple species that provide

information about the phylogeny of our lineage allows for detailed studies on when and how some human-defining traits appeared along evolution.

As for the study of human cultural and behavioural evolution, primarily addressed through archaeological studies, disciplinary challenges should attempt to redefine the role of technology in shaping past societies but also in transforming the environment, as well as the nature of interactions between biotic and abiotic agents throughout the evolution of our Genus. We should aim at the identification of patterns, rather than events, in the course of the behavioural and cultural evolution of our species.

To what extent biotic versus abiotic evolutionary factors dictate divergent trajectories in the human past is one of the top questions in the study of past human behaviour. From the role of climate in early human adaptations to its relevance in the emergence of food production, the secular interest in the influence of abiotic causes in social evolution is now leading to a shift in perspective. The new grand challenges in social evolution should highlight, for example, the importance of mammal community structure in the shaping of early human behaviour, or the impact of human actions over fauna and flora and, in recent times, even over the climate.

Although often neglected, the emerging field of evolutionary medicine is providing fresh answer to long sought-after questions. Diseases result from the disturbance of the physiological homeostasis of organisms, and they can be endogenous (e.g., heritable diseases, developmental constraints or cancer) or exogenous (e.g., infections or intoxications). In as much as organisms themselves, and the way they interact with their biotic and abiotic environments, are the result of evolutionary forces, their diseases are also the result of complex co-evolutionary processes. By incorporating an evolutionary perspective, we would better understand and combat disease. This can be applied not only to human, but also to farm animals and crop diseases. Important areas of research are the study of the origins and spreading of new infectious agents, the appearance of multi-drug resistant microorganisms, or the dynamics of cancer cells and tumour growth. Also, the study of the evolution of microbial communities will be of great importance, in light of the role of the microbiome in multiple aspects of health and disease. Furthermore, understanding the origin of broad multigenic medical conditions such as diabetes, obesity, cardiovascular disease, autoimmune diseases, and in the long run aging, will sure be instructive to design novel therapeutic approaches.

All these processes are governed by the same basic and universal forces of evolution: mutation, genetic drift, migration, natural selection, and adaptive trade-offs. It is becoming increasingly clear that solutions to such complex problems potentially involve every level of organization, from molecules to populations.

Future life: mastering evolution? As we have discussed above, evolution is a complex, multilevel process that operates at long time scales. Therefore, it becomes obvious that it can only be understood from a systems perspective. Though we have a considerable wealth of experimental data, the major challenge is to develop models and theoretical frameworks to understand empirical results and to pose better focused experimental questions. One of the first unsolved questions, as previously described, is how phenotypes arise from genotypes. Can complex biological functions be constructed from simpler modules? How do regulatory circuits emerge? What are the limits to the design of robust and portable functional modules? Answers to these questions will assess the validity of reductionist approaches, as opposed to viewing innovation as an emergent phenomenon, arising from network-like distributed properties. In a broader framework, we should be concerned about the mechanistic origin of evolutionary transitions, and on the role played by external forcing versus contingent or stochastic phenomena in their generation.

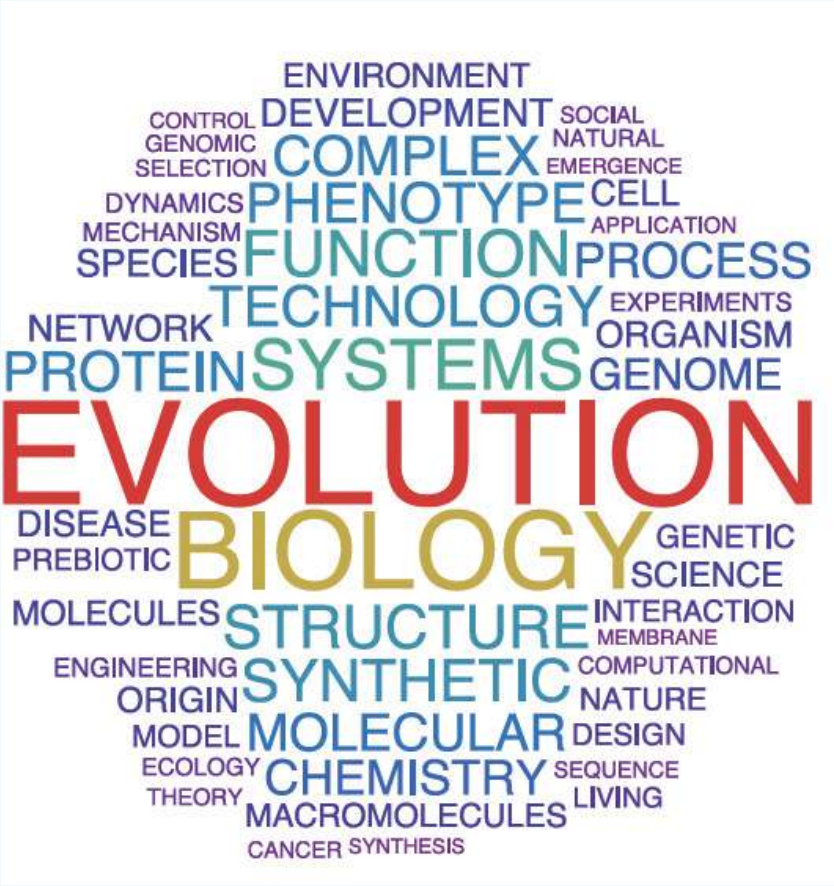
Evolution itself can be viewed as a tool for Synthetic Biology, since directed evolutionary selection is a way to attain desired functions. A major goal is to integrate design with a selection-driven exploration of phenotypic spaces. Advances will be conditional on the construction of large evolutionary platforms where experimental evolution informed by design and theory can proceed *en masse*. Research at the frontiers between evolutionary science, climate, and ecology address not only how will the ecosystem be affected by these organisms, but also how will the organisms be affected by the ecosystem in a rapidly changing environment.

Obviously, the questions posed above impact directly on the aims of synthetic biology, that are to assemble a minimal living unit with programmable functionality. Achieving this grand challenge, building a synthetic cell from scratch, will contribute to our understanding of the basic principles of life and its origin from lifeless components. It will also provide novel solutions to outstanding environmental and health-related problems.

We still do not understand how these pieces interact in a coordinated way to develop cellular functions. In this line, the generation of life from the molecular components existing on the primitive earth is one of the great enigmas of life, and thus a major scientific challenge. Synthetic biology offers new strategies for its resolution. From a fundamental perspective, the integration of molecular modules that will give rise to functional synthetic cells will help to reveal the limits of life. In this regard, we can envision that the merging of synthetic biology with molecular and cellular evolution may end up in the synthesis of living cells from scratch, the function of which will be tuned by controlled evolutionary mechanisms. Other novel approaches that take into consideration evolutionary principles and integration of smaller independent units is that of swarm robotics. To put it simply, understanding how evolution produced living systems as those around us, will allow us to design and generate artificial living systems from scratch.

A quantitative analysis of a collective effort. A quantitative approach using a wordcloud analysis (Fig. 1), which identified the most represented terms in the collection of the eight chapters that follow, allows to draw some general conclusions about the most important concepts stressed in this white book. As one might have guessed, evolution in biology, is the main theme of these chapters thus justifying the quote from Theodosius Dobzhansky that we used to start this introduction (“*Nothing makes sense in biology except in the light of evolution”*).

Interestingly, underlying this main theme, we can clearly see the foundations and approaches that can help to understand evolution in a broad sense. Thus, these chapters or challenges (C) have revealed that the likely most valid approach to understand the “evolution problem” must be necessarily multi-scale and integrative: from the genome to the whole organism, including their ecological and social relations, and including all the intermediate scales (e.g., molecules, cells, ...). This idea is beautifully captured and stressed by the term “Systems” in the wordcloud. The reasons that justify this methodological viewpoint is the complexity, and the intricate network of interactions between components at different scales, that leads to the relations between structure (in a broad sense) and function. A second conclusion refers to the tools needed to implement such progress of the field: newer technologies, computational and engineering approaches, and theoretical frameworks able to propose quantitative models and mechanisms, are needed in combination with genomic and synthetic biology approaches. A third conclusion refers to the possible



applications derived from these studies. Thus, a deep understanding of "biological evolution" by these means is the way to gain knowledge and to propose solutions about diseases (e.g., cancer), their possible environmental drivers, and, eventually, to shed light into the origin of the emergent phenomenon that we call Life.

Evolution @ CSIC. As outlined above and explored in more detail in the following pages, only the collective effort among researchers belonging to the three domains in which the CSIC is currently divided -Life, Materia, Society- can lead to a better understanding of Life, its origin, diversity and long history of co-evolution, meaning with this term, that the evolution of each single

organism depends on that of the surrounding environment, including organisms, though the time scale might be different. The CSIC stands at a unique position among nationwide research organizations to tackle the challenge of understanding evolution with all its implications and ramifications. The CSIC counts with specialized research institutes dedicated to the study of living and fossil biodiversity, multiple institutes where the study of molecular mechanisms acting during evolution is a mayor goal, and dedicate research institutes to genomic evolution or system biology as well as with centres that address the different branches of mathematics, physics, chemistry, social sciences and humanities. This fragmented organization however should evolve at the same pace that research evolves. Mayor challenges can only be addressed from multidisciplinary perspective. The participation of researchers belonging to all the three major CSIC domains to the assembly of this topic is a clear reflection of this trend. Thus, the CSIC has the tools, the manpower, the knowledge and, more importantly, collectively has identified the most salient and unexplored challenges for the next decades within this topic. The grand challenge that now needs to be addressed is to efficiently coordinate the activity of a large number of researchers with different background and expertise, to create the appropriate settings for their activity, to foster without hesitation the emergent approaches, to increase the critical mass, to educate the scientists of the future and give them freedom and independence to explore their own ideas from the very beginning and to provide them with the financial support to be internationally competitive.

Physical proximity and continuous exchange of ideas is still the best breeding ground for innovative research. The creation of a dynamic Evolutionary Biology centre where to develop the activities here proposed should be thus among the mid-term goals of this grand challenge.

CHALLENGE 1

ABSTRACT

The origin and early evolution of life is one of the most challenging and interdisciplinary scientific topics. It aims at explaining the synthesis of biochemical building blocks and their polymerization to give rise to functional and information-carrying biopolymers responsible for life's biochemistry, as well as their coupling in autonomous systems with open-ended evolution capacities. CSIC has the potential to actively contribute to this field thanks to the highly relevant work already being developed by several of its research groups, the strong complementarities detected and the synergies that are bound to emerge among them, as well as with their collaborators in Spanish universities, other OPIs and international partners.

KEYWORDS

astrochemistry | chirality | complex systems | early Earth | evolution | genotype-to-phenotype map | *in vitro* evolution | interstellar medium | LUCA | metabolism | non-coding RNAs | origins of life | prebiotic chemistry | prebiotic systems chemistry | protocell | replication | RNA world | surface science | synthetic biology | vesicles | viroids | viruses

CHALLENGE 1

THE ORIGINS OF LIFE. FROM CHEMISTRY TO BIOLOGY

Coordinators
Carlos Briones, (CAB, CSIC/INTA)
Izaskun Jiménez-Serra, (CAB, CSIC/INTA)

Participant researchers and centers:
Jacobo Aguirre, (CAB, CSIC/INTA)
Alfredo Berzal-Herranz, (IPBLN)
Pedro Cintas, (Univ. Extremadura)
Emilio J. Cocinero, (IBF, CSIC-UPV-EHU)
Elena R. Alonso, (IBF, CSIC-UPV-EHU)
Andrés de la Escosura, (UAM)
Juan Manuel García-Ruiz, (IACT, CSIC-UG)
Jordi Gómez, (IPBLN)
David Hochberg, (CAB, CSIC/INTA)
Marta Ruiz, (CAB, CSIC/INTA)
Eva Mateo-Martí, (CAB, CSIC/INTA)
Susanna Manrubia, (CNB)
José Ángel Martín-Gago, (ICMM)
César Menor-Salván, (Univ. Alcalá de Henares)
Juli Peretó, (I2SYSBIO, CSIC-UV)
Germán Rivas, (CIB)
Kepa Ruiz-Mirazo, (IBF, CSIC-UPV-EHU)

EXECUTIVE SUMMARY

The origin(s) of life was a complex process that allowed the transition from chemistry to biology on Earth, perhaps even beyond our planet. Multiple mechanisms could have contributed to biogenesis, including the synthetic reactions in astrophysical environments, the delivery of organics by meteorites and cometary nuclei, and the geophysical and geochemical processes that took place on early Earth, about four billion of years ago. The discovery of over 200 small molecules in interstellar space and extra-terrestrial bodies suggests that basic prebiotic processes are ubiquitous in the Universe. Low molecular weight compounds have indeed been found in asteroids (parental bodies of meteorites) and comets: therefore, such minor bodies could have decisively contributed to the enrichment of the so-called "primordial soup" with organics and CHONPS-containing molecules. The combination of endogenous and exogenous compounds in different environments on early Earth gave rise to a growing number of biomonomers, which could interact through self-assembly processes and catalytic activities operating in aqueous solution and reactive interfaces (including water-ice and water-lipid) as well as on mineral and metal surfaces.

Open questions in the field of prebiotic chemistry include: the abiotic synthesis of the required organic molecules in plausible physicochemical conditions; origin of the homochirality of current biomonomers (e.g., sugars, nucleotides and amino acids); self-assembly and polymerization of certain biomonomers into longer polymers, such as RNA or peptides, in homogeneous or heterogeneous media; formation of amphiphile-based vesicles with enough permeability; rise of biochemical functions in compartmented microenvironments; origin of basic metabolic activities and heritable information; supramolecular association of the complex molecules generated; establishment of a genetic code and the origin of ribosomes, as well as the progressive genotype/phenotype decoupling. The first self-reproducing and evolvable cellular organisms should have been able to: i) keep their molecular components together and distinguish themselves from their environment; ii) stay away from thermodynamic equilibrium by capturing energy and material resources from the environment; and iii) transmit heritable information to their progeny. Membrane compartments, metabolic machineries and replication mechanisms should have thus originated and combined in the transition from complex (though still thermodynamically driven) chemical systems into proto-biological ones and, eventually, into (kinetically and spatially controlled) living organisms.

Prebiotic chemistry and biochemistry have dealt with partial aspects of such a complex problem. However, the physicochemical mechanisms involved in the formation of those infra-biological subsystems, when implemented and developed separately, often turned out to be incompatible. As discussed below, higher levels of molecular heterogeneity, together with the use of systems chemistry-based approaches, provide a more realistic scenario for the origins of life. In this framework, the traditional replication-first (i.e. the RNA world model) vs. metabolism-first (either autotrophic or heterotrophic) controversy, can be substituted by a scenario in which the key molecules (including monomers and oligomers, metal and mineral catalysts, or reactive interfaces with water-based media) could all co-evolve from the beginning, forming mixed, pre-biochemical interaction networks.

The implications of this new approach will be described in this chapter, both regarding the individual units (as self-maintaining systems with an internal organization) and in relation to their collective and long-term evolutionary dynamics (based on competition, collaboration and selection processes among those individuals, including the proto-ecological networks and syntrophic

relationships they must have established). Such an interdisciplinary approach can be accomplished by the CSIC groups and their partners in this scientific challenge.

1. INTRODUCTION AND GENERAL DESCRIPTION

From astrochemistry to astro-biochemistry: the quest for prebiotic molecules in space

The space in between stars is not empty but contains a high number and diversity of atoms, high-energy nuclei, (sub-)micron-sized dust particles and molecules. The study of their abundances and reactions occurring in space, and their interaction with radiation, is called Astrochemistry. This is a young discipline halfway between Astrophysics and Chemistry that started with the detection of the first molecules in space (the diatomic radicals CH and CN and the molecular ion CH^+) in the late 1930s with optical telescopes. We had to wait until the late 1960s to discover other molecular species such as OH, H_2O and NH_3 at centimetre wavelengths thanks to the development of radio-telescopes. Fifty years later, over 200 molecular species have been reported in space to date, which reveals that *our Universe is molecular* (see the Cologne Database for Molecular Spectroscopy, CDMS, for an up-to-date inventory of the molecular species detected in space).

Most of these molecules are organic and contain abundant interstellar elements such as H, C, N, and O. They are found across different environments (i.e., cold dark clouds, hot cores, circumstellar envelopes around evolved stars, etc.) through the analysis of the molecular line emission in spectroscopic surveys carried out at centimetre, millimetre and submillimetre wavelengths. These surveys are extremely congested with lines, known as "astrophysical weeds", which belong to the many molecules present in the interstellar medium (ISM).

A significant fraction of the species detected in space corresponds to large and heavy molecules (what, in this context, means that they have more than 5 atoms in their structures) named Complex Organic Molecules (COMs). The formation and survival of these large molecules is somewhat surprising. The ISM and the space around stars (the circumstellar medium, CSM), are harsh environments for molecules in general, and especially for COMs, because they are exposed to energetic phenomena such as UV/X-ray radiation, cosmic rays or shock waves.

Among the COMs detected in the ISM, there is a subset of species relevant for prebiotic chemistry and for theories of the origins of life (Patel et al., 2015). These include cyanamide (NH_2CN), cyanoacetylene (HC_3N), glycolaldehyde ($HOCH_2CHO$), formamide (NH_2CHO), or urea (NH_2CONH_2), which are nowadays routinely detected in the ISM (e.g. Belloche et al., 2008; Jorgensen et al., 2012; Jimenez-Serra et al., 2020). Despite this huge progress, the building blocks of terrestrial life such as amino acids (e.g. glycine), nucleobases (e.g. adenine) or relatively complex sugars (e.g. ribose) remain to be reported in ISM. Nevertheless, the presence of amino acids or monosaccharides precursors (Cocinero et al., 2012; Haykal et al., 2013; Peña et al., 2013; Alonso et al., 2019; Calabrese et al., 2020) and their prebiotic molecules in space (Alonso et al., 2016), is of utmost importance to understand the chemical processes that could lead to the formation of the building blocks of life, making their detection in the ISM almost as important as that of the amino acids or monosaccharides themselves.

Geochemical reactions and geological niches for the origin of life

The very name that geologists have given to the first 500 million years of the Earth, the Hadean eon (a Greek name that refers to the hell) is because the childhood of the planet was believed to be a hell, an inhospitable world of extreme conditions, of intense ultraviolet (UV) radiation, high temperature, no liquid water, and innumerable volcanoes and magma seas. This view has however changed drastically in recent years thanks to the information inferred from some rare preserved zircon crystals. The Earth and Earth-like planets and moons presented the required chemical and physical conditions for life to flourish much earlier than originally thought, i.e. about 4.4 billion years ago (Wilde et al., 2001). This is important when considering the time scales of the different processes involved in prebiotic chemistry.

Unveiling how a lifeless planet was like at that time is one of the inevitable tasks in the field of origin(s) of life. Water condensed on the surface of the planet soon after the solidification of the first ultramafic crust. The thermally-driven interaction between water and ultramafic minerals (serpentinization) unavoidably created an alkaline and reduced hydrosphere and a methane-rich atmosphere. Under these physicochemical conditions the planet should have been a global factory of simple and complex organic compounds (García-Ruiz et al., 2020). The question of why there is no evidence of the existence of life elsewhere may be then due to the difficulties to synthesize the simplest microorganisms, not so much the initial chemical building blocks, as explained next.

1.3. Prebiotic chemistry

Prebiotic chemistry encompasses the physico-chemical processes that occur within a given planetary environment, from its formation to the emergence of the first self-replicating systems on which the Darwinian process can begin to operate (Eschenmoser, 2007; Ruiz-Mirazo et al., 2014; Pross, 2016). Some of the key open questions include: i) the abiotic origin of the required organic molecules (the main field of prebiotic chemistry); ii) the origin of the homochirality of all the current bioorganic compounds, as it is currently accepted that it is impossible to construct any functional biopolymer using mixtures of enantiomeric compounds; iii) the polymerization of monomers, molecular self-assembly and reactivity on surfaces; iv) supramolecular association of the generated complex molecules and vesicle formation, within the framework of prebiotic systems chemistry (see below).

The hypothesis commonly accepted to explain the origin of life is that simple inorganic molecules (for instance, CH_4, CO_2, N_2, NH_3, or others single C molecules such as HCN, formamide or urea) reacted to form more complex organic compounds (Marín-Yaseli et al, 2016; Mompeán et al., 2019) or even oligomeric/polymeric substances (Mann, 2013; Ruiz-Bermejo et al., 2019) that then underwent subsequent reactions to yield biologically functional molecules. These compounds self-organized and self-assembled to increase the organization at molecular level. Such a succession of chemical processes is considered as the most likely pathway for the emergence of the first living organisms. Therefore, the first step to verify this theory is to check how the components of proteins and nucleic acids, or the constituents of other protobiological polymers, could be formed under abiotic conditions. Among the relevant physicochemical scenarios for prebiotic chemistry to operate, both "warm little ponds" and hydrothermal systems could have played a significant role. Also, oceanic organic and inorganic surface products could be transmitted within aerosols to the atmosphere, where they were then exposed to radiation and electrical discharges in the form of lightning, thus contributing to the origin of organic molecules under abiotic conditions.

Additionally, understanding how *single* chirality emerged on Earth (i.e., the homochiral L-amino acids and D-sugars in biopolymers) during molecular evolution is a crucial feature of any scientifically coherent proposal. Chemical evolution theory, therefore, not only should account for the transition from prebiotic chemistry (*i.e.*, a diverse pool of relatively simple achiral or racemic building blocks) to protobiology (supramolecular entities such as proteins,

polysaccharides, nucleic acids, and bilayer lipids, capable of mutual interaction, self-organization and Darwinian evolution) (Breslow and Cheng, 2009; Mann, 2013; Krishnamurthy, 2017), but it must also provide answers to outstanding questions such as: i) when did biological homochirality emerge?; ii) why is there a chiral bias for L-amino acids and D-sugars?; iii) how were chiral nonracemic amino acids and sugars formed from simple achiral molecules in the absence of asymmetric catalysts?

Answers to these questions require applying physical chemistry, chemical physics and non-equilibrium thermodynamics to understand chiral symmetry breaking and chiral amplification in prebiotically plausible, model reaction systems. This involves the study of autocatalytic reactions, chemical self-replicating systems, and chiral symmetry breaking processes, as precursors to the origin of biological homochirality (Ribó et al., 2017). Equally important are the conditions imposed by the thermodynamics of far from equilibrium systems and especially entropic aspects of molecular mirror symmetry breaking, and the application of maximum/minimum entropy production criteria as selection principles for these nonlinear chemical networks (Hochberg and Ribó, 2019; Hochberg and Cintas, 2020).

The interactions and reactivity between biomolecules and mineral surfaces are also important pieces in the puzzle of prebiotic chemistry, as minerals may adsorb and concentrate such biomolecules acting as catalysts for biochemical reactions. In addition to the widely accepted "primordial" or "prebiotic soup" theory, Huber and Wachtershauser proposed the "iron-sulphur world" theory, which states that the first reactions that led to the formation of amino acids did not occur in a bulk solution in the oceans but on the surface of minerals (such as pyrite), as they facilitate prebiotic polymerization and chemical reactions between organic molecules (Huber and Wachtershauser, 1998). Mineral surfaces could enable almost any type of catalysis, including those with low specificity and efficiency. However, the environmental conditions that favour molecular mineral adsorption, the reactions that can operate in such a heterogeneous system, and the precise nature of mineral-molecule interactions must be studied in detail. Surface science techniques are reliable tools to approach the study of molecular processes on catalytic minerals surfaces (Sanchez-Arenillas and Mateo-Marti 2016; Galvez-Martinez et al., 2019; Mateo-Marti et al., 2019). Simulations of polymerization processes using simple small molecules, especially under conditions thought to be plausible on early Earth and on other planets and moons, meteorites or comets,

are of intrinsic interest to understand chemical evolution towards biology (Lavado et al., 2018; Ruiz-Bermejo et al., 2019).

In parallel, on-surface synthesis has recently emerged as a promising approach to obtain carbonaceous nanostructures or complex molecules with atomic precision. Both one-dimensional polymers and intricate molecular structures have been synthesized directly on surfaces (Martín-Gago et al., 2011; Méndez et al., 2011). This method is based on the precise understanding of the interactions between adsorbed molecules and the surface of some materials, which catalyse coupling covalent reactions. One of the merits of on-surface synthesis is the ability to open new reaction pathways that are not possible via standard organic chemistry. This novel approach could thus provide an additional pathway, complementary to the proposed synthetic wet chemistry, to explain, for instance, the abiotic polymerization of nucleic acids from nucleotides, with or without the need of their previous chemical activation (Ferris, 2006). This is bound to have important implications for the origin of the RNA world.

1.4. The path to an RNA world

The discovery that certain specifically folded RNA molecules perform catalytic functions (acting as RNA enzymes or "ribozymes") put on the board that RNA can combine genotype and phenotype in a single molecular entity. This remarkable finding strengthened the hypothesis that the current world of living organisms, based on the flow of genetic information DNA→RNA→proteins, was preceded by an era when RNA was the sole genetic material and catalyst (Atkins et al., 2011; Ruiz-Mirazo et al., 2014).

Therefore, in the past few decades, experimental studies of the chemical roots of biochemical systems have been dominated by an RNA-centric perspective. In these studies, the construction of the molecular alphabet of RNA has focused primarily on the chemistry of hydrogen cyanide (HCN) or formamide ($CHO\text{-}NH_2$, a hydration product of HCN), as possible prebiotic precursors of nucleic acids (Oró and Kimball, 1961). However, essential questions remain to be solved: i) how were the building blocks of biopolymers (nucleotides and amino acids) formed?; ii) how were the first nucleosides formed and what was their fate in Earth's prebiotic environment?; iii) how were phosphate and/or other plausible linker groups, incorporated to chemical evolution (the so-called phosphate problem)?; iv) how were proto-biopolymers formed giving rise to supramolecular organizations which allowed the emergence of a system capable of Darwinian evolution?; and v) what are the plausible scenarios

in which the interaction between prebiotic chemistry and local geochemistry allowed the formation and chemical evolution of biopolymers?

In this field, one of the biggest challenges is to understand how the polymerization of the first proto-RNA structures took place, which led to supramolecular organization, selection and evolution towards macromolecular aggregates with essential biochemical functions, such a proto-ribosomes. The origin of the genetic code and full-fledged ribosomes allowed the information coded in RNA to be translated into proteins, thus decoupling the phenotype from genotype and triggering the broad exploration of the functional space (including catalytic activities required to establish a protometabolism) by a growing number of protein families (Ruiz-Mirazo et al., 2017).

However, protein-coding information only represents a small portion of the genetic load of modern-day living organisms. Indeed, an overwhelming amount of functional RNAs (named non-coding RNAs, ncRNAs) has been identified in current cells, providing evidence that the functional plasticity of the RNA has lasted until modern life. As a relevant example, RNA catalyses the peptide bond formation in ribosomes and provides the scaffold for the transfer of information from RNA to proteins, which supports the hypothesis that RNA preceded proteins in evolution, although peptides (and several low molecular weight molecules) could have modulated and complemented the catalytic properties of RNA (Ruiz-Mirazo et al., 2014).

Moreover, ncRNAs represent a critical piece in the process of gene expression regulation in most living organisms (Atkins et al., 2011). Additionally, RNA molecules with different capabilities have been produced from complex populations by *in vitro* selection/evolution experiments, leading to both artificial and engineered natural ribozymes (Puerta-Fernández, 2003; Joyce, 2004), as well as to aptamers with the desired target-binding activities (Briones, 2015) that show an increasing applicability in biotechnology and biomedicine (e.g., Sánchez-Luque et al., 2014; Moreno et al., 2019a). This methodology mimics the natural evolutionary processes in a test tube, under experimentally controlled conditions, and has greatly expanded the functional versatility of the RNA and its role in the origin of life (Joyce and Szostak 2018).

RNA function is an intrinsic consequence of its three-dimensional folding in solution, which in turn relies on the degenerated sequence-structure RNA map (different sequences can fold into the same secondary and tertiary structure, as it will be discussed below). Additionally, specific functions can be currently

performed by certain RNA structural domains that have been incorporated through evolution into much larger and versatile RNA molecules (that can contain both coding and non-coding regions), while, conversely, independently-structured short RNA molecules can have come together and play a new catalytic or regulatory role (Briones et al., 2009). Viral RNA genomes are good examples of such a functional versatility, fine-tuned through evolution. They encode structural/functional RNA domains that establish a close communication between them to render a complex and dynamic network of RNA-RNA interactions, the so-called RNA interactome, which ensures the completion of the viral cycle within the host cell as well as the virus adaptation to the cellular environment (Romero-López and Berzal-Herranz, 2020).

Different biochemical and biophysical methods have been used to search for ancient RNA elements that have preserved their specific secondary and/or tertiary structure in viral genomic RNAs and cellular mRNAs. This approach has been defined as "archaeological", as it can discover hidden evolutionary patterns through a non-phylogenetic and non-representational strategy (Ariza-Mateos et al., 2019). In particular, tRNA-like elements have been found in structurally or functionally relevant positions both in viral RNA and in certain mRNAs examined. The ligation-based concatenation (Briones et al., 2009) of these RNA motifs (among others) could have occurred in the RNA pools already present in the RNA world (Witzany, 2020). The extensive alteration of nucleotide sequences that likely triggered the transition from the predecessors of coding RNAs to the first fully functional mRNAs (which was not the case in the stepwise construction of ncRNAs), hinders the phylogenetic-based identification of RNA elements that could have been active before the advent of protein synthesis. Therefore, the archaeological method is a way to deepen into the structural/functional versatility of those RNA elements that had to adapt (losing their original function) to the selective pressures operating in the evolution from the RNA world to an RNA-protein world, in which the coding capacity of the progressively longer mRNAs was favoured (Ariza-Mateos et al., 2019).

1.5. Prebiotic systems chemistry as a new approach in the origins-of-life research

Systems chemistry represents a new research strategy in molecular sciences and encompasses a holistic view of complex chemical systems, understood as sets of diverse molecules interconnected through transformation and/or self-assembly processes. This field is called on to provide key insights into the

resolution of major open scientific questions, such as how living entities could emerge from inert matter (in connection with the origins of life research) and whether it is possible to engineer artificial life-like processes and materials (in relation with the field of synthetic biology and nanotechnology). Toward this goal, it is common to study the merging of different biological building blocks into synthetic systems, with properties arising from the combination of their biomolecular components. This approach is normally based on self-organization and/or chemistries that are dynamic in nature, and by looking closely at chiral effects of those components to reach specific morphologies. The long-term goal of this strategy is to create chemical networks and assemblies with emergent properties that are characteristic of life (Morales-Reina et al., 2020).

In the context of origins-of-life research, the concept of chemical evolution is central, as it encompasses plausible physicochemical mechanisms by which the first living protocells could have been assembled. Historically, this term began to be used shortly after the first steps in the field of prebiotic chemistry were taken, though its use has gained a renewed energy in recent years, thanks to the emergence of the field of prebiotic systems chemistry (Ruiz-Mirazo et al., 2014). The general view is that, in order to understand the transition from inanimate matter to living organisms, complexity must be embraced at the chemical level. We currently assume that the first living entities must have comprised, at least, a proto-cellular compartment, a proto-genome and an autocatalytic metabolic network supporting the system with energy and substrate molecules. Moreover, the replication dynamics of these three subsystems must have been coupled for the efficient reproduction of the system as a whole. Such requirements involve a great level of complexity, regarding both the molecular structure of the protocell components and their interaction dynamics, the establishment of which seems highly unlikely in the absence of an evolutionary driving force (de la Escosura, 2019). Indeed, the development of highly dynamic and integrated protocell populations (rather than complex reaction networks in bulk solution, sets of artificially evolvable replicating molecules, or even these same replicating molecules encapsulated in passive compartments) provides the most appropriate evolutionary framework to address the difficult problem of how prebiotic chemistry bridged the gap to full-fledged living organisms (Shirt-Ediss et al., 2017).

1.6. Emergence of function and molecular innovation in a pre-cellular world

One of the main difficulties in the way from simple chemical compounds to the first cell regards the emergence of functional molecules, which are basic bricks that may subsequently allow, through a process of trial-and-error, the construction of more complex ensembles endowed with a broader functional repertoire. The five essential concepts in the emergence of the first self-replicating molecular entities can be summarized as follows:

a. The phenotypic bias reflects the redundancy of the genotype-to-phenotype map. Sequences, in the form of random polymers, may uniformly explore the space of genotypes; however, phenotypes, here understood as molecular structures (and potentially functions) are not equally sampled. The fact that certain structures are way more likely than others (e.g. short open, random RNA sequences fold into hairpin-like structures; if circular, they preferentially adopt rod-like folds) may be behind the emergence of functions such as ligation (Briones et al., 2009) or of viroid-like replicators (Catalán et al., 2019).

b. The high dimensionality of sequence spaces entails the existence of astronomically large quasi-neutral networks of genotypes that guarantee a costless navigation and an efficient exploration of new molecular functions (Manrubia et al., 2020). A side effect of phenotypic bias is that the vast majority of sequences map into large, frequent phenotypes: these abundant phenotypes are parsimoniously attainable if evolution starts at any other viable phenotype. Thus, once any basic function is serendipitously found, the non-trivial topology of the genotype-to-phenotype map almost ensures that additional functionalities will emerge.

c. The emergence of a function does not imply its fixation. For a new function to become conspicuous, there should be an empty ecological niche. Sometimes, the molecular niche might be implicitly available (as could be the case of ligation, cleaving or replication once functional polymers are present), while in others it might be absent and it may appear as a new layer of complexity that depends on previously fixed functions (as in niche construction). An example could be molecular parasites, including viruses and viroids: selfish replicators that take advantage of the self-replicating machinery but do not give back to the system. Once parasites are in place, new parasites might find it difficult to displace the initial, opportunistic colonizers. Functions that increase

complexity, or the ability of the system to self-replicate, are subject to similar dynamics. The result of such first-come-first-served processes in molecular ecology is the well-known "frozen accidents" in evolution: suboptimal solutions whose only advantage is to have arrived first. The arrival of function is deeply related to phenotypic bias (Schaper and Louis, 2014).

d. An often forgotten mechanism favouring adaptation and the emergence of new function is molecular promiscuity. This is particularly true for RNA, as one given sequence may easily adopt different conformations (and, potentially, different functions) when the physical-chemical environment changes. Examples are RNA switches, which might be an important mechanism at early stages of innovation (Schultes and Bartel, 2000). Secondary functions are a potential source of innovation and, at the same time, confer robustness to the system by performing, if needed, the primary function of a different molecule. This can be also understood as a case of molecular mimicry, a feature exploited by certain viroids that are disguised as DNA double-stranded molecules to replicate by means of a DNA polymerase.

e. A final mechanism relies on synergistic interactions among functional molecules. The relevance of horizontal gene transfer (HGT) is difficult to overstate when discussing genome construction, as is the role that modularity might have played in the origin of complex functions (Briones et al., 2009). However, a simpler situation, not requiring the stable association of genes or modules, might arise when multiple functions are co-circulating in an open environment. A paradigmatic example of such transient associations might be extant multipartite viruses and virus-satellite associations (Lucía-Sanz and Manrubia, 2017). In the context of the origins of life, such associations might have been much more frequent and loose since the multi-layered, hierarchical architecture later fostered by the appearance of microorganisms' consortia, metazoa, and ecosystems at large, was not yet in place.

1.7. The role of the Theory of Complexity in origins-of-life research

While the application of the tools associated to the Theory of Complexity has been very fruitful in other contexts, their potentiality in unveiling the first steps of life has not been sufficiently assessed. This novel vision requires a cross-disciplinary approach, where theoretical tools taken from the combination of complex network theory and game theory are reinforced with both numerical work and *in vitro* experiments.

Extensive computer work has already been developed in the context of *digital life* (Wilke & Adami, 2002). The goal here was to model how different digital organisms struggle for resources, mutate and adapt, thus mimicking real life. This perspective has been especially useful to cast light on what Stephen Jay Gould called "replaying the tape of life" repeatedly to find out what kind of biosphere would be obtained in each case. Questions like whether life is an inevitable consequence of the chemical interaction among the ingredients of the primordial soup, or whether it is an extremely rare event, can benefit from this approach. Most of the work done, however, has been strictly numerical and based on computer platforms where digital organisms compete for resources such as memory and CPU time.

On the other hand, the combination of complex network theory, game theory and digital life remains vastly unexplored, both from the theoretical and the numerical perspective. The extent by which this combined approach can enlarge our knowledge and comprehension of the origins of life (not only on Earth, but in a generic environment) fulfilling the basic requirements for its appearance, is an open question.

1.8. Clues from synthetic biology

Synthetic biology involves the rational design and synthesis of complex biology-based or bio-inspired systems that display useful functions, even ones that do not exist in nature (Ausländer et al., 2017). Such an innovative approach could allow finding a second example of life, through the chemical synthesis of an artificial cell (Solé 2016). In this regard, the living cell is considered as a highly complex factory equipped with chemical machines that perform a variety of tasks, which can be engineered to originate novel systems with programmable functionality. One of the main goals of synthetic biology is to identify a minimal configuration of a biological cell capable to fulfil its most important tasks: growth, replication, metabolism and energy production. In this line, the question about the origins of life is addressed by an experimental bottom-up approach (complementary to the top-down approach, which modifies pre-existing, current cells) that involves the reverse-engineering of a particular biological function using a minimal set of natural, modified or synthetic molecular components in the absence of cells (Fletcher 2016; Schwille et al., 2018; Yewdall et al., 2018.).

The construction of such minimal or artificial cells using a targeted design will contribute to complete our understanding of how the building blocks of modern cells collectively operate to transition into a living system. Because a

basic requirement for the origin of life is the compartmentalization, considerable efforts have been done to establish cell-sized compartments in the form of controllable droplets and lipid vesicles, using a variety of technologies such as microfluidics (Supramanian et al., 2019). Significant progress has also been made in the bottom-up reconstitution and quantitative understanding of essential cellular machineries (including DNA processing, cell division, self-organized spatial protein patterns, cell-free gene expression systems) in cell-like containers (Schwille et al., 2018; see Challenge 8).

All these rapid advances, together with those in prebiotic systems chemistry and nanotechnology, allow us to tackle the problem of integrating basic individual systems to a (currently unknown) level of complexity where life-like properties can emerge (Schwille, 2017; Stano, 2018). To overcome this complexity barrier, theoretical and computational studies will be essential in guiding the experimental progress. Engineering sciences must also be involved, as they are very successful in dissecting complex systems into a controllable set of functional modules. The application of these approaches to cellular complexity will allow compiling the basic elements that can assemble the essential biochemical functions (Fletcher, 2016). In the field of the origins of life, the main challenge will be to understand how functional modules self-organize in time and space so that autonomous cell-like properties emerge.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

2.1. Basic and philosophical aspects to be addressed

A minimal list of philosophical issues that deserve attention in any research plan on origins of life (Mariscal et al., 2019) include the following: i) the nature of life and the quest for a definition of life; ii) the explanatory power of origins-of-life research, in which the universal (i.e., true for origins everywhere), historical (descriptions of life's origin on Earth), and synthetic (possible ways of originating life) research programs overlap, all with interesting scientific projections; iii) the strategies for such a research, typically thought of as either top-down (inferring from current life back to the last universal common ancestor, LUCA) or bottom-up (starting from non-life and working out how to get life started), which face different epistemological problems and require distinct philosophical commitments; iv) the metabolism-first vs. replication-first debate, which has been challenged by systemic approaches (see below); v) the nature of evolution prior to LUCA, which was certainly

different from contemporary evolution, i.e., the onset of selection; vi) the nature of entities prior to LUCA, which are sometimes thought of as loose communities; vii) the challenges of origins of life which are common to multidisciplinary sciences: competing research programs, diverse standards of evidence, and communicating across disciplinary divisions; viii) the development of new theories or tools, which offer opportunities for new avenues of research, but may also constrain others.

2.2. Looking for the key components of life in interstellar space

To deepen into the connections between the molecular repertoire in interstellar space and the origins of life, a novel approach has been recently initiated, by looking at the problem from the point of view of a (bio)chemist instead of an astrophysicist. It focuses on the search in the ISM of the precursor organic molecules required for the origin of nucleobases and complete ribonucleotides that could have given rise to an RNA world (Powner et al., 2009; Patel et al., 2015; Becker et al., 2019). This strategy has already delivered the first detections in space of key reactants and products within this scheme, such as glycolonitrile (OHCH2CN; Zeng et al., 2019), cyanomethanimine (HNCHCN; Rivilla et al., 2019) and urea (Jimenez-Serra et al., 2020).

However, this is just the starting point. Prebiotic chemical schemes also include other species such as carboxylic acids, fatty acids, fatty alcohols, amino acids, sugars and polyaromatic hydrocarbons (PAHs). A large fraction of these compounds has been detected in meteorites (as e.g. ribose; Furukawa et al., 2019) and in comets (as e.g. glycine; Altwegg et al., 2016). Interestingly, nucleobases were also identified in carbonaceous chondrites (Martins et al., 2008; Callahan et al., 2011). This implies that such relatively large molecules, and/or their precursors, could have formed originally in interstellar space, being later on delivered to Earth by impacts of meteorites and cometary nuclei.

The goal of the ISM group at CAB for the next decades, is thus to populate the *interstellar prebiotic chemistry scheme*, so to determine what organic compounds could be formed already in the ISM. With that aim, the group has developed new astronomical data analysis tools (the MAdrid Data CUBe Analysis package, MADCUBA) and experimental setups (the ultra-high-vacuum chamber ISAC), not only to search for these molecules, but also to try to understand their formation pathways in space (Muñoz-Caro et al., 2002; see C9.6). These efforts will be combined with astrochemical modelling of complex organics in the ISM (Quenard et al., 2018).

To accomplish this goal, spectroscopy plays an essential role. The identification of a certain molecule requires the direct comparison of the frequencies (molecular rotational lines) observed using radio-telescopes, with the experimental ones obtained in the laboratory. Computational methods improve day by day, but they are not able to predict the frequencies of a molecule with the needed accuracy and, therefore, there is no other option than to characterize the molecule in the laboratory. The Spectroscopy Group of the University of the Basque Country (UPV-EHU) and Instituto Biofisika (CSIC/UPV-EHU) has designed and built instrumentation in the centimetric microwave region. The goal is to combine different time and frequency domain spectroscopic tools at centimetre, millimetre and sub-millimetre wavelengths, to provide the precise set of spectroscopic constants that could be used to search for these species in the ISM. This work will be essential for the scientific exploitation of future instrumentation such as the Square Kilometre Array (SKA), in which Spain (and more in particular, the ISM group at CAB), is heavily involved (Cradle of Life Science Working Group).

2.3. The role of mineral surfaces in the origin of life

Both thermodynamic and kinetic reasons sustain the proposal that minerals played a key role in the chemical and protobiological routes to first living organisms. Montmorillonite (Ferris, 2006), as well as many other minerals (including zeolites, sulphides, iron-oxide and silica) have been proposed to speed up the synthesis of molecules required for life as we know it, as well as the polymerization reactions that gave rise to RNA and peptides (Saladino et al., 2019).

Experimental work has been done to support these claims but still misses a good understanding of the structural reasons of why and how mineral surfaces catalyse reactions relevant to prebiotic chemistry and early metabolism (Hazen, 2010). This challenge will require mineralogists, organic chemists, material scientists, biochemists, catalysis chemists and microbiologists, to work synergistically in the following tasks: i) to search shortcuts to chemical routes towards polymerization, compartmentalization, UV protection, and the formation of complex molecules eventually capable to replicate, such as polypeptides, nucleic acids, or other molecules not "invented" by life; ii) to search for autocatalytic chemical processes that could provide energy to trigger reactions showing high kinetic barriers; iii) to understand the demonstrated ability of silica to adsorb organic compounds and to catalyse their reactions. This, together with the exploration of other iron and magnesium minerals, will open new avenues for discovery of the pre- and protobiological history of this planet.

Within this framework, to search for oldest remnants of life is also crucial for setting the younger limit of the lifeless planet and the geochemical conditions triggering the appearance of life and Darwinian evolution (see Challenge 3). Unfortunately, this is by no means an easy task (Javaux, 2019). Since the very first micro-palaeontological studies on the oldest remnants of life, it was obvious the difficulty to rigorously assess biogenicity, i.e. whether a feature has been formed by life or not. Neither isotopic signatures nor organic matter itself are unambiguous biosignature, since it may be synthesized abiotically in many different environments. Putative microfossils and microbially-induced sedimentary structures, such as stromatolites, are neither reliable as they can be the results of abiotic processes. The ability of purely inorganic physical and chemical processes to mimic cellular remnants of early life has been demonstrated experimentally, so that morphology by itself cannot be used as a criterion for biogenicity (García-Ruiz et al., 2003). The problem of primitive life detection here or elsewhere must be approached with an unbiased search, i.e. we should look for mineral bizarre patterns (either morphological, physical or chemical) and then apply a protocol containing criteria for biogenicity and non-biogenicity. Finally, the unexplored subject of bacterial taphonomy should be developed in parallel to the investigation of abiotic mineral self-organization.

2.4. Scientific objectives in the field of prebiotic chemistry

Ongoing research in the field of prebiotic chemistry includes the influence of the physico-chemical environmental conditions on the primary processes that generate complex organic compounds from simple molecules in gas and liquid phases, as well as in ices. Different sources of energy will be studied, including a broad range of temperatures, cold and hot plasmas, UV photolysis, impact shocks and vaporization. Furthermore, the interaction of biomolecules on metallic and mineral surfaces will inform about the stability of adsorbed molecules under different environmental conditions, self-assembling processes and heterogeneous catalysis in connection with the origin of life. This could give clues on the emergence of far-from-equilibrium systems based on the oscillatory Krebs-type cycles, eventually paving the way for a proto-metabolism.

Of particular importance is the elucidation of single handedness of biomonomers (e.g., amino acids and nucleotides) and polymers (e.g., peptides or nucleic acids) and whether symmetry breaking was produced by physical forces (able to cause small, yet significant, enantiomeric imbalances in chiral organics formed in extra-terrestrial media) or took place after polymerization

(Cintas, 2016; Hochberg and Ribó, 2019; Hochberg and Cintas, 2020). Indeed, molecular asymmetry represents a biosignature of life (Glavin et al., 2020) that motivates investigations of abiotic mechanisms and places life in a broader perspective. Thus, there is need to study the enantiomeric selection driven by chiral molecules adsorbed on surfaces and that triggered by crystallization of non-chiral molecules (e.g., sodium chlorate), usually following autocatalytic pathways, which are dominant in our planet owing to the existence of a hydrosphere (Viedma et al., 2015).

A key aspect in this framework is to develop the theoretical aspects of prebiotic chemistry and the origin of molecular chirality. A major focus will be studying autocatalytic reaction systems, chemical self-replicating systems, and chiral symmetry breaking systems, as precursors to the origin of biological homochirality. The goal is to analyse relevant reaction networks, using physics, physical chemistry and numerical simulations to study their dynamic stability properties (perturbation theory), their critical properties (via the dynamic renormalization group), the role of intrinsic diffusion-limited noise (stochastic methods), and their spatial and temporal evolution (numerical simulations). Another promising direction is to exploit astrochemical models for reactions in the ISM considering naturally occurring physical processes, such as chiral photochemistry, that may influence the formation of small enantiomeric excesses in simple chiral organic molecules found on dust grains.

2.5. Deeping into the role of RNA in the origin and early evolution of life

There are two main schools of thought regarding the origin of RNA: RNA is either the product of sequential prebiotic reactions that gave rise to ribonucleotides (Sutherland, 2016) or the outcome of the chemical/pre-biological evolution from other genetic polymers (Hud et al., 2013; Higgs and Lehman, 2015). Although key advances have been produced in this field (Powner et al., 2009; Patel et al., 2015; Becker et al., 2019), the demonstration of a direct prebiotic synthesis of RNA is facing numerous challenges, which reinforces the hypothesis that the availability of heritable information may have started with an ancestral proto-RNA polymer. Finding a plausible ancestor of RNA that might have resulted from molecular self-assembly would conciliate the prebiotic chemistry and the geochemistry found in planetary environments (Ruiz-Mirazo et al., 2014). One possibility relies on the introduction of non-canonical bases that fit well in a nucleic acid structure (such as barbituric acid,

a prebiotic analogue of uracil, Menor-Salván et al., 2009; Menor-Salván and Marin-Yaseli, 2013) and form nucleosides spontaneously, overcoming the challenges of the direct prebiotic synthesis of canonical nucleosides.

When discussing about the origins of life, modern authors very often reference the famous letter which Charles R. Darwin sent to Joseph D. Hooker in 1871, in which he mused about a "warm little pond, with all sorts of ammonia and phosphoric salts... that a protein compound was chemically formed ready to undergo still more complex changes". Using the physical and chemical properties of urea, a key molecule in prebiotic chemistry, the warm little pond model combines into a single scenario the geochemistry of the Archean/Hadean Earth, the unique chemistry in the atmosphere in this era, and the chemistry of phosphates. The problematic nature of phosphates, the predominant form of phosphorus, stems from their extremely low solubility and their inability to readily form organophosphates. This makes it difficult to propose a meaningful geochemical model and gives rise to the phosphorylation problem in the origins of life field. Prebiotic cyanide could be the precursor of key prebiotic chemicals, such as urea and ammonium formate, and can directly promote the synthesis of nucleotides by mobilization of phosphate from minerals. Hence, if we assume the presence of cyanide, urea and phosphate minerals on Early Earth, it could be possible that prebiotic chemistry and minerals established a system in which the formation of nucleobases, nucleosides and prebiotic phosphorylation could have been readily achieved in a simple scenario consistent with geochemistry and planetary geology (Burcar et al., 2016; Burcar et al., 2019; Burcar and Menor-Salván, 2020).

Additionally, to gain insight into the knowledge of the RNA role in the origin of life would require a combination of biochemical, biophysical, mutational and phylogenetic analyses, together with the use of novel computational tools, in the context of different biological models (Briones et al., 2009; Atkins et al., 2011; Romero-López and Berzal-Herranz, 2013; Takahashi et al., 2016; Moreno et al., 2019b). Understanding the function of the plethora of current ncRNAs should be addressed to trace some of the key biological processes of life. Moreover, knowing the functional diversity of RNA and their mechanisms of action, as well as the architecture of its interactome (i.e. the relationship between the structure and the biological function in viral and cellular RNAs), should enlighten the role of this biopolymer in the origin and early evolution of life (Atkins et al., 2011; Ariza-Mateos et al., 2019). The additional development of new molecular tools by *in vitro* selection strategies would also help to elucidate the function of

structural elements contained in cellular RNA molecules.

A quantitative assessment of the issues related to the emergence of complex function and molecular innovation in a pre-cellular world, will require substantial trans-disciplinary research. Mathematical analyses of genotype-to-phenotype-to-function map, new computational approaches that permit an unbiased characterization of sequence spaces, and experiments tailored to respond specific questions have to be developed (Manrubia et al., 2020; Challenge 5).

2.6. The application of Theory of Complexity in the field of origin of life

This novel line of research aims at analysing life as a process that emerges spontaneously from the complex interaction (competition/cooperation) among networked systems that represent the biochemical structures existing in early Earth. This research will benefit from the current understanding of how networked systems are organized, interact and evolve obtained by the application for more than a decade of complex network theory to real systems (Boccaletti et al., 2006). Indeed in many cases, a network has been shown to be a network of networks, and although the detection of modules within a network has been thoroughly studied, the influence of the networks' interconnections on their dynamical processes, still represents a challenge. For instance, these interactions are crucial in processes such as spread of diseases, knowledge and even rumours (Aguirre et al., 2013). The interaction between the building bricks of life is not an exception. Thus, developing a theory that fully describes the competition and cooperation between biological and/or chemical networks, making use of a synergy between network and game theory (Iranzo et al., 2016) – both in equilibrium and out of the equilibrium (Buldú et al., 2019) – becomes a necessary requirement to understand the first steps of life. In addition, each one of the many biochemical networks already described in the literature (protein, genetic, metabolic, or neutral genotype networks, among others), has peculiarities and it seems difficult – if not impossible – to extract general behaviours and properties. A thorough analysis of which of these networks might be applicable (and how) to describe the transition from chemistry to biology, or the necessity of developing *ad hoc* networks to better describe the idiosyncrasy of the prebiotic world, will be also of critical importance to achieve the goal.

2.7. Challenges in the field of prebiotic systems chemistry

There are currently various lines of experimental and computational work,

which focus on the construction of proto-cellular assemblies through the integration of the three basic subsystems of cellular life through different kinds of physicochemical processes. Research on self-replication, autocatalytic networks and self-reproducing compartments are the most important ones, but they all face several inherent problems and limitations. First, most of the work has been performed with molecular components taken from existing living organisms (e.g., phospholipids, peptides, oligonucleotides, etc.), assuming they would have been available on the prebiotic Earth. This is a useful approach to study functional models of the first protocells, yet it is highly improbable that those biomolecules could have been produced spontaneously through non-evolutionary processes (condensation reactions, random amino acid or nucleotide polymerizations, etc.), in sufficient quantities and with adequate structure/sequence to exert their role. Moreover, a strong limitation lies in the difficulty of integrating the complex dynamic behaviours of each separate subsystem (de la Escosura, 2019). In order to overcome such limitations, many scientists have realized that unravelling the main pathways and mechanisms for chemical evolution will require investigating molecular assemblies that are able to regulate the production of their own ingredients from the simplest building blocks (e.g., fatty acids, simple sugars, amino acids, nucleobases, etc.). This implies great challenges in both synthetic and supramolecular studies dealing with organic and inorganic reaction networks, as well as in analytical chemistry, due to the extremely complex messy mixtures involving hundreds of dynamic reactions operating in parallel. Recent, significant steps have been made in the challenge to connect synthetic organic chemistry and the biochemistry related to 'core metabolic pathways' (Keller et al., 2016; Coggins and Powner 2017; Muchowska et al., 2017; 2019; Springsteen et al., 2018) involving a remarkable diversity of species and in the absence of enzymatic catalysts. However, the current standard molecule-collecting analytics (nuclear magnetic resonance, chromatography with detection, mass spectrometry, etc) do not seem sufficient to address such complexity, and they will have to be complemented with high-throughput screening methods, adapted from the biological "omic" disciplines (genomics, proteomics, metabolomics, etc) and computational modelling to unravel emergent behaviours within those mixtures (Ruiz-Mirazo et al., 2014).

A second key aspect is evolvability. This relates to the capacity of living systems to process heritable information, coded in DNA and used to instruct how the cell works. The emerging overall picture shows that information processing in cells occurs through a hierarchy of genes regulating the activity of other genes

through a complex interactome and metabolic networks. There is an implicit semiotic character in this way of dealing with information, based on functional molecules that act as signs to achieve self-regulation of the whole network (de la Escosura, 2019; Ariza-Mateos et al., 2019). In contrast to cells, chemical systems are not capable to process information. Yet, they must have preceded biological organisms, and evolved into them. Hence, there must have been prebiotic molecular assemblies that could somehow process information, in order to regulate their own constituent reactions and supramolecular organization processes. A great challenge for the near future is thus to investigate if there are universal principles pervading the way information processing determines an ever-increasing dynamic complexity of the material world, particularly in (and possibly towards) its living manifestations. This perspective also raises the question of whether there is a causal relationship between information processing and evolution, not only in living organisms but also in abiotic stages of chemical evolution (de la Escosura et al., 2015; Ruiz-Mirazo et al., 2017). Furthermore, considering that protocell populations should be considered the proper units of prebiotic evolution, three experimental challenges aimed at constructing protocell systems should be considered: coupling chemical reaction networks with vesicle dynamics, finding conditions and mechanisms for minimal functional integration, and characterizing the evolutionary dynamics of pre-Darwinian protocells (Shirt-Ediss et al., 2017).

2.8. Towards the first protocells

Prebiotic systems chemistry is superseding the classical controversies around the origin of life based on dichotomies such as the early vs. late emergence of compartments, metabolism-first vs. genetics-first models, or autotrophic vs. heterotrophic scenarios (Peretó, 2005; Ruiz-Mirazo et al., 2014). Experimental results in prebiotic systems chemistry suggest that proto-metabolism on Archaean Earth generated a small set of biomonomers that prefigured extant biochemistry (Peretó, 2019; see Challeng 5 and Challenge 8).

There are, nevertheless, relevant challenges ahead when trying to implement experimental systems with increasing levels of complexity. 1) We need to know more about the dynamic coupling of lipid vesicles with heterogeneous chemical mixtures of small molecules and polymers (Murillo-Sánchez et al., 2016), as well as their experimental evolution into more robust systems that simulate protocells with pre-Darwinian or Darwinian behaviour. This would include the exploration of evolutionary processes in the context of protocell population/selection experiments (Budin and Szostak, 2011; Adamala and

Szostak 2013) or even the emergence of the first proto-ecological or syntrophic relationships among protocells, as it is recently argued in (Ruiz-Mirazo et al., 2020). 2) The merging of systems chemistry with experimental evolution methodologies is an absolute must if we are to bridge the gap between protometabolism and functionally integrated protocells (Ruiz- Mirazo et al., 2017; Shirt-Ediss et al., 2017). In short, solutions hardly accessible by rational design can be found by the own system's dynamics. Therefore, merging laboratory, computational, robotics, microfluidics and synthetic biology technologies is compulsory to explore in depth the transition from simple chemical mixtures to systems with complex behaviours (Scharf et al., 2015). 3) Certain aspects, however, still require significant and persistent research efforts, including finding the paths to the emergence of autocatalytic networks of reactions exhibiting dynamics of prebiotic relevance.

3. KEY CHALLENGING POINTS

The origin of life is a multidisciplinary field involving physics, chemistry, biology, and geology, with relevant implications for social sciences and humanities. The approach proposed here combines the knowledge of different groups at CSIC. This should position our institution not only as a national leader in the field, but also will place it in a privileged situation to compete and collaborate with other leading international institutions. Within this framework, the key points to be addressed are the following:

- Search in the ISM of the precursor organic molecules required for the origin of life and to try to understand their formation pathways in interstellar space.
- Analysis of mineral patterns (morphological, physical or chemical) present in some of the oldest remnants of life and application of a protocol containing criteria for biogenicity and non-biogenicity.
- Investigation of chemical routes towards the synthesis of biomonomers, polymerization of peptides and nucleic acids, as well as molecular self-assembly on mineral or metal surfaces.
- Design and implementation of prebiotic chemistry experiments that generate complex organic compounds from simple molecules in gas and liquid phases, as well as in ices, using sources of energy such as different temperatures, cold and hot plasmas, UV photolysis, impact shocks and vaporization.
- Study of the enantiomeric selection driven by chiral molecules adsorbed

on surfaces, as well as that triggered by crystallization of non-chiral molecules.

- Development of theoretical aspects of prebiotic chemistry and the origin of molecular homochirality.
- Investigation of how the phosphate and other plausible linker groups were incorporated to chemical evolution, in particular to RNA.
- To deepen into the sequence-structure-function-evolution relationships in RNA molecules and populations.
- Experimental analysis of the structure and function of ncRNAs as well as their interaction patterns in selected viral and cellular RNAs.
- To perform in vitro evolution experiments using complex molecular mixtures, within the framework of prebiotic systems chemistry.
- Investigation of the self-assembly of complex molecules and vesicle formation in different experimental settings, as well as the coupling of chemical reaction networks with vesicle dynamics.
- Integration of experimental, computational and synthetic biology technologies to explore the transition from simple chemical mixtures to systems endowed with complex behaviours.
- To identify the universal principles that allow information processing to determine the increase in dynamic complexity of the living world.
- Use of new computational approaches that permit an unbiased characterization of sequence spaces and the analysis of genotype-to-phenotype map.
- Development of a theory to describe the competition and cooperation between chemical and/or biological networks, making use of network and game theories, both in equilibrium and out of the equilibrium.

CHALLENGE 1 REFERENCES

Adamala, K., Szostak, J.W. (2013). Competition between model protocells driven by an encapsulated catalyst. *Nat. Chem. 5*, 495–501.

Aguirre, J., Papo, D., Buldú, J.M. (2013). Successful strategies for competing networks. *Nat. Phys. 9*, 230–234.

Alonso, E.R., Kolesniková, L., Tercero, B. et al. (2016). Millimeter Wave Spectrum and Astronomical Search for Vinyl Formate. *The Astrophysical Journal 832*, 42.

Alonso, E.R., McGuire, B.A., Kolesniková, L. et al. (2019). The Laboratory Millimeter and Submillimeter Rotational Spectrum of Lactaldehyde and an Astronomical Search in Sgr B2(N), Orion-KL, and NGC 6334I. *The Astrophysical Journal 883*, 18.

Altwegg, K., Balsiger, H., Bar-Nun, A. et al. (2016). Prebiotic chemicals-amino acid and phosphorus-in the coma of comet 67P/Churyumov-Gerasimenko. *Science Adv. 2*, 5.

Ariza-Mateos, A., Briones, C., Perales, C., Domingo, E., Gómez, J. (2019). The archaeology of coding RNA. *Annals of the New York Academy of Sciences 1447*, 119–134.

Atkins, J.F., Gesteland, R.F., Cech, T.R., (Ed). (2011). *RNA worlds: from life's origins to diversity in gene regulation*. CSHL Press, NY, US.

Ausländer, S., Ausländer, D., Fussenegger, M. (2017). Synthetic biology- the synthesis of biology. *Angew. Chem. Int. Ed. Engl. 56*, 6396–6419.

Becker, S., Feldmann, J., Wiedemann, S. et al. (2019). Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science 366*, 76–82.

Belloche, A., Menten, K.M., Comito, C. et al. (2008). Detection of amino acetonitrile in Sgr B2(N). *Astronomy & Astrophysics 482*, 179B.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U. (2006). Complex networks: structure and dynamics. *Phys. Rep. 424*, 175–308.

Breslow, R., Cheng, Z-L. (2009). On the origin of terrestrian homochirality for nucleosides and amino acids, *Proc. Natl. Acad. Sci. USA 106*, 9144–9146.

Briones, C. (2015). Aptamer. In *Encyclopedia of Astrobiology*, 2nd Ed. M. Gargaud and W.M. Irvine, Eds., Springer, Heidelberg. Pp. 112–113.

Briones, C., Stich, M., Manrubia, S.C. (2009). The dawn of the RNA world: Towards functional complexity through ligation of random RNA oligomers. *RNA 15*, 743–749.

Budin, I., Szostak, J.W. (2011). Physical effects underlying the transition from primitive to modern cell membranes. *Proc. Natl. Acad. Sci. USA 108*, 5249–5254.

Buldú, J.M., Pablo-Martí, F., Aguirre, J. (2019). Taming out-of-equilibrium dynamics on interconnected networks. *Nat. Commun. 10*, 5314.

Burcar, B., Castañeda, A., Lago, J. et al. (2019). A Stark Contrast to Modern Earth: Phosphate Mineral Transformation and Nucleoside Phosphorylation in an Iron- and Cyanide-Rich Early Earth Scenario. *Angew. Chemie Int. Ed. 58*, 16981–16987.

Burcar, B., Menor-Salván, C. (2020). The multiple roles of urea in chemical evolution. *Chem. Rev*., in press.

Burcar, B., Pasek, M., Gull, M. et al. (2016). Darwin's Warm Little Pond: A One-Pot Reaction for Prebiotic Phosphorylation and the Mobilization of Phosphate from Minerals in a Urea-Based Solvent. *Angew Chemie Int. Ed. 55*, 13249–13253.

Calabrese, C., Uriarte, I., Insausti, A. et al. (2020). *Observation of the Unbiased Conformers of Putative DNA-Scaffold Ribosugars*. ACS Central Science 0.

Callahan, M.P., Smith, K.E., Cleaves, H.J. et al (2011). Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc. Natl. Acad. Sci. 108*, 13995–13998.

Catalán, P., Elena, S.F., Cuesta, J.A., Manrubia, S. (2019). Parsimonious scenario for the emergence of viroid-like replicons de novo. *Viruses 11*, 425.

Cintas, P. (2016). Homochirogenesis and the emergence of life-like structures. In *Chirality in Supramolecular Assemblies-Causes and Consequences* (Ed.: R. Keene), John Wiley & Sons, NY, Ch. 2, pp. 44–64.

Cocinero, E.J., Lesarri, A., Écija, P. et al. (2012). Ribose Found in the Gas Phase. *Angew. Chem. Int. Ed. 51*, 3119.

Coggins, A.J., Powner, M.W. (2017). Prebiotic synthesis of phosphoenol pyruvate by alpha-phosphorylation-controlled triose glycolysis. *Nat. Chem. 9,* 310–317.

Cologne Database for Molecular Spectroscopy (CDMS): https://cdms.astro.uni-koeln.de/classic/molecules

De la Escosura, A. (2019). The informational substrate of chemical evolution: implications for abiogenesis. *Life 9,* 66.

De la Escosura, A., Briones, C., Ruiz-Mirazo, K. (2015). The systems perspective at the crossroad between chemistry and biology. *J. Theor. Biol. 381,* 11–22.

Eschenmoser, A. (2007). The search for the chemistry of life's origins. *Tetrahedron 63,* 12821–12844.

Ferris, J.P. (2006). Montmorillonite-catalysed formation of RNA oligomers: the possible role of catalysis in the origins of life. *Phil. Trans. R. Soc. B 361,* 1777–1786.

Fletcher, D.A. (2016). Bottom-up biology: Harnessing engineering to understand nature. *Dev. Cell. 38,* 587–589

Furukawa, Y., Chikaraishi, Y., Ohkouchi, N. et al. (2019). Extraterrestrial ribose and other sugars in primitive meteorites. *PNAS 116,* 49

Galvez-Martinez, S., Escamilla-Roa, E., Zorzano, M.P., Mateo-Marti, E. (2019). Defects on a pyrite(100) surface produce chemical evolution of glycine under inert conditions: experimental and theoretical approaches. *Phys. Chem. Chem. Phys. 21,* 24535.

Garcia-Ruiz, J.M., Melero-Garcia, E., Hyde, S.T. (2009). Morphogenesis of self-assembled nanocrystalline materials of barium carbonate and silica. *Science 323,* 362-365.

García-Ruiz, J.M., Van Zuilen, M.A., Bach, W. (2020). Mineral self-organization on a lifeless planet. *Physics of Life Reviews,* in press.

Glavin, D.P., Burton, A.S., Elsila, J.E., Aponte, J.C., Dworkin, J.P. (2020). The search for chiral asymmetry as a potential biosignature in our solar system. *Chem. Rev.*, in press.

Haykal, I., Margulès, L., Huet, T.R. et al. (2013). The CM-, MM-, and SUB-MM-WAVE spectrum of allyl isocyanide and radioastronomical observations in Orion KL and the SgrB2 line surveys. *The Astrophysical Journal 777,* 120.

Hazen, R.M., Sverjensky, D.A. (2010). Mineral surfaces, geochemical complexities, and the origins of life. *Cold Spring Harb Perspect Biol. 2,* a002162.

Higgs, P.G., Lehman, N. (2015). The RNA World: Molecular Cooperation at the Origins of Life. *Nat. Rev. Genet. 16,* 7–17.

Hochberg, D., Cintas, P. (2020). Does pressure break mirror-image symmetry? A perspective and new insights. *Chem. Phys. Chem. 21,* 633–642.

Hochberg, D., Ribó, J.M. (2019). Entropic Analysis of Mirror Symmetry Breaking in Chiral Hypercycles. *Life 9,* 28.

Huber, C., Wachtershauser, G. (1998). Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: implications for the origin of life. *Science 281,* 670–672.

Hud, N.V., Cafferty, B.J., Krishnamurthy, R., Williams, L.D. (2013). The Origin of RNA and "My Grandfather's Axe". *Chem. Biol. 20,* 466-474.

Iranzo, J., Buldú, J.M., Aguirre, J. (2016). Competition among networks highlights the power of the weak. *Nat. Commun. 7,* 13273.

Javaux, E. (2019). Challenges in evidencing the earliest traces of life. *Nature 572,* 451–460.

Jimenez-Serra, I., Martin-Pintado, J., Rivilla, V.M. et al. (2020). CAB: Towards the RNA-world in the interstellar medium – detection of urea, and search of 2-amino-oxazole and simple sugars. *Astrobiology 20,* 7.

Jorgensen, J., van der Wiel, M., Coutens, A. et al. (2012). Detection of the Simplest Sugar, Glycolaldehyde, in a Solar-type Protostar with ALMA. *Astronomy & Astrophysics 595,* A117

Joyce, G.F., Szostak, J.W. (2018). Protocells and RNA Self-Replication. *Cold Spring Harb. Perspect. Biol. 10,* a034801.

Joyce, G.F. (2004). Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem. 73,* 791–836.

Keller, M.A., Zylstra, A., Castro, C., Turchyn, A.V., Griffin, J.L., Ralser, M. (2016). Conditional iron and pH-dependent activity of a non-enzymatic glycolysis and pentose phosphate pathway. *Sci. Adv. 2,* e1501235.

Krishnamurthy, R. (2017). Giving rise to Life: Transition from Prebiotic Chemistry to Protobiology. *Acc. Chem. Res. 50,* 455–459.

Lavado, N., García de la Concepción, J., Babiano, R., Cintas, P. (2018). Formation of cyanamide-glyoxal oligomers in aqueous environments relevant to primeval and astrochemical scenarios: a spectroscopic and theoretical study. *Chem. Eur. J. 24*, 4069–4085.

Lucía-Sanz, A., Manrubia, S. (2017). Multipartite viruses: adaptive trick or evolutionary treat? *Systems Biology and Applications 3*, 34.

Mann, S. (2013). The Origins of Life: Old Problems, New Chemistries. *Angew. Chem. Int. Ed. 52*, 155–162.

Manrubia, S., Cuesta, J.A., Aguirre, J. et al. (2020). From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Journal of the Royal Society Interface*, submitted.

Marín-Yaseli, M.R., González-Toril, E., Mompeán, C., Ruiz-Bermejo, M. (2016). The role of the aqueos aerosols in the "Glyoxylate Scenario": An experimental approach. *Chem. Eur. J. 22*, 12785–12799.

Mariscal, C., Barahona, A., Aubert-Kato, N. et al. (2019). Hidden Concepts in the History and Philosophy of Origins-of-Life Studies: a Workshop Report. *Orig. Life Evol. Biosph. 49*, 111–145.

Martin-Gago, J.A. (2011). Polycyclic aromatics: On-surface molecular engineering. *Nature Chemistry 3*, 11–12.

Martins, Z., Botta, O., Fogel, M.L. (2008). Extraterrestrial nucleobases in the Murchison meteorite. *Earth Planet. Sci. Lett. 270*, 130–136.

Mateo-Marti, E., Galvez-Martinez, S., Gil-Lozano, C., Zorzano, M.P. (2019). Pyrite-induced UV-photocatalytic abiotic nitrogen fixation: implications for early atmospheres and Life. *Scientific Reports 9*, 15311.

Méndez, J., López, M., Martín-Gago, J. (2011). On-surface synthesis of cyclic organic molecules. *Chemical Society Reviews 40*, 4578.

Menor-Salván, C., Marín-Yaseli, M.R. (2012). Prebiotic Chemistry in Eutectic Solutions at the Water–Ice Matrix. *Chem. Soc. Rev. 41*, 5404.

Menor-Salván, C., Ruiz-Bermejo, M., Guzmán, M.I., Osuna-Esteban, S., Veintemillas-Verdaguer, S. (2009). Synthesis of Pyrimidines and Triazines in Ice: Implications for the Prebiotic Chemistry of Nucleobases. *Chem. - A Eur. J. 15*, 4411–4418.

Mompeán, C., Marín-Yaseli, M.R., Espigares, P., González-Toril, E., Zorzano, M.P., Ruiz-Bermejo, M. (2019). Prebiotic chemistry in neutral/reduced-alkaline gas-liquid interfaces. *Scientific Reports 9*, 1916.

Morales-Reina, S., Giri, C., Leclercq, M. et al. (2020). Programmed recognition between complementary dinucleolipids to control the self-assembly of lipidic amphiphiles. *Chem. Eur. J. 26*, 1082–1090.

Moreno, M., Fernández-Algar, M., Fernández-Chamorro, J., Ramajo, J., Martínez-Salas, E., Briones, C. (2019a). A combined ELONA-(RT) qPCR approach for characterizing DNA and RNA aptamers selected against PCBP-2. *Molecules 24*, 1213.

Moreno, M., Vázquez, L., López-Carrasco, M.A., Martín-Gago, J.A., Flores, R., Briones, C. (2019b). Direct visualization of the native structure of viroid RNA at single-molecule resolution by atomic force microscopy. *RNA Biology 16*, 295–308.

Muchowska, K.B., Varma, S.J., Chevallot-Beroux, E., Lethuillier-Karl, L., Li, G., Moran, J. (2017). Metals promote sequences of the reverse Krebs cycle. *Nat. Ecol. Evol. 1*, 1716–1721.

Muchowska, K.B., Varma, S.J., Moran, J. (2019). Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature 569*, 104–107.

Muñoz-Caro, G., Meierhenrich, U.J., Schutte, W.A. et al. (2002). Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature 416*, 403.

Murillo-Sánchez, S., Beaufils, D., González-Mañas, J.M., Pascal, R., Ruiz-Mirazo, K. (2016). Fatty acids' double role in the prebiotic formation of a hydrophobic dipeptide. *Chem. Sci. 7*, 3406–3414.

Oro, J., Kimball, A.P. (1961). Synthesis of Purines under Possible Primitive Earth I. Adenine from Hydrogen Cyanide. *Arch. Biochem. Biophys. 94*, 217–227.

Patel, B.H., Percivalle, C., Ritson, D.J. et al. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chemistry 7*, 301.

Peña, I., Cocinero, E.J., Cabezas, C. et al. (2013). Six Pyranoside Forms of Free 2-Deoxy-D -Ribose. *Angewandte Chemie - International Edition 52*, 11840–11845.

Peretó, J. (2005). Controversies on the origin of life. *Int. Microbiol. 8*, 23–31.

Peretó, J. (2019). Prebiotic chemistry that led to life. In: *Handbook of Astrobiology* (V Kolb ed.) CRC Press.

Powner, M.W., Gerland, B., Sutherland, J.D. (2009). Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature 459*, 239–242.

Pross, A. (2016). *What is Life? How Chemistry Becomes Biology*, Oxford University Press: Oxford.

Puerta-Fernández, C., Romero-López, C., Barroso-delJesus, A., Berzal-Herranz, A. (2003). Ribozymes: recent advances in the development of RNA tools. *FEMS Microbiology Reviews 27*, 75-97.

Ribó, J.M., Crusats, J., El-Hachemi, Z., Moyano, A., Hochberg, D. (2017). Spontaneous mirror symmetry breaking in heterocatalytically coupled enantioselective replicators. *Chem. Sci. 8*, 763-769.

Rivilla, V.M., Martin-Pintado, J., Jimenez-Serra, I. et al. (2019). Abundant Z-cyano-methanimine in the interstellar medium: paving the way to the synthesis of adenine. *Monthly Notices of the Royal Astronomical Society 483*, L114-L119.

Romero-López, C., Berzal-Herranz, A. (2013). Unmasking the information encoded as structural motifs of viral RNA genomes: a potential antiviral target. *Rev. Med. Virol. 23*, 340–354.

Romero-López, C., Berzal-Herranz, A. (2020). The role of the RNA-RNA interactome in the hepatitis C virus life cycle. *Int. J. Mol. Sci. 21*, 1479.

Ruiz-Bermejo, M., de la Fuente, J.L., Carretero-González, J., García-Fernández, L., Aguilar, M.R. (2019). A comparative study on HCN polymers synthesized by polymerization of NH_4CN or diaminomaleonitrile in aqueous media: new perspectives for prebiotic chemistry and materials science. *Chem. Eur. J. 25*, 11437–11455.

Ruiz-Bermejo, M., de la Fuente, J.L., Carretero-González, J., García-Fernández, L., Aguilar, M.R. (2019). A comparative study of HCN polymers synthesized from NH4CN or DAMN polymerization in aqueous media: New perspectives for prebiotic chemistry and material science. *Chem. Eur. J. 25*, 11437–11455.

Ruiz-Mirazo, K., Briones, C., de la Escosura, A. (2014). Prebiotic systems chemistry: new perspectives for the origins of life. *Chem. Rev. 114*, 285–366.

Ruiz-Mirazo, K., Briones, C., de la Escosura, A. (2017). Chemical roots of biological evolution: the origins of life as a process of development of autonomous functional systems. *Open Biology 7*, 170050.

Ruiz-Mirazo, K., Shirt-Ediss, B., Escribano-Cabeza, M., Moreno, A. (2020). The construction of biological 'inter-identity' as the outcome of a complex process of protocell development in prebiotic evolution. *Frontiers in Physiology (Systems Biology),* in press.

Saladino, R., Di Mauro, E., García-Ruiz, J.M. (2019). A Universal Geochemical Scenario for Formamide Condensation and Prebiotic Chemistry. *Chemistry–A European Journal. 25*, 3181–3189.

Sánchez-Arenillas, M., Mateo-Martí, E. (2016). Pyrite surface environment drives molecular adsorption: cystine on pyrite(100) investigated by X-ray photoemission spectroscopy and low energy electron diffraction. *Phys. Chem. Chem. Phys.* 18, 27219.

Sánchez-Luque, F.J., Stich, M., Manrubia, S., Briones, C., Berzal-Herranz, A. (2014). Efficient HIV-1 inhibition by a 16 nt-long RNA aptamer designed by combining in vitro selection and in silico optimisation strategies. *Scientific Reports 4*, 6242.

Schaper, S., Louis, A.A. (2014). The Arrival of the Frequent: How Bias in Genotype-Phenotype Maps Can Steer Populations to Local Optima. *PLoS ONE 9*, e86635.

Scharf, C., Virgo, N., Cleaves, H.J. 2nd et al. (2015). A Strategy for Origins of Life Research. *Astrobiology 15*, 1031–1042.

Schultes, E.A., Bartel, D.P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science 289*, 448–452.

Schwille, P. (2017). How simple could life be? *Angew Chem. Int. Ed. Engl. 56*, 10998–11002.

Schwille, P., Spatz, J., Landfester, K. (2018). MaxSynBio: Avenues towards creating cells from the bottom up. *Angew Chem. Int. Ed. Engl. 57*, 13382–13392.

Shirt-Ediss, B., Murillo-Sánchez, S., Ruiz-Mirazo, S. (2017). Framing major prebiotic transitions as stages of protocell development: three challenges for origins-of-life research. *Beilstein J. Org. Chem. 13*, 1388–1395.

Solé, R. (2016). Synthetic transitions: Towards a new synthesis. *Phil. Trans. R. Soc. B: Biol. Sci. 371*, 20150438.

Springsteen, G., Yerabolu, J.R., Nelson, J., Rhea, C.J., Krishnamurthy, R. (2018). Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun. 9*, 1–8.

Stano, P. (2018). Is research on "synthetic cells" moving to the next level? *Life (Basel) 9*(1), 3.

Supramaniam, P., Ces, O., Salehi-Reyhani, A. (2019). Microfluidics for artificial fife: Techniques for bottom-up synthetic biology. *Micromachines (Basel) 10*, 299.

Sutherland, J.D. (2016). The Origin of Life - Out of the Blue. *Angew. Chemie Int. Ed. 55*, 104–121.

Takahashi, M.K., Watters, K.E., Gasper, P.M., Abbott, T.R., Carlson, P.D., Chen, A.A., Lucks, J.B. (2016). Using in-cell SHAPE-Seq and simulations to probe structure-function design principles of RNA transcriptional regulators. *RNA 22*, 920–933.

Viedma, C., Coquerel, G., Cintas, P. (2015). Crystallization of chiral molecules. In *Handbook of Crystal Growth* (Ed.: T. Nishinaga), Elsevier, Amsterdam, Vol. 1B, Ch. 22, pp. 952–1002.

Wilde, S.A., Valley, J.W., Peck, W.H., Graham, C.M. (2001). Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature 409*, 175–178.

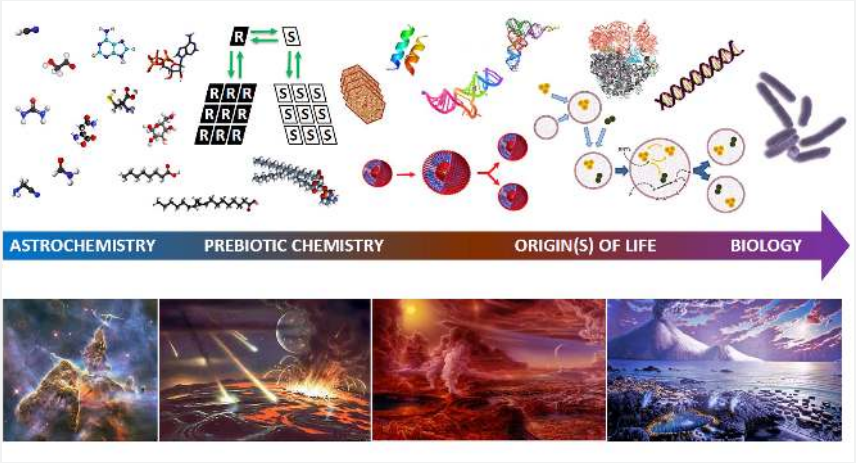
Wilke, C.O., Adami, C. (2002). The biology of digital organisms. *Trends in Ecology and Evolution 17*, 528–532.

Witzany, G. (2020). What is Life? *Front. in Astron. Space Sci. 7*, 7.

Yewdall, N.A., Mason, A.F., van Hest, J.C.M. (2018). The hallmarks of living systems: towards creating artificial cells. *Interface Focus. 8*, 20180023.

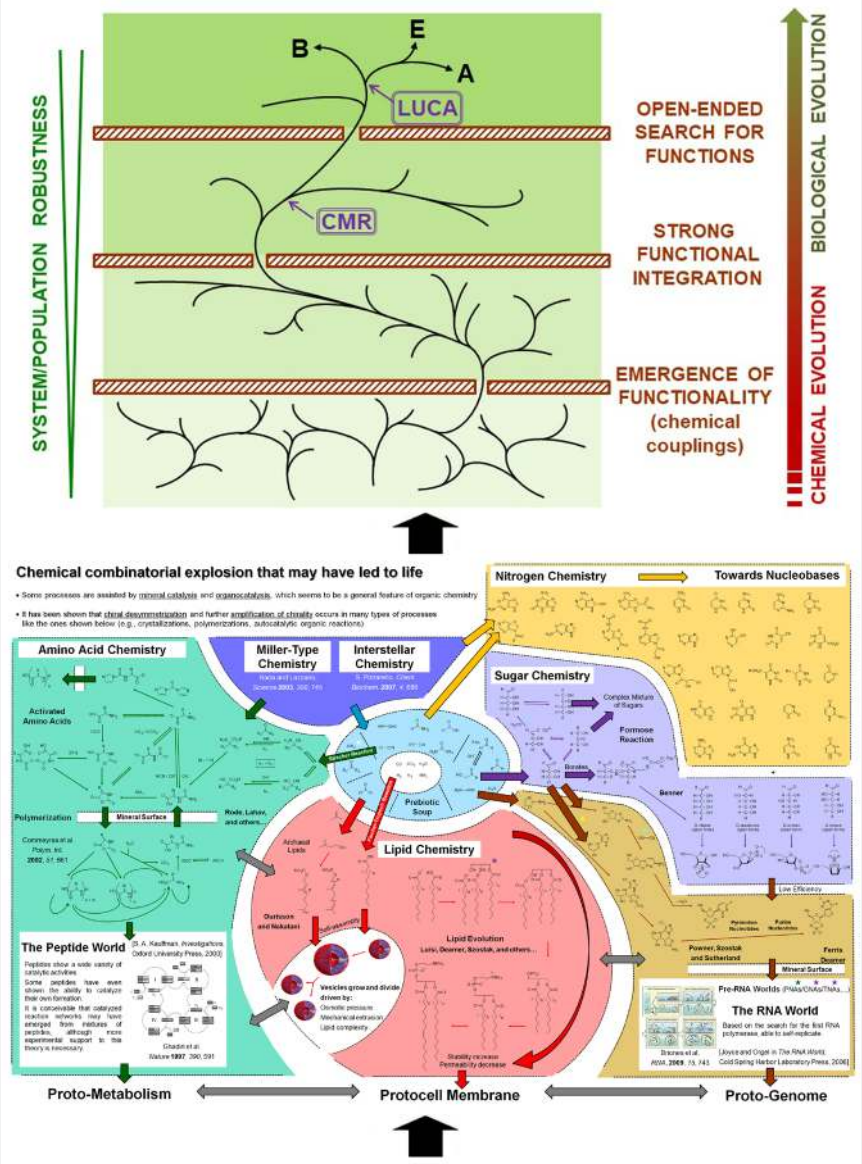
Zeng, S., Quenard, D., Jimenez-Serra, I. et al. (2019). First detection of the pre-biotic molecule glycolonitrile ($HOCH_2CN$) in the interstellar medium. *Monthly Notices of the Royal Astronomical Society 484*, L43Z.

SUMMARY FOR THE GENERAL PUBLIC



A simplified view of the path from astrochemistry to biology, showing some examples of the molecules and aggregates involved (adapted from Ruiz-Mirazo et al., 2014). The bottom panels depict putative scenarios for astrochemistry, prebiotic chemistry on early Earth and the origin(s) of life.

SUMMARY FOR EXPERTS



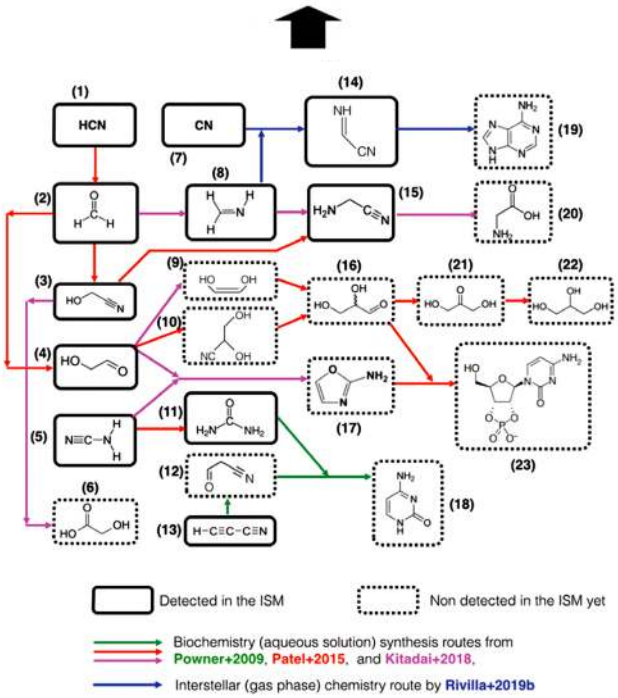


FIG. 1. Summary of the chemical scheme of the RNA-world scenario. Extracted from Powner *et al.* (2009), Patel *et al.* (2015), Kitadai and Maruyama (2018), and Rivilla *et al.* (2019b; see the different colors for the arrows). Solid-line boxes indicate molecules that have been detected in space, while dotted-line boxes denote those species that remain undetected in space. The goal of this diagram is to visualize at once how many primordial RNA-world prebiotic species have been detected in the ISM, and how they are related to each other. The names of the molecules shown are as follows: (1) hydrogen cyanide; (2) formaldehyde; (3) glycolonitrile; (4) glycolaldehyde; (5) cyanamide; (6) glycolic acid; (7) cyanide; (8) methenamine; (9) enol form of glycolaldehyde; (10) cyanohydrin; (11) urea; (12) 3-oxopropanenitrile; (13) cyanoacetylene; (14) cyanomethanimine; (15) aminoacetonitrile; (16) glyceraldehyde; (17) 2-amino-oxazole; (18) cytosine; (19) adenine; (20) glycine; (21) dihydroxyacetone (DHA); (22) glycerol; (23) beta-ribocytidine-2′,3′-cyclic phosphate (pyrimidine ribonucleotide). HCN, hydrogen cyanide; ISM, interstellar medium. Color images are available online.

Schematic representation of the relationship between the scientific fields of astrochemistry (bottom panel: Jimenez-Serra et al., 2020) and prebiotic chemistry (middle panel: de la Escosura, Briones and Ruiz-Mirazo, unpublished), showing examples of the molecules and processes involved. The top panel depicts the transition from chemical to biological evolution, in which the compartment-metabolism-replication coupling (CMR) is identified as a key step in the origin of life (Ruiz-Mirazo et al., 2017).

CHALLENGE 2

ABSTRACT

Understanding the bases of life and its evolutionary diversity requires deciphering the three-dimensional structure and dynamic nature of all macromolecules underlying living processes, and how they ensemble and function in a coordinated, timely and precise manner. Addressing this challenge will allow us to understand and treat diseases, harness biological processes for biotechnological purposes and synthetically design new biological entities.

KEYWORDS

Proteins | nucleic acids | macromolecular complexes | membrane-proteins | intrinsically disordered macromolecules | three-dimensional structure | dynamics molecular bases of disease | structural drug-design | cryoelectron microscopy | correlative microscopy | cryoelectron tomography | X-ray crystallography | nuclear magnetic resonance | single-molecule methods | proteomics

CHALLENGE 2

STRUCTURAL BASES OF LIFE AND EVOLUTION OF MACRO-MOLECULAR COMPLEXITY

Coordinators

José M. Valpuesta
(CNB)

Santiago Ramón-Maiques
(CBMSO)

Participating researchers and research centers

José M. Carazo
(CNB)

Carlos González
(IQFR-CSIC)

Juan A. Hermoso
(IQFR-CSIC)

Fernando Moreno-Herrero
(CNB)

Carmen San Martín
(CNB)

EXECUTIVE SUMMARY

Life relies on a myriad of processes carried out by macromolecular machines, formed mostly by proteins and nucleic acids. These machines have been perfected through evolution, becoming more complex, sophisticatedly regulated and integrated into dedicated operative pathways. Understanding the bases of life and its evolutionary diversity requires deciphering the atomic nature of these macromolecules, to explain how they are synthesized and assembled, and how they function in a coordinated, timely and precise manner. In the case of proteins, the major working molecules of life, their functions are largely dependent on their three-dimensional shape and dynamics. It is therefore necessary to learn these processes at atomic level to understand the molecular mechanism of action in depth, including its dynamics, which would allow us to find solutions for their malfunctioning and ultimately create new activities for our benefit in biomedicine and biotechnology. In the case of the nucleic acids, the structural landscape is much more diverse than previously thought. Non-regular DNAs are emerging as key structures in a variety of biological processes, such as genome transcription, repair or telomere maintenance, and RNA transcripts, including small and long non-coding RNAs appear to regulate almost every step of gene expression and have broad impacts on development and disease. Paradoxically, the function of an increasingly

number of proteins, protein regions and also RNAs appears to reside in their ability to remain unstructured. These intrinsically disordered macromolecules are involved, among other things, in promoting changes in the physical state of the cell milieu (phase separation), allowing the formation of membraneless cell compartments with multiple purposes such as transient storage or stress-response functions. Future studies should attempt to understand what triggers the formation of these novel membraneless bodies, and what is their relationship with their aqueous surrounding.

A further level of complexity is the analysis of how, where and when these macromolecular machines assemble and act in concert. The cell can be considered as a factory with multiple and sometimes short-lived compartments where the macromolecular content depends on cell needs and must be carefully controlled. This subcellular arrangement of macromolecules and their corresponding associated functions – termed "molecular sociology" of the cell – has a purpose that needs to be recognised. Of particular interest is the role of membranes not only as barriers to separate functions but also as concentrating points of specific ones in which membrane proteins have key roles.

The knowledge provided by structural biology and biophysical techniques will shed light on the evolutive changes that have generally driven macromolecules towards a higher complexity order, in most cases resulting in a finer tuning, and a higher control of their activities. A thorough comparative structural analysis of proteins, RNA and macromolecular complexes working in similar processes in different organisms will provide essential information to reconstruct the history and evolution of life. Ultimately, this knowledge will allow us to design new biological objects and entities and harness synthetic biology.

1. INTRODUCTION AND GENERAL DESCRIPTION

Life as we know it, with its extraordinary diversity of forms, their seemingly purposeful behaviour, and their ability to grow and reproduce, relies on a myriad of processes carried out by macromolecular machines formed mostly by proteins and nucleic acids. Life was born when a minimum set of these functional macromolecules assembled, and somehow compartmentalized using the first membrane-like structures, into self-perpetuating systems. These macromolecular machines have been perfected through evolution, becoming more complex, sophisticatedly regulated and integrated into dedicated

operative pathways. Understanding the bases of life and its evolutionary diversity requires deciphering the atomic nature of these macromolecules, explaining how they are synthesized and assembled, and how they function in a coordinated, timely and precise manner. This goal can only be achieved by the joint venture of research groups covering a broad range of disciplines. The CSIC is in a privileged position to tackle this ambitious challenge thanks to a thriving and coordinated community of qualified researchers focusing on different aspects of the structure and function of macromolecules.

Proteins are the major working molecules of life, performing key functions, which are largely dependent on their three-dimensional shape and dynamics. To carry out their activity, proteins undertake molecular recognition processes with other biomolecules, in which non-covalent interactions are crucial. Thus, it is necessary to understand these processes at atomic level to 1) understand the molecular mechanism of action in depth, including its dynamics; 2) find solutions for their malfunctioning; and 3) create new activities for our benefit in biomedicine and biotechnology. Our scientific community has numerous groups devoted to study the link between the structure and function of proteins as well as how they acquire their final 3D structure —sometimes assisted by other macromolecules—, and how chemical energy fuels conformational and mechanical changes to exert their action. Many diseases stem from defective proteins with altered functions or impaired structural properties, and thus, our research has a clear biomedical orientation. Besides, proteins are the common targets of medical treatments, either to restore function or to combat infectious diseases or tumours, and can also be manipulated for biotechnological purposes. Therefore, obtaining structural information of proteins is key for a better design of drugs with therapeutic or biotechnological applications. The same can be said about nucleic acids. The landscape of DNA structures is much more diverse than thought some years ago, and non-regular DNAs are emerging as key structures in a variety of biological processes, such as genome replication, transcription and repair, telomere maintenance, etc. On the other hand, RNA transcripts, including small and long non-coding RNAs appear to regulate almost every step of gene expression and have broad impacts on development and disease (Djebali et al., 2012). Understanding the structure of proteins and nucleic acids has made possible, for instance, to design novel genome editing tools such as CRISPR/Cas. Despite decades of intense research effort, we are far from understanding many of the general rules that determine how biomacromolecules acquire their final shape, and experimental structural determination is a basic need.

Naturally, experimentation and computational analysis go side by side helping us to decipher the structural and dynamical basis behind biomolecular interactions. This knowledge will allow us to design and synthetize new catalysts, scaffolds or drugs *a la carte*, with higher efficiency, stability or new purposes to intervene or emulate biological processes, bringing us closer to the challenge of constructing artificial (synthetic) life.

Paradoxically, the function of an increasingly number of proteins, protein regions and also RNAs appears to reside in their ability to remain unstructured. We have very limited knowledge on these intrinsically disordered macromolecules, which might acquire defined conformations upon interaction with other cellular components, or promote changes in the physical state of the cell milieu (phase separation), allowing the formation of biomolecular condensates (or membraneless cell compartments) with multiple purposes such as transient storage or stress-response functions. Future studies should tackle the fine analysis (*in vitro* and *in vivo)* of how disordered macromolecules introduce order, what triggers the formation of these novel membraneless bodies, and what is their relationship with their aqueous surrounding. Despite the advances in solving 3D structures of isolated proteins or assemblies, we are far from understanding the number and versatility of macromolecular complexes in the cell and when, where, and how they assemble and act in concert. The cell can be considered as a factory with multiple and sometimes short-lived compartments where the macromolecular content depends on cell needs and must be carefully controlled. A precise knowledge of protein and nucleic acids homeostasis – the fine balance between synthesis, folding, misfolding and degradation – is needed to extract some general rules on when and where macromolecules are present. Then, these macromolecules associate in more or less dynamic and transient complexes, and likely interact with other macromolecular assemblies either free or anchored to membranes, DNA or other cellular structures, to execute myriads of cellular functions that do not occur randomly. This subcellular arrangement of macromolecules and their corresponding associated functions – dubbed "molecular sociology" of the cell (Robinson et al., 2007) – has a purpose that we aim to recognise. An integrative approach combining numerous techniques, and the development of new ones, will be needed for understanding the dynamic arrangement of subcellular structures, and their relationship with different functions and the macromolecular complexes that execute them. Of particular interest is the role of membranes and cell walls as concentrating points of specific functions, other than the physical separation that they impose. Besides, membrane

proteins are key molecules in cellular communications, from signal transduction to transport of ions, metabolites and other molecules, and they protect living organisms from toxic factors. However, the structure of integral membrane proteins and the nature of membrane-lipid interactions are still poorly understood, despite its clear importance not only for understanding the very existence of the cell, but also as obvious targets for drug design. Membrane proteins are among the most difficult types of proteins to work with, from expression and purification to their analysis by almost any technology, creating the paradox that we know the least about one of the most important protein types (Cheng, 2018).

The knowledge provided by structural biology and biophysical techniques will shed light on the evolutionary changes that have generally driven macromolecules towards a higher order of complexity, in most cases resulting in finer tuning, and higher control of their activities. As far as we know, this is usually achieved through the insertion of new functional domains —in most cases coming from other proteins—, the incorporation of new subunits into complexes, and through the stable or most times short-lived interactions of the different complexes that form part of a given process, to make it more efficient and controllable. A thorough comparative structural analysis of proteins, RNA and macromolecular complexes working in similar processes in different organisms will provide essential information to reconstruct the history and evolution of life. Ultimately, this knowledge will allow us to design new biological objects and entities and harness synthetic biology.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Since the discovery of the 3D structure of the DNA, the ability to elucidate the structure of biological macromolecules has caused a broad and profound impact in biology, providing comprehensive understanding of reactions and processes that are central for life.

The importance of this field of research resides in part in the finding that misfolded and defective macromolecular structures are responsible for many diseases. The first “molecular disease”, sickle-cell anaemia, was recognized to be a result of a single mutation causing haemoglobin to polymerize into filaments that distort and destroy red blood cells. Since then, many molecular diseases, such as cancer or inborn errors of metabolism are known to result from

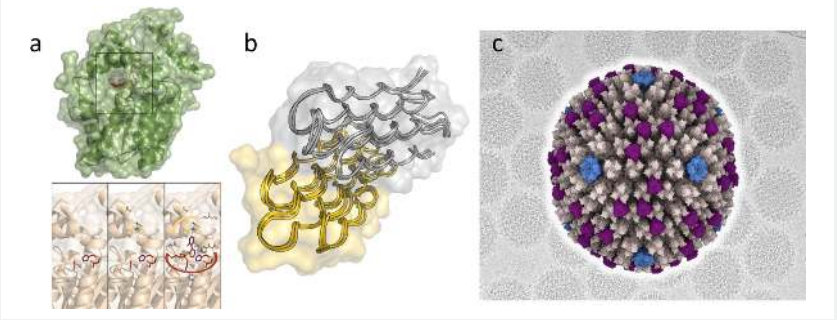
mutations that alter protein structure and function. In only two decades, the field of structural biology has expanded from the study of soluble proteins to membrane proteins, from isolated domains to multidomain and large macromolecular complexes, and from fixed portrays to dynamic views of macromolecular behaviours (Palamini et al., 2016). This knowledge has contributed centrally to the understanding of both common and rare human diseases, the mechanisms of pathogenic infections and to the discovery of remedies for these maladies. Structural techniques are well established in the pharmaceutical industry for drug-discovery, with a recognized impact in the identification of both ortho- and allosteric targets, and in the rational-design and optimization of small-molecule drugs and therapeutic proteins. Of the 210 new drugs approved by the US Food and Drug Administration between 2010 and 2016, 184 were informed by structural biology (88%) (Westbrook and Burley, 2019). In combination with powerful genomic and proteomic technologies, future advances in structural biology, particularly in the study of membrane proteins and large flexible molecular assemblies, will lead to a deeper understanding of pathologies (e.g., chemotherapy resistance mutations, antibiotic resistant bacteria), uncover more complex targets, yet unknown "druggable" pathways and new modes of action that will boost the development of new drugs and help to customize tailored therapies in personalized medicine.

Besides the medical interest, the structural knowledge of macromolecules and our ability to modify them, modulate their activity or synthetize at will, offers unlimited biotechnological applications. The structural-guided modification of proteins with increased efficiency/stability is a common practice in industrial processes, as well as in crop improvement or microbial engineering. Modifying macromolecules to produce self-assembled nanometric devices (nanomachines) with particular properties and controlled structure is also an innovative and exciting field at the frontier between biotechnology and synthetic biology for generating new materials, sensors or "smart" drugs (Schwarz et al., 2017).

3. KEY CHALLENGING POINTS

There are several challenges that need to be tackled for a proper understanding of how biological molecules originate, assemble, function, interact and evolve.

FIGURE 1–a) Crystal structure of the enzyme ester-hydrolase eh1ab1 from the metagenome of lake Arreo (Alonso et al., 2019). This is first evidence of a plurizyme with two engineered biological active sites, as an approach to endow new properties to enzyme scaffolds. The goal is the generation of highly effective biocatalysts for more sustainable industrial processes, and with enormous potential in many other therapeutic and diagnostic applications in biomedicine. **b)** Solution structure of the "snow flea" antifreeze protein dimer obtained from NMR data. Each monomer builds on six polyproline II helices, creating a flat and very rigid structure that will potentially be an excellent new material for biotech and synthetic biology applications (Treviño et al., 2018) **c)** Structure of the reptilian adenovirus at near atomic resolution by single particle cryoEM. Image provided by Dr. R. Marabini.



3.1. Folding and homeostasis of biological macromolecules

Proteins are large and complex molecules essential to the overwhelming majority of life processes. What proteins can do depend on their shape, on how they assemble, move and change. Over the past five decades, different experimental techniques (X-ray crystallography (Fig. 1a), nuclear magnetic resonance (NMR) (Fig. 1b) and cryoelectron microscopy (cryoEM) (Fig. 1c) revealed the 3D structures of a large number of proteins, and many more await to be determined. This knowledge has revealed the delicate connection between structure and function that controls all the life processes, and have shed light into the mechanisms of how proteins acquire their unique form.

The ability to predict a protein shape from its linear sequence of amino acids is known as the "protein folding problem". Recent fundamental breakthroughs in the experimental and computational fronts have awakened the interest to harness this important challenge (Senior et al., 2020), pushing the folding problem to proteins of increasing size, of disordered nature (see below), or capable of adopting more than one native form (e.g. morpheeins and metamorphic proteins). In addition, it is essential to consider that understanding how 3D structures are built and maintained requires to comprehend a

variety of molecular mechanisms controlling all aspects of protein homeostasis, from synthesis and nascent folding, through misfolding and aggregation to final degradation. Besides increasing fundamental basic knowledge, attaining this challenge will have a profound medical impact. Stress-, mutated-, or damaged-induced protein misfolding causes disorders. A number of higher-ordered misfolded protein assemblies (e.g. aggregates, inclusion bodies, aggresomes, stress granules) are linked to neurodegenerative diseases (e.g. Alzheimer's, Parkinson's, Huntington's disease, amyotrophic lateral sclerosis) and aging. Accurate prediction and understanding of protein folding will boost the research on any biological process and will guide the development of new therapeutic strategies.

We face a similar challenge in the comprehension of RNA molecules. The repertoire of structural and functional RNA domains has been neglected for decades and remains mostly uncharacterized. The great flexibility of RNA hampers structure prediction and functional inference, particularly when the molecules are long. Understanding the 2D RNA scaffolds and their stable or transient assembly into one or multiple 3D tertiary structures is mandatory to describe fundamental biological processes with an impact in medicine and biotechnology. New knowledge about the complexity and versatility of protein and nucleic acid folding will open opportunities for *de novo* design of synthetic macromolecules with specific shapes and particular functions of medical and technological relevance (see Challenge 7 and Challenge 8).

3.2. Structural and functional dynamics of molecular machines

To understand function, we need to study the atomic structure of macromolecules in action (Fig. 1). Structural biology, with the relevant contribution of X-ray crystallography and NMR, has elucidated 3D structures of a vast number of macromolecules (proteins, DNA, RNA) and their complexes, revealing the nitty-gritty of catalytic and regulatory mechanisms. Still, our current vision of the cell macromolecular landscape is very limited, and the shape and organization of many macromolecules, including the large majority of membrane proteins (see below), remain unknown. In recent years, technological advances such as high-resolution cryoEM and automatized processes have accelerated structural research of large macromolecular assemblies. We foresee that in the next decades, a combination of experimental and structural computational approaches will achieve the ultimate challenge of obtaining 3D information of all proteins and RNAs encoded in a genome.

In the coming years, the usage of state-of-the-art structural methodologies – mostly cryoEM and X-ray crystallography –, together with our ability to produce and isolate macromolecules, are expected to contribute a vast number of new structures, portraying fixed models, or at best, a limited number of populated conformational states. However, going beyond the static picture poses an additional major challenge, encompassing key outstanding questions that we should aim to solve:

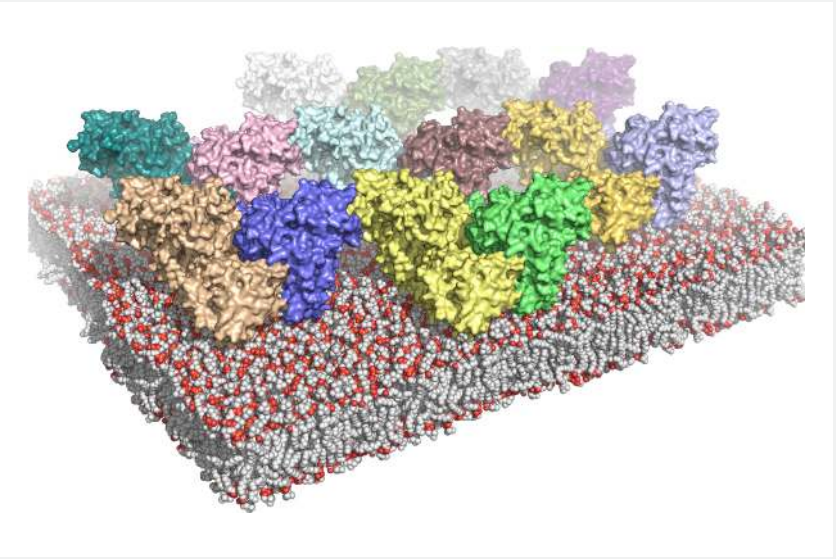
a. What are the internal dynamics and conformational flexibility of a macromolecule?
b. What are the active and inactive conformational states and their relative energies?
c. What barriers need to be crossed to switch between states and what is the kinetics of these transformations?
d. How do mutations, ligands and other elements in the cell environment alter the structural landscape?
e. How can we use this knowledge to model macromolecular properties?

Understanding the dynamic behaviour of macromolecular structures will involve the development of experimental techniques with spatiotemporal resolution including NMR, fluorescent spectroscopy, single-molecule biophysical methods such as optical and magnetic tweezers (Miller et al., 2018), ultra-fast EM, time-lapse and time-resolved X-ray crystallography, and the increased power and robustness of molecular dynamics and structure computational approaches. The integration of all available experimental and computational methods into hybrid structural approaches will be mandatory to: i) elucidate the function and dysfunction of proteins with well-structured domains and disordered regions; ii) understand how biological macromolecules interact with water, ions, lipids, and small molecule effectors in solution or at the membrane; iii) understand the consequences of modifications and mutations; and iv) develop new therapeutic drugs to combat for instance, antibiotic resistant bacteria, emerging pathogens, or chemotherapy resistant tumours.

3.3. Membrane proteins: folding, structure and dynamics

The study of membrane proteins is one of the biggest challenges at the frontier of structural biology, and deserves a special mention. Approximately one third of our genetic information encodes proteins (e.g. enzymes, transporters, ion channels, signalling receptors or energy transducers) that exert their

FIGURE 2–Polymerization of the Focal Adhesion Kinase (FAK) protein on the cell membrane, where it is activated upon interaction with other proteins. FAK regulates cell adhesion, migration and survival. Understanding this membrane-associated machinery will guide in the design of new strategies to fight cancer cell invasion and metastasis. Image provided by Dr. D. Lietha (CIB-CSIC).



function integrated within the lipid bilayers delimiting the cellular compartments. These proteins constitute prime targets for therapeutic drugs because they perform essential functions in the cell, such as controlling the flow of information and materials, or mediating hormone action and nerve impulses (Fig. 2). With membrane proteins being implicated in many different diseases (e.g. heart disease, Alzheimer's, cystic fibrosis) it is of crucial importance that their structures are characterized for novel therapeutic strategies and treatments to come to light. However, out of the ~8.000 known membrane proteins found in human cells, less than 300 have their atomic structure determined, mostly due to the difficulty in producing and isolating these macromolecules in soluble conditions.

Thus, a major hurdle will be to optimize or develop new techniques to engineer and produce membrane proteins that can be later characterized by structural methods. Some membrane proteins are synthesised by the ribosome just like other soluble proteins and then have to make their way to different

membrane locations within a cell, while others follow specific translation and translocation strategies. How does the cell cope with the unique and sometimes conflicting demands on membrane proteins for folding, translocation and stability? How are these proteins transported to the membrane? How do membranes shape protein structure? How do specific lipids interact and modulate protein activity? How do membrane proteins interact and assemble with other proteins either soluble or membrane-bound? How signal transmission is produced across the membrane? Hence, membrane proteins present three important challenges for which structural biology could be of paramount relevance:

a. To understand biogenesis, maturation and trafficking of membrane proteins (e.g. structural biology could contribute by reporting the intermediate states in biogenesis and/or the proteins implicated in translocation and transport of organelles).
b. To characterize the structure-function relationship of membrane proteins. A relevant derived question is to understand the role of specific lipids in the structure and function of membrane proteins. In this sense isolation of membrane proteins with their own lipids avoiding the use of detergents could be crucial.
c. To understand dynamics of membrane proteins and transient interactions among them. Relevant in this dynamic is to understand both, the internal dynamics of membrane protein itself and the external dynamics, i.e. their mobility across the membrane. Structural biology can make relevant contributions by characterization of transient complexes, such as a receptor bound to a channel or a receptor bound to a transporter. Characterization of such oligomers could be of critical relevance for their validation *in vivo* (e.g. by site directed mutagenesis) and to the development of new drugs (e.g. by specific design of drugs modulating interactions between proteins).

3.4. Liquid-liquid phase separation and the formation of biomolecular condensates or membraneless bodies

Contrary to previous thinking, intracellular organization is not only based on the formation of membrane-based compartments. Others compartments exist that rely on liquid-liquid phase transitions, a process by which a group of biomolecules – proteins and/or RNA – change from one physical state to another that separates from the general aqueous environment, triggering the formation of transient membraneless structures. Among these

membraneless organelles, the best known are the centrosome, viroplasms and P bodies in the cytoplasm, or the Cajal bodies and the nucleolus in the nucleus. These not well-understood compartments serve for multiple purposes such as storage, localised functions or virus assembly (Heinrich et al., 2018). The goal here is to analyse the fine structure (*in vitro* and *in vivo)* of membraneless bodies, to understand what triggers their formation and what is their relationship with their aqueous surrounding. A rapidly growing group of proteins with part or all of their sequence in a naturally disordered state is involved in the formation of membraneless bodies. Not much is known about these intrinsically disordered proteins (IDP), and different techniques must be used to study how they interact among them or with structured proteins, and whether they acquire secondary structure upon interaction.

3.5. Assembly of macromolecular networks

Macromolecules rarely act as single entities. Most of life processes, including cell signalling, metabolic pathways, immune response, or the replication and transmission of genetic material, depend on the often transient association of macromolecular complexes. To truly understand the structural bases of life we face the challenge of describing these macromolecular complexes and the inherent processes that control their assembly and disassembly. What are the stable and transient interactions between macromolecules and how do these influence their functions? How do these "interactomes" change in time, with cell type (e.g. neuron vs. hepatocyte, quiescent vs. proliferating/tumoral) or environmental conditions (e.g. metabolic state, stress, pathologies)? How can they be isolated or studied *in vivo?* Are there common patterns for protein interaction? Can we predict recognition interfaces? This challenge requires pushing forward the technological frontiers of integrated structural hybrid methods, in combination with super-resolution cell imaging techniques, novel biophysical methods, and massive -omics data.

Elucidating the dynamic organization of macromolecule networks will shed light on central biological processes, such as: i) the assembly-disassembly of viruses, ii) the architecture and regulation of chromatin, iii) the intricate complexity of cell-walls in plants or bacteria, v) the organization, function and "druggability" of membrane anchored macromolecular ensembles, vi) signalling events triggered by secondary messengers such as $Ca2^{+}$ and hormones, or vii) the recognition of natural chemicals or drugs by membrane receptors, among others.

3.6. In cellulo observation of molecular machines at work

The structural study of the macromolecules of life needs to consider the complexity of the environment and its impact on the structural dynamics of any molecular system. Besides, some macromolecules are functional – or even exist – only in their native niche and some supramolecular assemblies cannot be extracted from the cell. Thus, improvement of current techniques and development of new tools are needed to study macromolecular structures in their native context. We need non-invasive methods that provide a detailed molecular view of the intracellular environment, of the molecular crowding and of the distribution and relative disposition of supramolecular assemblies, including those at the membrane. In sum, we want to take a look at the macromolecular landscape of the cell to further investigate how actions are regulated.

This challenge will require modelling across disparate scales, integrating high-resolution cryoelectron tomography and, in general, time-resolved cryoEM, *in vitro* molecular function, correlated light microscopy and EM, and the accumulated structural and dynamic information of isolated macromolecular components with cellular -omics data (Rout and Sali, 2019). Using all this information (which is a type of complex big data problem, involving massive amounts of heterogeneous information), it will be possible to make a huge qualitative jump in our ability to reconstruct cellular interactomes and shed light on the interconnected cellular pathways. In the past few decades, there have been incredible progress in deciphering macromolecule structures thanks to recombinant molecular biology and "classical" structural methods. But there is much more to learn for the next generation of structural biology approaches, starting from the rupture with the simplification and idealization of the cell as a machine composed of discrete elements, and the assumption that by defining the structure/function of each and every one of these elements it will be possible to reconstruct the function of the system as a whole. Our current knowledge only allows us to grasp the entangled molecular organization, and we are still largely unaware of the number and complexity of the interactions, reactions and conditions within the cellular milieu, and how they modulate known and yet to be discovered macromolecular functions. In the next decades, we will use the experimental structural, dynamic and topological data to apply artificial intelligence to infer molecular interactions, interconnect cellular pathways and construct models of biological systems (see Challenge 5).

3.7. Understanding macromolecular evolution complexity

The evolution of life complexity is poorly understood. Although we accept that there has been an increase in complexity along the history of life, the origin of this complexity and the concept itself are hard to define. Modern biology has proven a deep relationship in genes, basic molecular mechanisms and macromolecular interaction networks between very distant organisms. Comparative structural analysis of the machinery responsible for similar processes across species will help us to reconstruct the history and evolution of life, providing an explanatory framework for understanding how macromolecules evolved into more controllable and sophisticated systems (Bamford et al., 2005; Beck et al., 2018). This knowledge, in turn, can eventually be used to design synthetic, tailored enzymatic pathways, signalling modules, self-assembling nanomaterials or the construction of a synthetic cell from its constituent molecular components that could include new functional modules (see Challenge 8).

CHALLENGE 2 REFERENCES

Albert, S., Schaffer, M., Beck, F., Mosalaganti, S., Asano, S., Thomas, H.F., Plitzko, J.M., Beck, M., Baumeister, W. and Engel, B.D. (2017). Proteasomes tether to two distinct sites at the nuclear pore complex. *Proceedings of the National Academy of Sciences 114*, 13726.

Alonso, S., G Santiago, I. Cea-Rama, L. Fernández-López, Coscolín, J Modregger, AK Ressmann, M Martínez-Martínez, H Marrero, R Bargiela, M Pita, JL Gonzalez-Alfonso, ML Briand, D Rojo, C Barbas, FJ Plou, P N. Golyshin, P Shahgaldian, J Sanz-Aparicio, V Guallar, and M Ferrer. (2019). Genetically engineered proteins with two active sites for enhanced biocatalysis and synergistic chemo- and biocatalysis. *Nature Catalysis 3*, 319–328. doi:10.1038/s41929-019-0394-4.

Asano, S., Fukuda, Y., Beck, F., Aufderheide, A., Förster, F., Danev, R. and Baumeister, W. (2015). A molecular census of 26S proteasomes in intact neurons. *Science 347*, 439.

Bamford, D.H., Grimes, J.M. and Stuart, D.I. (2005). What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol. 15* 655–663.

Beck, M., Mosalaganti, S. and Kosinski, J. (2018). From the resolution revolution to evolution: structural insights into the evolutionary relationships between vesicle coats and the nuclear pore. *Current Opinion in Structural Biology 52*, 32–40.

Cheng, Y. (2018). Membrane protein structural biology in the era of single particle cryo-EM. *Curr. Opin. Struct. Biol. 52* 58–63.

Djebali, S., et al. (2012). Landscape of transcription in human cells. *Nature 489*(7414), 101-108.

Efremov, R.G., Gatsogiannis, C. and Raunser, S. (2017). Lipid Nanodiscs as a Tool for High-Resolution Structure Determination of Membrane Proteins by Single-Particle Cryo-EM. *Methods Enzymol. 594*, 1–30.

Frank, J. (2018). New Opportunities Created by Single-Particle Cryo-EM: The Mapping of Conformational Space. *Biochemistry 57*, 888.

Gilbert, J.A. and Dupont, C.L. (2010). Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science 3*, 347–371.

Heinrich, B.S., Maliga, Z., Stein, D.A., Hyman, A.A. and Whelan, S.P.J. (2018). Phase Transitions Drive the Formation of Vesicular Stomatitis Virus Replication Compartments. *mBio 9*, e02290–02217.

Miller, H., Zhou, Z., Shepherd, H., Wollman, A.J.M., Leake, M.C. (2018). Single-molecule Techniques in Biophysics: A Review of the Progress in Methods and Applications. *Rep. Prog. Phys. 81*(2), 024601.

Murray, D.T., Kato, M., Lin, Y., Thurber, K.R., Hung, I., McKnight, S.L. and Tycko, R. (2017). Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell, 171*(3), 615–627, e16. https://doi.org/10.1016/j.cell.2017.08.048

Palamini, M., Canciani, A. and Forneris, F. (2016). Identifying and Visualizing Macromolecular Flexibility in Structural Biology. *Front Mol. Biosci. 3*, 47.

Parmar, M., Rawson, S., Scarff, C.A., Goldman, A., Dafforn, T.R., Muench, S.P. and Postis, V.L.G. (2018). Using a SMALP platform to determine a sub-nm single particle cryo-EM membrane protein structure. *Biochim. Biophys. Acta Biomembr. 1860*, 378–383.

Pfeffer, S. and Mahamid, J. (2018). Unravelling molecular complexity in structural cell biology. *Current Opinion in Structural Biology 52*, 111–118.

Robinson, C.V., Sali, A., Baumeister, W. (2007). The molecular sociology of the cell. *Nature 450*, 973–982.

Rout, M.P. and Sali, A. (2019). Principles for Integrative Structural Biology Studies. *Cell, 177*, 1384–1403.

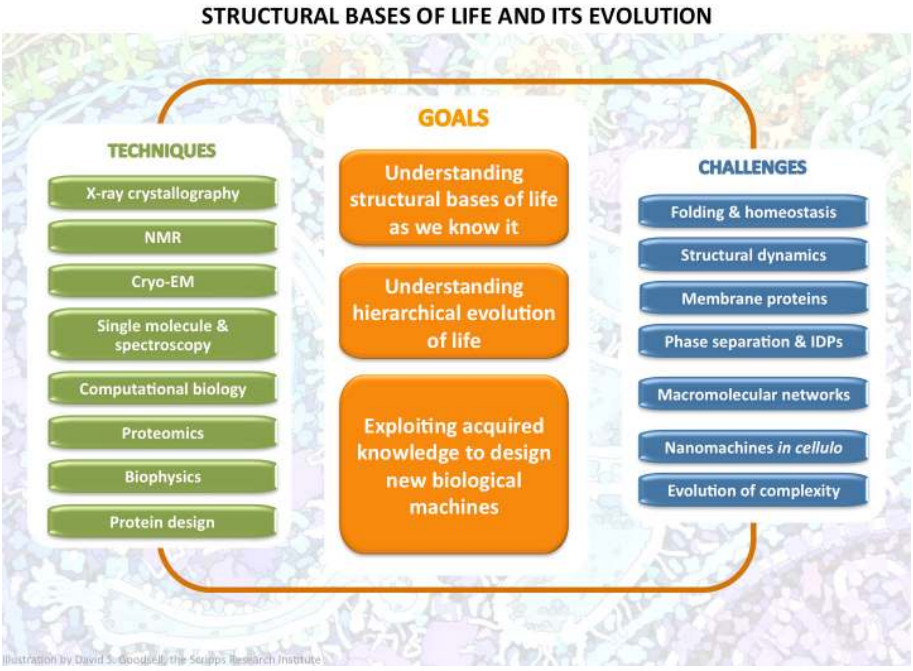
Schwarz, B., Uchida, M. and Douglas, T. (2017). Biomedical and Catalytic Opportunities of Virus-Like Particles in Nanotechnology. In Kielian, M., Mettenleiter, T.C. and Roossinck, M.J. (eds.), *Advances in Virus Research. Academic Press*, Vol. 97, pp. 1-60.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature 577*, 706–710.

Treviño, M.A., Pantoja-Uceda, D., Menéndez, M., Gomez, M. V., Mompeán, M. and Laurents, D.V. (2018). The Singular NMR Fingerprint of a Polyproline II Helical Bundle. *Journal of the American Chemical Society 140*, 16988–17000.

Westbrook, J.D. and Burley, S.K. (2019). How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drugs approval. *Structure 27*, 211–217.

SUMMARY FOR EXPERTS



SUMMARY FOR THE GENERAL PUBLIC

CHALLENGE 3

ABSTRACT

Evolutionary biology seeks to understand how biological diversity originates and is maintained. High-throughput sequencing permits assembling chromosome-level genomes, characterizing single-cell transcriptomes, and determining epigenomic modifications. Once widely applied to the diversity of living organisms, the reconstruction of the Tree of Life and the identification of the genomic targets of natural selection will be achieved.

KEYWORDS

Phylogenomics | Tree of Life | radiations | metagenomics | biotic frontiers | evolutionary genomics | comparative method | adaptation | homology | reference genomes | neglected taxa

CHALLENGE 3

THE TREE OF LIFE: INTERTWINING GENOMICS AND EVOLUTION

Coordinators
Rafael Zardoya, (MNCN-CSIC)
Ana Riesgo, (MNCN-CSIC)

Participant researchers and research centers
Silvia G. Acinas, (ICM-CSIC)
Paula Arribas, (IPNA-CSIC)
Damien P. Devos, (CABD-CSIC)
Rosa Fernández, (IBE-CSIC)
José M. Gomez-Reyes, (EEZA-CSIC)
Juan M. Gonzalez-Grau, (IRNAS-CSIC)
Jesús Lozano, (IBE-CSIC)
Borja Milá, (MNCN-CSIC)
Joaquín Ortego, (EBD-CSIC)
Jaume Pellicer, (IBB-CSIC)
Sergio Pérez-Ortega, (RJB-CSIC)
Ramon Rosselló-Mora, (IMEDEA-CSIC)
Gerard Talavera, (IBE-CSIC)
Miguel Verdú, (CIDE-CSIC)

EXECUTIVE SUMMARY

The continuous improvement of high-throughput sequencing opens the possibility of completing high-quality reference genomes for all living species in due time. This will allow reconstructing a robust Tree of Life (ToL) using phylogenomics and further our understanding of the genomic drivers underpinning the origin and diversification of life using evolutionary genomics. It is a cumbersome task not exempt of major challenges that require strong network collaboration and dedicated computer resources to manage and analyse big data. Main efforts will be centred on obtaining the samples (neglected taxa and uncultivated microbes) from biotic frontiers, dealing with giant genomes and important proportions of repetitive elements, identifying homology types and ploidy, detecting genomic hallmarks of selection, inferring candidate gene functions, and on gathering and incorporating long term natural history, geological, ecological, and environmental associated metadata under a phylogenetic framework. Certainly, the CSIC is in a privileged position to tackle such an endeavour, with renowned experts working across many microbial, animal, plant, and fungal lineages, conducting leading research in phylogenomics and evolutionary genomics. By setting a long-term programme under these auspices, the CSIC should be able to catalogue biodiversity, understand the

origin of species, unveil the mechanisms underlying evolutionary adaptation, enhance conservation of nature, and discover in related species within the ToL, numerous useful natural metabolites and drugs, which are the products of million of years of evolution and selection, contributing to human welfare and a better knowledge of global change on Earth.

1. INTRODUCTION AND GENERAL DESCRIPTION

Darwin's Theory of Evolution is a unifying principle in biology, which establishes natural selection as the main mechanism underpinning the origin and maintenance of biological diversity. For more than a century, evolutionary biologists have been documenting the endless pathways explored by natural selection to generate biodiversity, in order to infer general evolutionary laws. Understanding evolution requires a multilevel approach to determine ecosystem assembly and function, ecological interactions, and the genomic basis of adaptation. Finding appropriate model (or organismic) systems to study evolution is not easy, as it is a gradual process that takes many generations to become evident. One possibility is to draw upon artificial selection performed by humans on taxa either with fast evolutionary rates (viruses and bacteria), or that were domesticated in historical times. An alternative is focusing on cases in which natural selection either accelerated diversification rates (adaptive radiations) or ended in convergent solutions.

The top priority of evolutionary genomics for the coming years is to complete the genome sequences of every living organism in order to delimit species boundaries, reconstruct the Tree of Life (ToL), and perform comparative analyses aimed at determining the genomic drivers of adaptation (Richards, 2015). This is an ambitious task (there are at least 1.5 million named eukaryotes and between 1 and 10 million Archaea and Bacteria yet to be named; Yarza et al., 2014) that should become increasingly feasible as new sequencing technologies, bioinformatics tools, and computer resources improve beyond what is currently available (Lewin et al., 2018). For example, unlocking the access to microbial genomes without the need of purifying them was not possible until the metagenomic approach was developed (e.g., Almeida et al., 2019).

After the completion of the human genome, and in less than 20 years, evolutionary genomics has experienced an unprecedented momentum thanks to the continuous improvement of high-throughput sequencing technologies, which have steadily increased the yield of reads, decreased costs

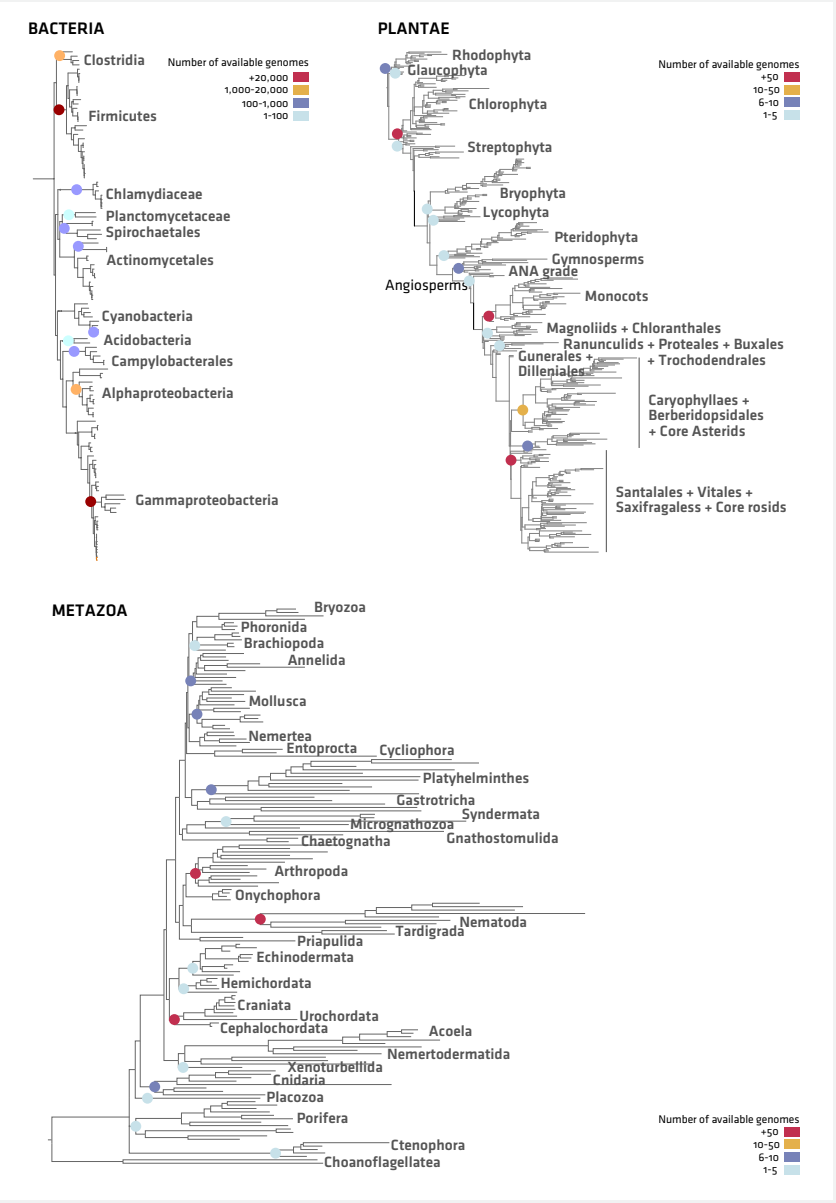
considerably, and improved the overall quality of results (Goodwin et al., 2016). At present, sequencing based on short reads is used widely to obtain transcriptomes, re-sequence genomes or generate thousands of phylogenetic markers through target enrichment. Moreover, it is now possible and increasingly affordable to obtain reference genomes with chromosome-level scaffolds through the combination of long reads (Amarashinge et al., 2020) and technologies that capture chromatin information.

While waiting for high-quality reference genomes covering all biodiversity, a plethora of other genomic resources was designed to address specific evolutionary questions in a cost-effective manner. These technologies are still useful and their efficiency will improve in the years to come. For instance, mitogenomes have been widely used to reconstruct robust phylogenetic trees (below the order level) for many years, and now they can be routinely sequenced on a species pool and subsequently assembled and separated into sequences corresponding to each species present (the so-called "mitochondrial metagenomics"; Arribas et al., 2019) or obtained as by-products of transcriptome projects (Plese et al., 2019).

Application of RNA sequencing to evolutionary genomics studies has also bloomed during the last decade, as transcriptomes constitute a good proxy of the whole gene repertoire of an organism, but cheaper to sequence and much faster to assemble, annotate, and analyse than a genome. Phylotranscriptomics has been used successfully to reconstruct robust large phylogenies (e.g., Laumer et al., 2019), but one of its caveats is the necessity of fresh specimens, as RNA degrades easily if not preserved appropriately. To circumvent this limitation, several genome skimming techniques were developed with the aim of high-throughput sequencing specific genes from DNA. These techniques have paved the way to "museomics", the high-throughput sequencing of preserved museum and herbarium samples, therefore unlocking a new wealth of precious material (type series, rare, and recently extinct specimens) key to illuminate the ToL (Trevisan et al., 2019).

Handling such massively generated sequence data has required the development of appropriate computational resources, which have flourished rapidly over the years. The cornucopia of bioinformatics tools now available to assemble genomes and transcriptomes allow the different research groups to construct pipelines tailored for their specific needs. There is also a wide variety of software packages at hand for automated annotation, although this step still renders many "hypothetical" protein-coding sequences due to the

FIGURE 1–Number of sequenced genomes in three selected clades of the ToL. The animal and plant phylogenies are modified from Laumer et al., 2019 and Chen et al., 2019. The setting of common high quality standards should be yet another advantage of building partnerships. International consortium-type efforts driven by researchers working in a given taxon should be complemented with broad programmes funded by institutions or governments.



incompleteness of the currently available gene databases. Phylogenetic methodologies are also readily adapting to the use of large multilocus sequence datasets. On the one hand, universal protein-coding genes are being selected on the basis of their relative evolutionary divergence to maximize phylogenetic signal (Parks et al., 2018). On the other hand, phylogenomics is shifting from concatenating all data to the use of coalescent-based approaches that infer trees from every single gene, thus obtaining a more complex vision of the evolutionary history of the organisms (Bravo et al., 2019). Finally, numerous statistical methods to deepen into patterns of evolution under the comparative method have been developed during the last years. These methods use phylogenetic frameworks not only to reconstruct the past (e.g., how a trait evolved, how fast a clade diversified) but also to inform about the present (e.g., how many uncultured microbial species are there), and to predict the future (e.g., how an alien species will invade an ecosystem, how a parasite will switch its host).

A genomic approach to global biodiversity requires collaboration of research communities at the international level given the great number of taxa that are currently targeted. Most efforts to date have been concentrated on vertebrates, and the vast majority of the ToL awaits attention (Fig 1). Research collaboration should work at different levels from coordination of the sampling effort to the sharing of computing resources and pipelines in the cloud.

As the sequencing of the human genome produced a paradigm shift in medical research, the sequencing of reference genomes representing life diversity on Earth should bring about a revolution in evolutionary biology in the coming years (Richards, 2015). It will be possible to address many of the current challenges in the field including reconstruction of robust phylogenetic relationships, improved determination of orthologous and paralogous relationships, characterization of the tempo and mode of gene family evolution, understanding of genome dynamics, identification of the genomic targets of natural selection, exhaustive detection of evolutionary innovations, recognition of causal connections between genotype and phenotype, recognition of the genomic regions and functions responding to environmental change, etc. Altogether, the availability of reference genome should set the basis to globally enhance biological research (and conservation) of previously neglected groups, as many latest technologies are only applicable if genomic data are available (Richards, 2015). A genomic approach to the ToL has also important applied deliverables and for instance, it should accelerate the discovery in

related species within the ToL of a variety of highly efficient and specific metabolites and natural drugs, which are the products of millions of years of evolution and selection (see also Challenge 2 and Challenge 7).

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

The completion of reference genomes representing the diversity of life should have a long-lasting impact on basic and applied biology. A first main outcome will be the full resolution of the ToL, which is essential for any downstream comparative biological research (see also Challenge 3 and Challenge 6). Resolving the ToL will generate enormous knowledge not only on biodiversity patterns and relationships, but also on ecosystem complexity and function, and will help discover the fundamental laws governing evolutionary processes (see also Challenge 4). This knowledge should enable conservation of biodiversity and maximize returns to society and human welfare (Lewin et al., 2018).

But there are still many unknown taxa to be incorporated into the ToL as well as many recalcitrant tree internal relationships to be resolved. A large portion of the unknown in the ToL is constituted by microbiota (Archaea, Bacteria, and unicellular eukaryotes; note that viruses cannot be included in the ToL, see Moreira and Lopez-Garcia, 2009) from extreme or highly inaccessible environments. A large survey of 16S rRNA diversity indicated that most of the known microbial diversity arises from the exploration of highly redundant environments, whereas all yet unexplored natural systems constitute a source of novelty (Yarza et al., 2014). The number of microbial species currently classified is seriously underestimated because many have never been brought to pure culture. In this regard, accepting a DNA sequence to become type material would open the door to classify metagenome assembled genomes (MAGs) and single cell amplified genomes (SAGs), enhancing the account of the real species diversity of the microbial world (Rossello-Mora et al., 2020).

One reason to unveil unknown microbial diversity is to broaden the general knowledge that has been gained from model organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*. As more microbial species are identified, it is becoming clear that the existing diversity of molecular, cellular, and functional biology in nature goes far beyond what has been learnt from model organisms (e.g., Rivas-Marín et al., 2016). Another important reason to enhance

microbial species discovery is to help discern the relative contribution of Archaea and Bacteria to the endosymbiotic origin of Eukarya (see also Challenge 1). For instance, a new phylum of Archaea, the Lokiarchaea, was recently identified through metagenomic analysis of deep marine sediments. The genomes of these cells contained the highest number of previously considered eukaryotic-specific features, although the cells lacked the eukaryotic-like cellular organization (Imachi et al., 2020). Moreover, various phylogenies consistently placed eukaryotes within the Lokiarchaea phylum, although the debate remains open. Finally, an applied reason to promote wide microbial prospecting is that they are likely the source of many new metabolic routes of importance in global cycles and many interesting products of biotechnological utility that could be detected as more efficient bioinformatics tools are developed.

Within eukaryotes, studying the origins of multicellularity will concentrate many efforts in the next decades. The evolutionary transition to complex multicellularity has occurred independently in plants, animals, fungi, (green, brown and red) algae, and some slime molds. The incorporation of unicellular relatives into phylogenetic studies has been of paramount importance to gain better knowledge on the origins of multicellularity. The transitions to complex multicellularity seem to require co-option of genes already present in the ancestral unicellular forms, which were already complex organisms, having extracellular matrix components and intricate signalling pathways (e.g., Sebé-Pedrós et al., 2017). The foreseen availability of reference quality genomes from more unicellular lineages, together with the implementation of genome editing technologies (e.g., CRISPR) will undoubtedly fuel studies on this topic.

One relevant source of unresolved nodes in the ToL is constituted by highly diversified taxa originated through rapid radiations, which are commonplace at different taxonomic levels. In addition to the difficulty of inferring phylogenetic relationships due to the intrinsic short internal branches, the challenge has been to determine whether shared polymorphism between radiated taxa is due to recent divergence and incomplete lineage sorting, or partly the result of hybridization and gene flow during speciation. Whole-genome assemblies will not only provide increased power to definitively resolve phylogenetic relationships in rapid radiations, but also to address the role of hybridization in promoting, not preventing, speciation (Stryjewski and Sorenson, 2017). In this sense, comparative genomics is and will be expanded enormously to the

population level (i.e., population genomics) aiming at understanding the relative role of natural selection, genetic drift, migration, hybridization, incomplete lineage sorting, and demography on the diversification of species. The continuum of speciation can be comprehensively studied at an unprecedented resolution from population fragmentation and ecological divergence to lineage split and species formation. This is of paramount importance to understand how organisms interact with the biotic and abiotic components of landscape heterogeneity, which has major implications to forecast their future responses to global change.

Connecting genotype, phenotype, and environment (see also Challenge 4) is still a major challenge that will benefit from evolutionary genomic studies (Edelaar et al., 2017). Phenotypic plasticity, the ability of a genotype to produce different phenotypes when exposed to different environments, is a pervasive feature of life. It may have important evolutionary and ecological consequences affecting biotic interactions and ecological niches, as well as shaping species coexistence and ecological network structure and dynamics (e.g., Sexton et al., 2017). However, the role of phenotypic plasticity in adaptation and the contribution of epigenomic changes remain largely unexplored and are topics that will be of particular relevance in the years to come due to anthropogenic pressure (habitat loss, global warming, invasive species, tolerance to pollutants, etc.).

Similarly, the genomics of adaptation is also a flourishing topic, fuelled by the increasing availability of high-quality genomes from a wide range of organisms. The broad implementation of genome-wide association studies (GWAS) such as those already performed to understand the diversity of dog breeds (Plassais et al., 2019), will be key in associating population genetic variants to phenotypic traits under selection, identifying the specific genomic regions involved in restricting gene flow among populations, understanding the relative importance of polygenic traits under the influence of many loci and those controlled by a few loci of large effect, as well as assessing the pleiotropic effects of single genes on different traits (e.g., Morris et al., 2019). Importantly, by comparing appropriate evolutionary model systems (e.g., adaptive radiations and/or cases of convergent phenotypes), a genome-wide approach to study adaptation and speciation will help revealing the relative importance of regulatory versus coding genomic regions as targets of natural selection, of key innovations versus multiple accumulative changes, of orphan genes versus gene families, etc.

At higher levels of biological organization, the use of phylogenies has broadened our understanding of ecological communities, being nowadays a standard approach in studies from many ecological disciplines such as conservation biology, community ecology, biogeography, and macroecology (Srivastava et al., 2012). Phylogenetic diversity may affect the functioning of ecosystems as intensively as taxonomic or functional diversity. The phylogenetic trait-based analysis of ecological networks emerges as a novel way of incorporating the evolutionary history of the interacting guilds to understand how they assemble. But the wider availability of genomic data for many taxa should definitely benefit conservation policies, planning, and management. Metabarcoding and mitochondrial metagenomics exploit all the potential offered by high-throughput sequencing to detect and identify anywhere thousands of species at a time from mass-collected, bulk samples of organisms or from environmental DNA (Deiner et al., 2017). These tools are applied to the study of manifold questions about spatial and temporal biodiversity patterns, as well as for biodiversity conservation and management. In fact, their combination with Earth Observation technology has been proposed as the most promising and efficient way of monitoring management impacts on biodiversity, its functions and services (Bush et al., 2017). Providing a phylogenomic context to the massive community level datasets generated by metagenomics, particularly from the so-called "biotic frontiers", opens a window to new analyses over these datasets, including the study of phylobetadiversity, diversification dynamics, or co-occurrence networks (e.g. Goberna et al., 2019).

3. KEY CHALLENGING POINTS

Sampling efforts at a global scale: The sampling of representatives of the different living species is and will likely be one of the most critical problems to solve. The access to taxa in the field is generally difficult and in the worst cases could be particularly costly in extreme environments such as the deep sea, or risky in politically unstable regions. There are important ecosystems that have been barely explored and their diversity is largely unknown. Among them, the ultimate "biotic frontiers" are probably within the microbial world, as well as the soil and deep-sea sediment mesofauna. At present, individual researchers mostly accomplish field sampling without further coordination. Therefore, there is an urgent need for large, multidisciplinary, and collaborative expeditions concentrated on biodiversity hotspots and biotic frontiers. Sound examples are the Our Planet Reviewed, Tara Oceans and Malaspina expeditions focused on diverse marine environments (e.g., Acinas et al. 2019)

Also, wide implementation of site-based approaches to characterise genomic diversity at the community scale could play an important role in sample acquisition (e.g., the Genomic Observatories; Davies et al., 2012).

Sampling for downstream genomic and transcriptomic analyses needs preservation methodologies that ensure obtaining high molecular weight DNA and intact RNA, respectively. This includes the proper handling in the field (including laborious tissue dissections) and adequate preservation in collections according to standardized protocols not yet widely implemented. Although there are automated sequence-based barcoding solutions for the identification of well-known species, the unambiguous identification of poorly known, cryptic, and unknown species ultimately requires sound reference collections of type material and the dedicated work of experienced taxonomists, often unavailable for neglected, highly diversified groups. The need for vouchers (representative samples deposited and stored in collections) and curated metadata as well as for protocols of data processing and sharing are then important issues that await coordination. Organized efforts are underway to sequence Bacteria and Archaea (the Earth Microbiome project; Gilbert et al., 2014) and Eukarya (the Earth BioGenome project; Lewin et al., 2018). These global initiatives use a taxonomically driven format, for which the contributions of natural history museums, botanical gardens, zoos, and aquaria are essential. To accelerate sampling, they intend to capitalize on the burgeoning citizen scientist movement (fuelled by the internet and social media) and new autonomous robotic technologies (Lewin et al., 2018).

Genome sizes, repetitive content, and ploidy. Despite the progress in sequencing technologies, there are some genomes that due to their large size, high content of repetitive elements, and/or polyploidy, remain a major challenge in terms of assembly and annotation. Genome size plays a key role as an evolutionary driver, given its implications in the biology of organisms (e.g., Pellicer et al., 2018), and it is a fundamental trait to consider when designing a sequencing project, as it provides essential information for estimating overall costs, needed resources, and expected drawbacks. For example, the assemblies of the giant genomes of the marbled lungfish (*Protopterus aethiopicus*, 1C=129.90 Gb) and the monocot lily *Paris japonica* (1C=148.80 Gb) are challenging and will require the development of new technologies and computer tools.

Genome size dynamics are mainly regulated by the relative frequency of amplification versus deletion of repetitive DNA and/or the incidence of

polyploidy. Repetitive elements (transposable elements and tandem repeats) constitute a significant fraction of animal and plant genomes. Given the ubiquitous nature of transposable elements, they may alter gene expression through insertions, activate responses to stress enabling genetic adaptations, or have an influence on chromosomal restructuring, among others. As repetitive elements accumulate in the genome through time, they are more likely to undergo erosion resulting in an overall landscape of degraded repeats, often called the "dark matter" of the genome (Maumus and Quesneville, 2014). Therefore, young genomes would have more homogenous repeat profiles, whereas in giant genomes the repetitive fraction of the genome would show a substantial proportion of uncharacterised and probably defunct elements. Deciphering the structure, function, and dynamics of the dark matter of genomes will be one of the major challenges in evolutionary genomics in the coming years.

Genomes resulting from polyploidy and/or whole genome duplication (WGD) events are particularly interesting for evolutionary studies but also challenging in terms of assembly and annotation due to their large sizes and important levels of paralogy. Polyploidy has been frequently associated with ancestral hybridisation episodes, and it has been largely studied in plants because of its consequences at the genomic and phenotypic levels. The increasing transcriptomic and genomic data being made available in recent years has evidenced that WGD has been a recurrent phenomenon in the evolution of plants (Landis et al., 2018) and occurred early in vertebrate diversification. Both in plants and animals, polyploidy is usually counterbalanced with genomic restructurings resulting in the loss of a large fraction of the duplicated genome. The retained duplicated genes may acquire new functions resulting in novel forms of adaptation. However, establishing a link between such processes and diversification bursts through the rise of new phenotypic acquisitions has proven complex (Landis et al., 2018) and constitute an interesting line of research for the near future. Finally, it is important to note that humans, through domestication, have long selected polyploids to improve aquaculture and agricultural systems (e.g., strawberries, cotton, salmon). The evolutionary dynamics of domesticated polyploid genomes and their adaptive consequences are and will be fascinating topics of research with applied deliverables.

Homology assignment. Homology, or the similarity due to shared ancestry, is a central concept in evolutionary biology. Identifying homology is key to understanding what has been retained by selection, and what has changed in

structure and/ or function during evolution. Two genes can be homologous if arisen through speciation (i.e., orthologous) but also if arisen through duplication (i.e., paralogous). Sequence similarity searches cannot distinguish orthologous from paralogous genes, and both from functionally convergent genes. The only way to assess homology is through the reconstruction of phylogenetic trees, and this is particularly challenging when analysing the complete gene set of a genome. A remarkable number of algorithms have been developed in the last decade to infer homology types. An immediate undesirable consequence is that homology assignments are in most cases not comparable. The past decade has seen a burst of genome and transcriptome sequences from non-model organisms, but often, these datasets are incomplete, and contain errors and unresolved isoforms. These can severely violate the assumptions underlying some homology inference methods. Hence, it is expected that as more high-quality genomes are assembled, homology determination will become more reliable, which is fundamental, as evolutionary genomics needs to distinguish the different types of homology, and the reconstruction of the ToL can only be based on orthologous genes.

A related problem is the accurate identification of orphan genes, i.e. genes restricted to a taxon that do not possess homologs in any other lineage (see also Challenge 2). Some animal lineages can have up to 30% orphan genes in their genomes (Fernández and Gabaldón, 2020). Orphan genes can arise from duplication, rearrangement (including fusion and fission) and further fast divergence, but also from *de novo* evolution of non-coding regions, including translation of neutrally evolving peptides (Rödelsperger et al., 2019). Orphan genes could also result from loss in stem lineages during evolution. For instance, a massive gene loss has been described recently in all lineages of animals (Fernández and Gabaldón, 2020), and some genes remaining in restricted clades might have then become orphan. The rapid population of databases with nearly complete genome sequences of rare, neglected or previously difficult-to-sequence taxa, will potentially modify the predictions for the number of orphan genes in most lineages.

Horizontal gene transfer. One of the biggest challenges of the current decade is the evaluation of the extent of horizontal gene transfer (HGT) occurring among Archaea and Bacteria in their natural environments, and how that would affect our view of the real diversity of these taxa. It has been proposed that prokaryote taxa evolution cannot be fully described without HGT (Palmer et al., 2019), and that genetic exchanges are so rampant that would blur the ToL at least for

these taxa. The latter may be too extreme, as the most recent phylogeny based on almost 100k genomes fairly mirrored the reconstructed trees based on the 16S rRNA gene, which is assumed to be inherited only vertically (Parks et al., 2018). At present, there are different platforms such as the MiGA database (Rodriguez-R et al., 2019), which acquire and organise high quality genomes from new microbial isolates, as well as MAGs from environmental samples and species microbiomes. Therefore, this will provide access to statistically sound datasets to both reconstruct the Archaea and Bacteria ToL and trace HGT. In eukaryotes, detection and validation of HGT have demonstrated to be far more complex, and often require high coverage of the genomes as well as strict controls for bacterial contamination. That said, there are plenty of examples of HGT to eukaryotes (Husnik and McCutcheon, 2018). In eukaryotes, when a complete gene is transferred from Archaea or Bacteria, the gene function can be retained, widening the functional complements of the organism, including nutritional improvements, toxin delivery, adaptation to extreme environments, and protection from archaeal or bacterial pathogens. There is no doubt that as genome quality improves and more neglected taxa are sequenced, the extent and diversity of HGT in eukaryotes will be finally unveiled.

Tree discordance. Phylogenetic analyses based on the different genes within and genome or a transcriptome render gene trees, which may differ from each other and may depart from the species tree (Degnan and Rosenberg, 2009). The study of tree discordance can provide useful insights on the effective population size of ancestors, rates of species divergence, and comparative information on how different genes evolved through time. Two non-exclusive evolutionary processes account for tree discordance: incomplete lineage sorting (ILS) and introgressive hybridization. ILS or deep coalescence occurs when intraspecific gene polymorphism lasts longer than speciation events. It is a widespread phenomenon, predicted to be highest when ancestral effective population sizes are large, as well as for cases of rapid species divergence. Post-speciation introgression, or the incorporation of genetic material from one lineage or deme into the gene pool of another by means of hybridization and backcrossing is also a common phenomenon widespread across the ToL (Mallet et al., 2016).

The accurate discrimination of factors that lead to tree discordance is one the major challenges in phylogenomics. Because ILS and introgression between closely related taxa may produce similar discordant phylogenetic trees, distinguishing between both is complex, and requires developing appropriate

statistical approaches that use whole-genome data in reduced taxon datasets (Durand et al., 2011). These tests are in their infancy and their improvement (e.g., by inferring the direction of gene flow in large taxon datasets) represents an open and active field.

Beyond phylogenetic inference, ILS and introgression are of great importance to understand the evolutionary processes promoting or limiting species divergence. This is fundamental for accurate species delimitation, and thus for properly assessing biodiversity and managing conservation units. Most research on ILS and introgression has been conducted under simulated scenarios or relatively small empirical datasets. However, in the coming years, there will be an unprecedented bloom of population genomics studies thanks to the increasing possibility of sequencing genomes at the population level, including genome phasing (i.e., distinguishing alleles), which provides a promising scenario to study ILS and introgression at a wider scale. This implies the use of coalescence-based phylogenetic methods (Bravo et al., 2019), which are currently under active development. Future implementations involving co-estimation of phylogeny and coalescence time at the genomic scale could largely address current caveats (e.g., systematic error). This is not yet possible and needs to become viable.

Incorporation of fossil data and molecular clock analyses. Extinct taxa represented in the fossil record allow understanding the stepwise evolution of characters and body plans, better constrain ancestral character states and infer the timing of major diversification events. In this regard, large-scale morphological matrices have been generated, even including stratigraphic ranges or biogeographic events. Incorporating paleontological data into phylogenies of extant taxa has a remarkable effect in topology inference (Koch and Parry, 2019), but it is not straightforward. During the last decade, a plethora of total-evidence methods have been developed that allow the simultaneous estimation and dating of the relationships among living and fossil taxa using molecular and morphological data, respectively (e.g., Ronquist et al., 2012). Whereas realistic models of evolution exist for the molecular partitions, the models for morphological evolution are not yet fully developed and this field is still in its infancy. Importantly, the computational burden of these approaches is prohibitive. The upcoming years will see a revolution as total evidence approaches resolve the above-mentioned problems and enter into the genomic era.

A phylogenetic tree represents both the relationships among taxa (topology) and their relative divergence from most common ancestors (branch lengths).

The latter can be used to infer the evolutionary timescale for the origin and splits of lineages, and thus inform about what were the circumstances (e.g., climatic, geological) surrounding the diversification processes. The most widely used strategy to date a phylogenetic tree is to transform branch lengths into time by calibrating certain nodes in the tree using the age of fossil taxa (past geological events could also be used); the so-called molecular clock analysis. This method has resulted in the establishment of timeframes for many lineages in the ToL (Blair-Hedges et al., 2015). Dating divergences ultimately depends on a robust phylogenetic justification and an accurate geological age of the fossils. During the last years, there has been a considerable effort to gather fossil data into curated catalogues such as the Paleobiology database (PaleoDB), which would facilitate the possibility of adding ages to nodes at the same pace as the ToL is fully reconstructed.

Computational resources. Overall, the number of genomic datasets sequenced in the last decade (including mitogenomes, chlorogenomes, nuclear genomes, transcriptomes, and target-enriched datasets) probably account for a few thousands (Fig. 1). This number will increase exponentially in the years to come, as high throughput sequencing becomes cheaper and widely available. Hence, the expected main challenges in this new era of big data are related to their storage and analysis. For example, raw data from a single transcriptome can occupy a few gigabytes whereas a genome needs half a terabyte of storage space. Computing clusters are not conceived as storage units. Thus, scientists face the challenge of finding a suitable and affordable storage space to keep their data (including backups) in the long term, and have them readily accessible through the cloud.

High throughput data analysis will be another major challenge in the years to come as computing time is a critical limiting factor. The assembly and annotation of large genomes currently need months of computing to be performed. Similarly, it is now feasible to obtain large phylogenomic datasets (including hundreds of taxa and thousands of genes) that demand important loads of computation time. For example, a phylogenomic analysis of the animal ToL containing 201 species and 422 orthologous groups required 1.5 years (run in parallel in 64 cores in a computing cluster under one of the most complex mixture models of amino acid substitution; Laumer et al., 2019). Moreover, this is applicable too to spatially-explicit landscape and phylogeographic models, time calibration analyses, population genomic analyses, etc. Thus, a revolutionary transformation in the analytical power is needed.

CHALLENGE 3 REFERENCES

Acinas, S.G., Sánchez, P., Salazar, P.G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., Sunagawa, S., Hingamp, P., Ogata, H., Lima-Mendez, G., Roux, S., González, J.M., Arrieta, J.M., Alam, I.S., Kamau, A., Bowler, C., Raes, J., Pesant, S., Bork, P., Agustí, S., Gojobori, T., Bajic, V., Vaqué, D., Sullivan, M. B., Pedrós-Alió, C., Massana, R., Duarte, C. M. and Gasol, J. M. (2019). Metabolic Architecture of the Deep Ocean Microbiome. *bioRxiv*, P. 635680.

Arribas, P., Andújar, C., Moraza, M.L., Linard, B., Emerson, B.C. and Vogler, A.P. (2020). Mitochondrial metagenomics reveals the ancient origin and phylodiversity of soil mites and provides a phylogeny of the Acari. *Molecular Biology and Evolution 37*, 683–694.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 21*, 30.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D. and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. *Nature 568*, 499.

Blair-Hedges, S., Marin, S., Suleski, M., Paymer, M. and Kumar, S. (2015). Tree of Life reveals clock-like speciation and diversification. *Molecular Biology and Evolution 32*, 835–845.

Bravo, G.A., Antonelli, A., Bacon, C.D., Bartoszek, K., Blom, M.P., Huynh, S., Jones, G., Knowles, L.L., Lamichhaney, S., Marcussen, T. and Morlon, H. (2019). Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ* 7, e6399.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., Martius, C., Zlinszky, A., Calvignac-Spencer, S., Cobbold, C.A. and Dawson, T.P. (2017). Connecting Earth observation to high-throughput biodiversity data. *Nature Ecology and Evolution 1*, 176.

Chen, J., Hao, Z., Guang, X. Zhao, C., Wang, P. et al., (2019). Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nature Plants 5*, 18–25.

Davies, N., Meyer, C., Gilbert, J. A., Amaral-Zettler, L., Deck, J., Bicak, M., Rocca-Serra, P., Assunta-Sansone, S., Willis, K. and Field, D. (2012). A call for an international network of genomic observatories (GOs). *GigaScience 1*, 2047-217X-1-5.

Degnan J.H., and Rosenberg N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution 24*, 332–340.

Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., De Vere, N. and Pfrender, M.E. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology 26*, 5872–5895.

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution 28*, 2239–2252.

Edelaar, P., Jovani, R. and Gomez-Mestre, I. (2017). Should I change or should I go? Phenotypic plasticity and matching habitat choice in the adaptation to environmental heterogeneity. *The American Naturalist 190*, 506–520.

Fernández, R. and Gabaldón, T. (2020). Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution 4*, 524–533.

Gilbert, J.A., Jansson, J.K., and Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology 12*(1), 69.

Goberna, M., Montesinos-Navarro, A., Valiente-Banuet, A., Colin, Y., Gómez-Fernández, A., Donat, S., Navarro-Cano, J.A. and Verdú, M. (2019). Incorporating phylogenetic metrics to microbial co-occurrence networks based on amplicon sequences to discern community assembly processes. *Molecular Ecology Resources 19*, 1552–1564.

Goodwin, S., McPherson, J. and McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Review Genetics 17*, 333–351.

Husnik, F. and McCutcheon, J.P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology 16*, 67.

Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M. and Matsui, Y. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature 577*, 519–525.

Koch, N.M. and Parry, L.A. (2019). Death is on our side: paleontological data drastically modify phylogenetic hypotheses. *BioRxiv*, 723882.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C. and Soltis, P.S. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany 105*, 348–363.

Laumer, C.E., Fernández, R., Lemer, S., Combosch, D., Kocot, K.M., Riesgo, A., Andrade, S.C., Sterrer, W., Sørensen, M.V. and Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B 286*, 20190831.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P. and Goldstein, M.M. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences USA 115*, 4325–4333.

Mallet, J., Besansky, N., and Hahn, M. (2015). How reticulated are species? *Bioessays 38*, 140–149.

Maumus, F., and Quesneville, H. (2014). Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One 9*, e94101.

Morris, J., Navarro, N., Rastas, P., Rawlins, L.D., Sammy, J., Mallet, J. and Dasmahapatra, K.K. (2019). The genetic architecture of adaptation: convergence and pleiotropy in Heliconius wing pattern evolution. *Heredity 123*, 138–152.

Moreira, D. and López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology 7*, 306–311.

Palmer, M., Venter, S. N., Coetzee, M. P. and Steenkamp, E. T. (2019). Prokaryotic species are sui generis evolutionary units. *Systematic and Applied Microbiology 42*, 145–158.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A. and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology 36*, 996–1004.

Pellicer, J., Hidalgo, O., Dodsworth, S. and Leitch, I.J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes 9*, 88.

Plassais, J., Kim, J., Davis, B.W., Karyadi, D.M., Hogan, A.N., Harris, A.C., Decker, B., Heidi G. Parker, H.G. and Ostrander, E.A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature Communications 10*, 1489.

Plese, B., Rossi, M. E., Kenny, N. J., Taboada, S., Koutsouveli, V. and Riesgo, A. (2019). Trimitomics: an efficient pipeline for mitochondrial assembly from transcriptomic reads in nonmodel species. *Molecular Ecology Resources 19*, 1230–1239.

Richards, S. (2015). It's more than stamp collecting: How genome sequencing can unify biological research. *Trends in Genetics 31*, 411–421.

Rivas-Marín, E., Canosa, I., and Devos, D.P. (2016). Evolutionary cell biology of division mode in the bacterial Planctomycetes-Verrucomicrobia- Chlamydiae superphylum. *Frontiers in Microbiology 7*, 1964.

Rödelsperger, C., Prabh, N. and Sommer, R.J. (2019). New gene origin and deep taxon phylogenomics: opportunities and challenges. *Trends in Genetics 35*, 914–922.

Rodriguez-R, L. M., Gunturu, S., Harvey, W. T., Rosselló-Mora, R., Tiedje, J. M., Cole, J. R. and Konstantinidis, K. T. (2018). The Microbial Genome Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Research 46*, W282-W288.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D.L. and Rasnitsyn, A.P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology 61*, 973–999.

Rosselló-Mora, R., Konstantinidis, K.T., Sutcliffe, I. and Whitman, W. (2020). Opinion: Response to concerns about the use of DNA sequences as types in the nomenclature of prokaryotes. *Systematic and Applied Microbiology 43,* 126070.

Sebé-Pedrós, A., Degnan, B.M., and Ruiz-Trillo, I. (2017). The origin of Metazoa: a unicellular perspective. *Nature Reviews Genetics 18,* 498–512.

Sexton, J.P., Montiel, J., Shay, J.E., Stephens, M.R. and Slatyer, R.A. (2017). Evolution of ecological niche breadth. *Annual Review of Ecology, Evolution, and Systematics 48,* 183–206.

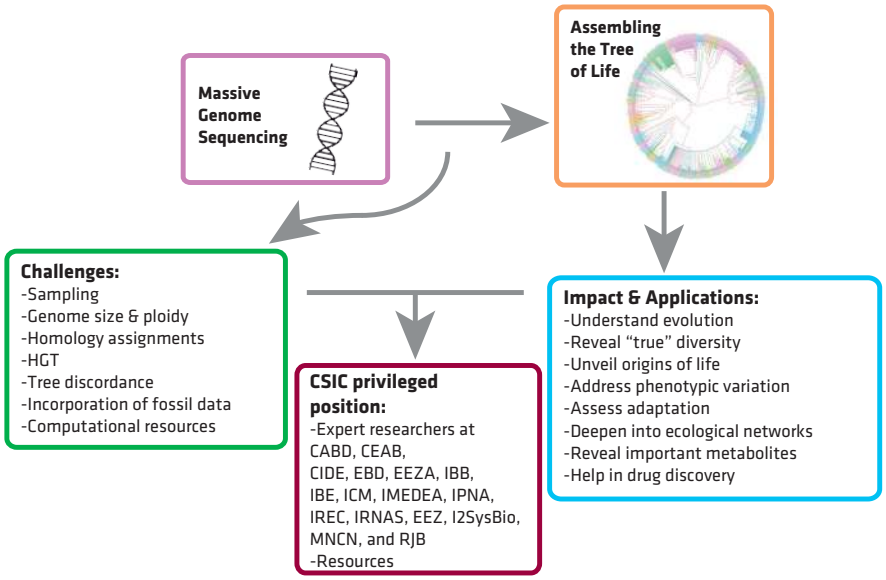
Srivastava, D.S., Cadotte, M.W., MacDonald, A.A.M., Marushia, R.G. and Mirotchnick, N. (2012). Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters 15,* 637–648.

Stryjewski, K.F. and Sorenson, M.D. (2017). Mosaic genome evolution in a recent and rapid avian radiation. *Nature Ecology and Evolution 1,* 1912–1922.

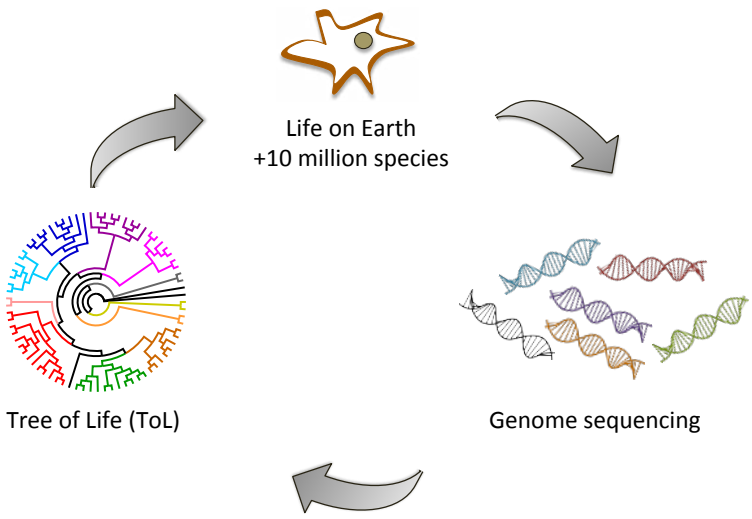
Trevisan, B., Alcantara, D.M., Machado, D.J., Marques, F.P. and Lahr, D.J. (2019). Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ 7,* e7543.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzéby, J., Amann, R. and Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology 12,* 635–645

SUMMARY FOR EXPERTS



SUMMARY FOR THE GENERAL PUBLIC



CHALLENGE 4

ABSTRACT

Understanding the development, maintenance, decline and evolution, *i.e.* the *genesis*, of the *phenotype* is a fundamental question in biology, with practical implications for human health, food production, or climate change, and that also impacts various areas of Engineering and Social Science. Specific challenges linked to this question have been identified and an ambitious strategy to position CSIC as a world-leading institution in solving these challenges is devised.

KEYWORDS

phenotype | encoding | plasticity | evolution | developmental biology | complex systems | mathematical and computational modelling | biological theory | health | engineering

CHALLENGE 4

THE GENESIS OF THE PHENOTYPE

Coordinators
Fernando Casares
(CABD, CSIC-UPO)
Ignacio Maeso
(CABD, CSIC-UPO)

Participants researchers and research centers
Josefa González
(IBE, CSIC-UPF)
Iván Gómez-Mestre
(EBD-CSIC)
Angela Nieto
(IN, CSIC-UMH)
Juan F. Poyatos
(CNB-CSIC)

EXECUTIVE SUMMARY

Biology has traditionally been divided into several subdisciplines that focus on different biological aspects and scales with the ultimate goal of understanding how living matter generates, replicates, and evolves form and function. However, such complex questions ultimately require multidisciplinary approaches that take advantage of the knowledge generated by the individual disciplines and integrates this knowledge to propose new questions that should allow to push the frontiers of knowledge.

Four topics have been identified that due to their complexity should benefit from such-multidisciplinary approach: (i) how biological information is stored and how this information is maintained, inherited, and changed; (ii) how this encoded information drives the emergence of the phenotype, *i.e.* of all perceivable traits and processes of living systems; (iii) how predictable is the phenotype and what is our capacity for engineering living systems; and (iv) how new phenotypes evolve and what are the temporal and spatial scales of evolutionary phenotypic changes.

Answering these questions should allow understanding the development, maintenance, decline, and evolution of phenotypes. These basic research questions have far-reaching implications not only for human health, food production or climate change, but also for various areas of engineering and social science.

Finally, while CSIC has the potential to be a leading research institution in solving these challenges, this will demand an important shift in the organization, function, and management of its existing centres. A transformation of this magnitude could only be accomplished with the establishment of a new multidisciplinar research institute.

1. INTRODUCTION AND GENERAL DESCRIPTION

The way in which living matter generates, replicates and evolves form and function at all levels of biological organization, is one of the central questions in biology (Thompson, 1945), yet one that is far from being answered. Living matter is characterized by a high degree of interdependence among its components and between them and the environment. It also has a strong dynamic nature, undergoing short-term cycles of development and reproduction, and long-term changes through geological time-scales (evolution), which themselves rely on the heritable variations in development. Life is a **dynamic, complex process** (Goodwin and Sole, 2002; von Bertalanffy, 1968). To examine this complexity, Biology branched into several subdisciplines that focused on different biological aspects and scales. Notably, and regardless of whether discussing the conformational structure of proteins or nucleic acids, different cell types, tissue architecture and homeostasis, body parts allometry, or the behavioural repertoire of organisms (all attributes of the organism and part of its *phenotype)*, all biological disciplines address comparable questions.

These questions include the examination of the many elements that can store biological information, how they vary, how these elements interact and contribute to defining the phenotype, and how this entire system evolves. The phenotype of an organism is what determines its relative success with respect to other conspecific individuals (i.e. its *fitness)* and how it interacts with other organisms, hence shaping the network of its ecological interactions. Despite its central role, our knowledge about the emergence of the phenotype is still very limited. Explaining the genesis, maintenance, and decline of phenotypes represents, in this way, the ultimate goal of Biological Research in the 21st century. CSIC aims to play a leading role in such new Biology. Furthermore, the principles that will be established by following through with this challenge could have implications in many other scientific disciplines, ranging from engineering to the social sciences. For instance, many aspects of robotics eventually confront the problem of defining the phenotype or the design of procedures of adaptation of this to fluctuating conditions.

Similarly, many social structures, from companies to cities, experience the problem of how encoded information emerges dynamically as a function of the whole.

Genetics has been a leading discipline in pursuing some of these questions, with the discovery that the main elements that store biological information are the "genes", understood as specific DNA sequences that contribute to the "genotype". DNA sequence variation is a major source of evolutionary change and takes many forms including from single nucleotide polymorphisms (SNPs), to copy number variants (CNVs) such as insertions, deletions, genomic duplications, and genomic rearrangements, to incorporation of foreign DNA, for example through species hybridization (introgression), or invasions of transposable elements and other lateral gene transfer phenomena. Even within this seemingly simple framework, many questions on the emergence of the phenotype remain. Genes interact with each other during development and across physiological processes at multiple levels (epistasis) (Lehner, 2011). Also, it must be recognized that the function of most genes is not unique and that they can influence several distinct phenotypic traits (pleiotropy) (Saltz et al., 2017). The effect of genes on the phenotype often depends on the environment (gene-environment interactions) (Eguchi et al., 2019), and that environmentally-induced changes in gene expression can lead to novel phenotypes and evolutionary innovations (Moczek et al., 2011; West-Eberhard, 2005). Thus, the combination of different input "drivers" (internal and external) throughout the life of an individual will influence the phenotype and, especially relevant for humans, the quality of life and the expectations for healthy ageing (Connallon and Hall, 2018; Lehner, 2013). It is now recognized that the experience of progenitors can carry on consequences to the progeny (transgenerational effects). Therefore, the resulting output in each individual could also have an impact on the future population (Jablonka and Lamb, 2005). These insights help recognizing the complexity of the questions posed in this challenge. **Four research topics** have been identified that should allow researchers at CSIC to push the frontier of knowledge on the genesis of the phenotype.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATION

The first topic corresponds to **how biological information is stored and how this information is maintained, inherited, and changed**. This **"encoding"** is not linked anymore to a "list of genes". Indeed, the very idea of the

genotype is under scrutiny since it is clear that the primary structure of DNA (i.e. its linear sequence) is not the sole template to store biological information. For instance, in addition to variation in the DNA primary structure (both nuclear and organelle DNA), epigenetic modifications of various sorts, including histone post-translational modifications and regulatory non-coding RNA, contribute to the heritable basis of our phenotypes, at least in short temporal scales (Harvey et al., 2018; Skvortsova et al., 2018).

Several additional factors contribute to biological information, such as maternal effects in the form of subcellular organelles, yolk nutrient provisioning and hormones, parental inter- and trans-generational effects, etc. Which elements should then be consider to fully capture the encoding of the phenotype? Is there a natural encoding hierarchy that helps store core pieces of the phenotypic information fixed? Are there particular encoding strategies to promote both qualitative and quantitative variation? Solving the questions in this topic will lead to a redefinition of the "genotype" which will have a significant impact on the future of genetics.

The second topic goes from the questions of the information storing to **how this encoded information drives the emergence of the phenotype** (the **"interpretation")**, and how environmental conditions modulate that driving. Which type of mechanisms regulate the emergence of phenotypes in biological systems? It is likely that not all encoded information is observable in all individuals, or in many of the environments these individuals may occupy. If this is the case, the range of phenotypes that an organism can display should be addressed. What are the limits to the phenotypes an organism can develop? To what extent is the environment an inductor of phenotypes in addition to a selective filter? And what are the temporal scales at which the environment exerts its effects? How do the different factors shaping the phenotype interact over the course of development? These questions lead, of course, to more specific problems: how are deleterious mutations tolerated in cells (Eyre-Walker and Keightley, 2007), how do multiple traits change simultaneously (Cheverud, 1988), is the mapping from encoding to phenotype modular (Wagner et al., 2007); how do environmental conditions adjust the expression of information (Arnold et al., 2008); what are the costs and limits associated with phenotypic plasticity (Murren et al., 2015); is there a hierarchical relationship between contributing sources of phenotypic variation (Phillips and Arnold, 1999), and how phenotypes vary along the life cycle of the individual and degenerate during disease and ageing (Bahar et al., 2006), to name a few.

All these factors, internal and external to the organism, contribute to determine how the phenotype is built and maintained, and they do so in a very complex network of interactions, which often act synergistically and with feedbacks, therefore resulting in non-linear dynamics. This second topic will, therefore, significantly explore how the phenotype *ages* (Metcalfe and Alonso-Alvarez, 2010; Wiley and Campisi, 2016), and how it deviates from a normal standard, healthy phenotype, into a diseased state.

The third topic addresses **the predictability of the phenotype and the capacity for engineering living systems** to produce designer phenotypes. Given a better knowledge of the genesis, maintenance, and decline of phenotypes, will it be feasible to predict the phenotype from knowledge about the environment and a complete genotype (defined beyond the current DNA sequence)? Can new phenotypes be designed and their production engineered? One could anticipate that this is the case and assume two possible strategies. In the first one, a phenomenological description of biological phenotypes can be produced. This requires identifying the appropriate variables to represent phenotypes, i.e., to reduce complexity (Anderson, 1972). The identification of the variables will not be enough, though, as they need to be integrated in a dynamical system. This will require new mathematical and computational modelling approaches. For instance, by incorporating modern data assimilation techniques, which connect models and experimental data to better understand dynamics. A second strategy might require the development of sophisticated artificial intelligence tools that help identifying all the encoded information influencing the expression of the phenotype so that patterns can be identified (Topol, 2019). It might be that the capacity to engineer biological systems is achievable, but with no clear fundamental understanding (e.g., airplanes fly but without a precise explanation of why birds do so). Besides, it could be that there exist intrinsic limitations in the capacity of predicting phenotypes (Waddington, 1957). One extension of this research will lead to engineering phenotypes *in vitro*. New techniques allowing the culture, differentiation and genetic editing of pluripotent stem cells might offer the possibility of implementing designed phenotypes for the improvement, repair, replacement of organs, or even the development of new "biobots" with programmed, built-in capabilities (Kriegman et al., 2020), which could also include details on the biomechanics of such cellular organizations. This topic will bring research to biomedical applications for the repair of damaged, diseased or aging tissues, the enhancement of existing organs and the design of novel phenotypes.

The fourth and final topic focuses on **evolution**. To explain the genesis of the phenotype means to understand how the amount of change in the elements encoding the information connects to the variation of phenotypes on which selection acts, and how novel phenotypes arise. It has been argued before that the mapping from the stored information to the phenotype can constrain some of this variability, but which mechanisms contribute to the properties defining this mapping, and how they evolve, are unknown (Uller et al., 2018). Is the architecture of biological systems promoting the evolution of the phenotype (Watson et al., 2016)? Is this evolution predictable? A good comprehension of the spatial and temporal scales of the evolutionary dynamics is also needed (Catullo et al., 2019; Segar et al., 2020), and how evolution could transform the encoding to the phenotype map (Martin et al., 2015). Conversely, are there special peculiarities of this map that stimulate the evolution of complex architectures? This map will not be a static map of "final phenotypes", but a dynamic one, in which each phenotype will be seen as a temporal *trajectory*.

Broader Impacts. There are a number of areas onto which this research program will have a significant impact:

a. *Human biology and health*

- **a.1** *Personalized medicine*. Understanding the decoding of the drivers of the phenotype will allow, in the case of human disease, to use the specific genetic and non-genetic components of each patient in order to understand the individual specificities of the pathology and to propose personalized interventions to steer the diseased phenotype towards a healthy one.
- **a.2** *Tissue repair*. Understanding the mechanisms that regulate phenotypic homeostasis will help in the design of better strategies to promote tissue repair and regeneration.
- **a.3** *Tissue and organ engineering*. The understanding of the mechanisms controlling developmental operations, combined with computational studies and *in vitro* systems will allow the controlled design and synthesis of tailored tissues and organs for replacement. New mathematical and computational modelling strategies will be essential in this pursuit.
- **a.4** *Quality of life and healthy ageing*. All of the knowledge obtained from the above can have a big impact on the design of better therapeutic strategies for pathological states. Importantly, it will also help in the

knowledge-mediated promotion of quality of life and particularly, healthy ageing.

b. *Ecology and global change*. By identifying the drivers of the phenotype and their effect on the development of the organism, it may be possible to predict the effects of ecological change (including temperature) on the organisms and thus anticipate the consequences of global change. Understanding how these changes impact the phenotype should also enable anticipating which species are more vulnerable and which more capable of coping with environmental change.

c. *Animal and vegetal production*. Farm production of animal and vegetal products will benefit from a better understanding of the generation of some of the traits with commercial/nutritional/biomedical value and from a better understanding of the engineering required for improving these traits or generating new ones.

d. *Exobiology*. In the search for extra-terrestrial life, understanding the principle of storage, encoding and decoding the information required for the building, maintenance and evolution of terrestrial life might help looking for similar organizational/informational principles elsewhere, regardless of their physical implementation.

e. *Transdisciplinary impacts*. A number of contemporary disciplines investigate the complex dynamics of many interacting parts, with the aim of predicting and engineering their behaviour, including robotics (e.g. robot swarm behaviour), social sciences (e.g. smart cities, social organization systems, social networks) and engineering (e.g. smart (energy) grids; information networks). The principles derived from the study of the drivers of the phenotype, their decoding and timescales of action may help establishing useful analogies in those fields. Reciprocally, the study of biological organization and the emergence and variation of the phenotype should benefit from the cross-fertilization of these other areas. The large-scale integration of the mutual and constant interactions between biological systems and the environment (as complex as cities, health systems, cultural factors, or technological development) will give a global view of how the phenotype emerges and changes.

f. *Societal & ethical impact and public policies*. Among the different theoretical frameworks and concepts developed within the life sciences, those associated to the present challenge (evolution, genotype-phenotype maps, etc.) are probably the ones with the strongest and most long-lasting impact on sociology, philosophy and even politics and

religious beliefs, such as the profound cultural transformation triggered by Darwin with the publication of *On the Origin of Species*. The same applies to many of the technological advances associated with this challenge (bioengineering, CRISPR, personalized medicine). This high societal impact comes with a cost, however, and there are multiple historical examples in which owing to insufficient or inadequate efforts in science communication or to cases of misconduct within or outside the scientific community, the misuse of this impact has led to dramatic consequences in society and in the design of public policies (ranging from eugenics during the 20th century to the use of genetic determinism as a justification of educational policies and income inequalities in the present). Thus, the future knowledge gained in the next 50 years on the genesis of the phenotype should be accompanied by interdisciplinary efforts with social scientists (such as those from CSIC's Instituto de Políticas y Bienes Públicos) to prevent these problems and to promote a beneficial impact on society. Not only our philosophical ideas will change. The possibility of predicting the course of human development or the capacity to design organs or organisms -ourselves included- will have a huge impact on how our societies are organized and evolve, and will demand major political, philosophical and ethical debates.

3. KEY CHALLENGING POINTS

Unveiling the full repertoire of phenotype drivers. During the past 70 years, our knowledge of the genotype-phenotype maps has been completely transformed since the discovery of DNA as the genetic material. This nucleic acid-centric view has been immensely productive, but to fully understand how phenotypes are generated and evolve a broader perspective will be needed in the future. Each day, additional ways in which living systems store the information for the building and maintenance of the organism's phenotype are described, ranging from histone post-translational modifications to prions and cell polarity (Ciabrelli et al., 2017; Duempelmann et al., 2019; Harvey et al., 2020; Jung et al., 2017; Klosin et al., 2017; Lev et al., 2019; Manzano-Lopez et al., 2019; Posner et al., 2019; Shirokawa and Shimada, 2016), many of which could not be investigated in detail before due to technical limitations. Can phenotypes be fully understood by focusing exclusively on DNA? How many other biological structures have been overlooked that could have the potential to act as hereditary factors? Do they have common properties? What is their relative contribution to the phenotype in comparison with DNA? Do all non-DNA

memory systems operate at the same temporal and organization scales? Have different lineages independently evolved different additional hereditary systems? Can novel synthetic systems be designed?

Experimental tests of the hierarchy of drivers of the phenotype. Different study systems amenable to experimentation will be used to understand how different drivers of phenotypic variation, such as genetic variants, maternal/paternal contributions, environmental induction, and biotic interactions, each contribute additively or synergistically to the generation of phenotypic variation. By focusing on the analysis of a single trait relevant across organisms (*e.g.* stress response), the understanding of how different organisms have found solutions to the same challenges should be possible. This knowledge on the different strategies available in nature could be used to engineer some of these solutions in our organisms of interest (*e.g.* humans in the context of disease, animals and plants in the context of food production).

Defining the scope of phenotypes. Biological organisms constitute the base on which the genotype-phenotype relationships have been classically framed. However, the traditional definitions and limits of biological organisms are currently under strong scrutiny. Currently, more inclusive concepts of biological organisms are being defined, such as the "holobiont" (Roughgarden et al., 2017), in an attempt to account for the complexity of co-constructed multi-species systems, and it is likely that even broader concepts will be developed in the future. How can the phenotypes of these supra-organismal systems be described? Is it possible to apply the concept of phenotype to much broader scales than has have done so far? If so, how are these supra-organismal phenotypes generated/developed? How do they evolve? Where and how are these phenotypes encoded? What kind of experimental approaches should be developed to study them?

Prediction of the phenotype. One potential approach to anticipate phenotypes is the use of large data sets including omics features, ecological parameters, morphometry (derived from using advanced imaging techniques), etc. These datasets will be analysed using deep learning and other artificial-intelligence (AI)-based tools with the aim of predicting the phenotype. These AI techniques may be used to define "laws" of mapping. It should also contribute to validate our trust in these approaches in relevant biomedical scenarios.

The homeostasis and decline of the phenotype. An extension of the study of the emergence of the phenotype with important implications for human health

is how this phenotype is maintained in health and how it degenerates in disease and ageing. This research program examines the time-scale of the life of the individual (the life-cycle time frame) and includes the specific analysis of all the factors that contribute to the maintenance of a healthy function and its changes as the individual ages. What is known about normal variations? Where is the boundary between physiological and pathological phenotypes? Given the known reactivation of developmental programs in adult disease (Nieto, 2013), how can this acquired knowledge on the emergence and maintenance of the phenotype be used to understand pathology? The knowledge of the mechanisms behind phenotype maintenance and tissue homeostasis should help understand those behind its decay or loss during disease and ageing. Major efforts will be devoted to understanding how other organisms age, repair, and regenerate their wounds or are less susceptible to certain diseases than humans are. This knowledge will be used in projects aimed at improving human health.

The requirements for self-organization. Complex biological systems, such as the active matter at the cell's cortex, the early embryo, or the retina, involve a number of interacting parts. This challenge will determine which are the requirements for the self-organization of biological phenotypes at different scales. For example, are there lower (and upper) limits to the number of cells necessary for organizing themselves into an early embryo-like structure, such as an embryoid, in which symmetry is broken and a basic germ layer organization arises?

4D characterization of development at single-cell resolution. The dynamics of specific developmental processes (as modular "developmental operations") will be analysed in space and time at a single-cell resolution, including transcriptomics, proteomics, and metabolomics and measuring the energetic costs and informational flows. This aim will require the development of **imaging techniques** capable of capturing the dynamic behaviour of many cells, coupled with **multiplexed detection** of molecules and **computational modelling.** The use of these models will be two-fold: on the one hand, they will be able to describe all interactions and determine parameters quantitatively. On the other hand, they will be used to derive global properties governing the dynamics of the developing process. Data meta-analyses will allow the comparison of normal and diseased states, or between different species showing distinct phenotypes.

*Synthetic developmental biology and evolution***.** The aim will be to synthesize organs with designed cellular composition and morpho-functional

characteristics. This challenge will require (1) defining the major developmental "operations" (such as cell division, tissue folding, collective cell migration); (2) obtaining a systems-level understanding of these developmental operations, including parameters that can modulate the behaviour of each of them (for example, two species may fold a neuroepithelium to develop one organ, but the fold may be deeper in one of the two species); (3) in silico evolutionary selection of phenotypes combining the developmental operations (in silico/computational evolution); and (4) biological implementation of these phenotypes (either in vitro or through genetic modification of specific species) through biological engineering ,either through modification of extant organisms, using organoids or by *de novo* design and synthesis.

Examining the breadth of phenotypic diversity while keeping the challenge in focus. The success of addressing this challenge will depend on the definition of clear goals and experimental systems. For that, it will be essential to have a two-fold strategy that is at the same time "centrifugal" and "centripetal". A centrifugal strategy will be needed in order to expand and diversify the extremely short list of organisms currently used as experimental systems. Features such as phenotypic plasticity, generation times, regeneration capabilities, biogeographical distribution, complexity of the life cycle, among others, are very different across species and therefore not all of them will be equally tractable to address the different topics and questions of *the genesis of the phenotype* challenge. Even more important, certain biological phenomena of particular relevance for this challenge (for example, adaptive prions or DNA/chromosome elimination), have only been described so far in a handful of species (Harvey et al., 2018; Smith, 2018). Thus, to have a complete picture, it will be important to target and study a broad range of organisms that is more representative of the full biological diversity, including species that cannot be currently cultured in the lab but should be amenable to experimentation in the near future. A centripetal strategy will be equally important, since a good understanding of the genesis of the phenotype will require a real integration and coordination of multidisciplinary approaches. Thus, to implement this type of combined actions it will be necessary to design specific projects that centralize these multidisciplinary efforts by focusing on particular sets of model organs/organisms/processes.

CHALLENGE 4 REFERENCES

Anderson, P.W. (1972). More is different. *Science 177*, 393–396.

Arnold, S.J., Burger, R., Hohenlohe, P.A., Ajie, B.C., Jones, A.G. (2008). Understanding the evolution and stability of the G-matrix. *Evolution 62*, 2451–2461.

Bahar, R., Hartmann, C.H., Rodriguez, K.A., Denny, A.D. et al. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature 441*, 1011–1014.

Catullo, R.A., Llewelyn, J., Phillips, B.L., Moritz, C.C. (2019). The Potential for Rapid Evolution under Anthropogenic Climate Change. *Curr Biol 29*, R996–R1007.

Cheverud, J.M. (1988). A Comparison of Genetic and Phenotypic Correlations. *Evolution 42*, 958–968.

Ciabrelli, F., Comoglio, F., Fellous, S., Bonev, B., Ninova, M. et al., (2017). Stable Polycomb-dependent transgenerational inheritance of chromatin states in Drosophila. *Nat. Genet. 49*, 876–886.

Connallon, T., Hall, M.D. (2018). Genetic constraints on adaptation: a theoretical primer for the genomics era. *Annals of the New York Academy of Sciences 1422*, 65–87.

Duempelmann, L., Mohn, F., Shimada, Y., Oberti, D., Andriollo, A., Lochs, S., Buhler, M. (2019). Inheritance of a Phenotypically Neutral Epimutation Evokes Gene Silencing in Later Generations. Mol. Cell. 74, 534–541 e534.

Eguchi, Y., Bilolikar, G., Geiler-Samerotte, K. (2019). Why and how to study genetic changes with context-dependent effects. *Curr. Opin. Genet. Dev. 58–59*, 95–102.

Eyre-Walker, A., Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8, 610–618.

Goodwin, B., Sole, R. (2002). *Signs of Life: How complexity pervades biology*. Basic Books.

Harvey, Z.H., Chakravarty, A.K., Futia, R.A., Jarosz, D.F. (2020). A Prion Epigenetic Switch Establishes an Active Chromatin State. *Cell 180*, 928–940 e914.

Harvey, Z.H., Chen, Y., Jarosz, D.F. (2018). Protein-Based Inheritance: Epigenetics beyond the Chromosome. *Mol. Cell. 69*, 195–202.

Jablonka, E., Lamb, M.J. (2005). *Evolution in Four Dimensions. Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. The MIT Press.

Jung, Y.H., Sauria, M.E.G., Lyu, X., Cheema, M.S., Ausio, J., Taylor, J., Corces, V.G. (2017). Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. *Cell reports 18*, 1366–1382.

Klosin, A., Casas, E., Hidalgo-Carcedo, C., Vavouri, T., Lehner. B. (2017). Transgenerational transmission of environmental information in *C. elegans*. *Science 356*, 320–323.

Kriegman, S., Blackiston, D., Levin, M., Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci. USA 117*, 1853–1859.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends Genet. 27*, 323–331.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics 14* 168–178.

Lev, I., Toker, I.A., Mor, Y., Nitzan, A., Weintraub, G., Antonova, O. et al. (2019). Germ Granules Govern Small RNA Inheritance. *Curr. Biol. 29*, 2880–2891 e2884.

Manzano-Lopez, J., Matellan, L., Alvarez-Llamas, A., Blanco-Mira, J.C., Monje-Casas, F. (2019). Asymmetric inheritance of spindle microtubule-organizing centres preserves replicative lifespan. *Nat. Cell. Biol. 21*, 952–965.

Martin, L.B., Ghalambor, C.K., Woods, H.A. (2015). *Integrative Organismal Biology*. Wiley Blackwell.

Metcalfe, N.B., Alonso-Alvarez, C. (2010). Oxidative stress as a life-history constraint: the role of reactive oxygen species in shaping phenotypes from conception to death. *Functional Ecology 24*, 984–996.

Moczek, A.P., Sultan, S., Foster, S., Ledon-Rettig, C. et al. (2011). The role of developmental plasticity in evolutionary innovation. *Proc. Biol. Sci. 278*, 2705–2713.

Murren, C.J., Auld, J.R., Callahan, H., Ghalambor, C.K., Handelsman, C.A. et al. (2015). Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity. *Heredity (Edinb) 115*, 293–301.

Nieto, M.A. (2013). Epithelial plasticity: a common theme in embryonic and cancer cells. *Science 342*, 1234850.

Phillips, P.C., Arnold, S.J. (1999). Hierarchical Comparison of Genetic Variance-Covariance Matrices. I. Using the Flury Hierarchy. *Evolution 53*, 1506–1515.

Posner, R., Toker, I.A., Antonova, O., Star, E. et al. (2019). Neuronal Small RNAs Control Behavior Transgenerationally. *Cell 177*, 1814–1826 e1815.

Roughgarden, J., Gilbert, S.F., Rosenberg, E., Zilber-Rosenberg, I., Lloyd, E.A. (2017). Holobionts as Units of Selection and a Model of Their Population Dynamics and Evolution. *Biological Theory 13*, 44–65.

Saltz, J.B., Hessel, F.C., Kelly, M.W. (2017). Trait Correlations in the Genomics Era. *Trends Ecol. Evol. 32*, 279–290.

Segar, S.T., Fayle, T.M., Srivastava, D.S., Lewinsohn, T.M., Lewis, O.T., Novotny, V., Kitching, R.L., Maunsell, S.C. (2020). The Role of Evolution in Shaping Ecological Networks. *Trends Ecol. Evol. 35*, 454–466.

Shirokawa, Y., Shimada, M. (2016). Cytoplasmic inheritance of parent-offspring cell structure in the clonal diatom *Cyclotella meneghiniana*. *Proc. Biol. Sci. 283*, 20161632

Skvortsova, K., Iovino, N., Bogdanovic, O. (2018). Functions and mechanisms of epigenetic inheritance in animals. *Nat. Rev. Mol. Cell. Biol. 19*, 774–790.

Smith, J.J. (2018). Programmed DNA Elimination: Keeping Germline Genes in their Place. *Curr. Biol. 28*, R601–R603.

Thompson, D.W. (1945). *On growth and form*. University Press, Cambridge.

Topol, E. (2019). *Deep Medicine. How artificial intelligence can make healthcare human again*. Basic Books.

Uller, T., Moczek, A.P., Watson, R.A., Brakefield, P.M., Laland, K.N. (2018). Developmental Bias and Evolution: A Regulatory Network Perspective. *Genetics 209*, 949–966.

Von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. George Braziller, New York.

Waddington, C.H. (1957). *The strategy of the genes: a discussion of some aspects of theoretical biology*. Allen & Unwin.

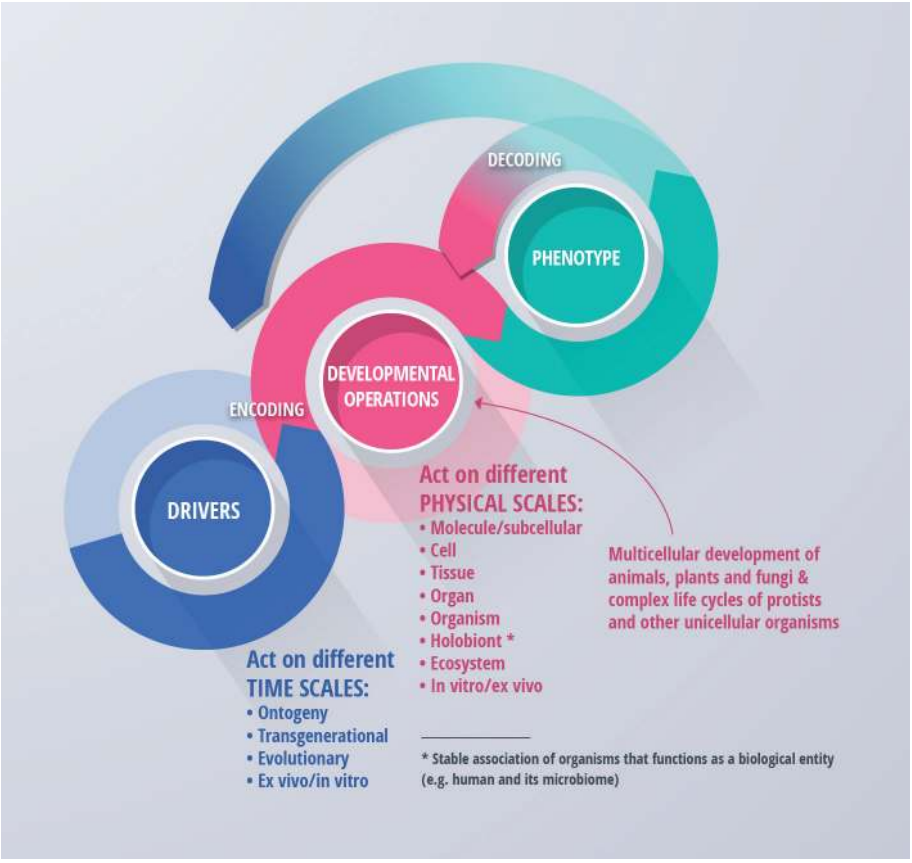
Wagner, G.P., Pavlicev, M., Cheverud, J.M. (2007). The road to modularity. *Nat. Rev. Genet. 8*, 921–931.

Watson, R.A., Mills, R., Buckley, C.L. et al. (2016). Evolutionary Connectionism: Algorithmic Principles Underlying the Evolution of Biological Organisation in Evo-Devo, Evo-Eco and Evolutionary Transitions. *Evol. Biol. 43*, 553–581.

West-Eberhard, M.J. (2005). Developmental plasticity and the origin of species differences. *Proc. Natl. Acad. Sci. USA 102*(Suppl 1), 6543–6549.

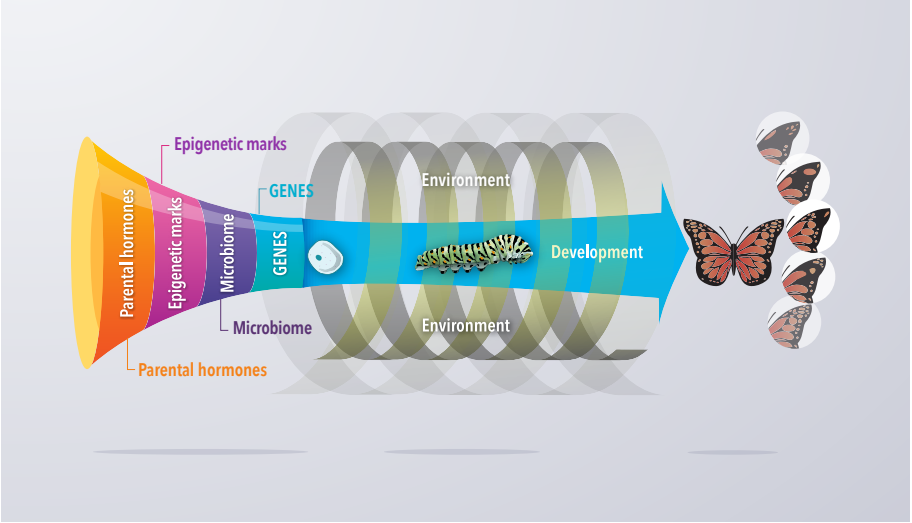
Wiley, C.D. Campisi, J. (2016). From Ancient Pathways to Aging Cells-Connecting Metabolism and Cellular Senescence. *Cell metabolism 23*, 1013–1021.

SUMMARY FOR EXPERTS



Graphic design: MÓVET (www.movet.es)

SUMMARY FOR THE GENERAL PUBLIC



Graphic design: MÓVET (www.movet.es)

CHALLENGE 5

ABSTRACT

Evolutionary Systems Biology (EvoSysBio) is a systemic approach to Biology that aims to generate mechanistic and evolutionary understanding of genotype-phenotype maps at multiple scales. By combining mathematical, molecular and cellular approaches to evolution, EvoSysBio will enhance the predictive and explanatory capacities of the modern evolutionary synthesis.

KEYWORDS

complex systems | evolution

genotype-phenotype map

synthetic biology systems biology theoretical biology

CHALLENGE 5

EVOLUTIONARY SYSTEM BIOLOGY

Coordinators
Sergi Valverde
(IBE)
Saúl Ares
(CNB)

Participants researchers and research centers
Eva Balsa-Canto
(IIM)
Julio R. Banga
(IIM)
Santiago F. Elena
(I2SysBio)
Jaime Iranzo
(CBGP)
Susanna C. Manrubia
(CNB)

EXECUTIVE SUMMARY

"Nothing in Biology makes sense except in the light of Evolution"
Theodosius Dobzhansky (1973).

Evolution is a complex, multilevel process that operates at long time scales and can only be understood from a systems perspective. Though we have a considerable wealth of experimental data, the major challenge is to develop models and theoretical frameworks to understand empirical results and to pose better focused experimental questions. One of the first unsolved questions is how phenotypes arise from genotypes. The full picture is lacking, and modelling is limited to the dynamics of molecules, for instance RNA or protein folding, or to the emergence of simple molecular interactions —as regulatory motifs. A deeper understanding of the structure of genotype spaces might lead, among others, to a quantification of the relative roles played by neutral and adaptive mechanisms. This research program will have to investigate the effects of evolutionary innovation. Can complex biological functions be constructed from previous simpler modules? How do regulatory circuits emerge? What are the limits to the design of robust and portable functional modules? Answers to these questions will assess the validity of reductionist approaches, as opposed to viewing innovation as an emergent phenomenon, arising from network-like distributed properties. In a broader framework, we should be concerned about the mechanistic origin of evolutionary transitions, and on the role played by external forcing versus

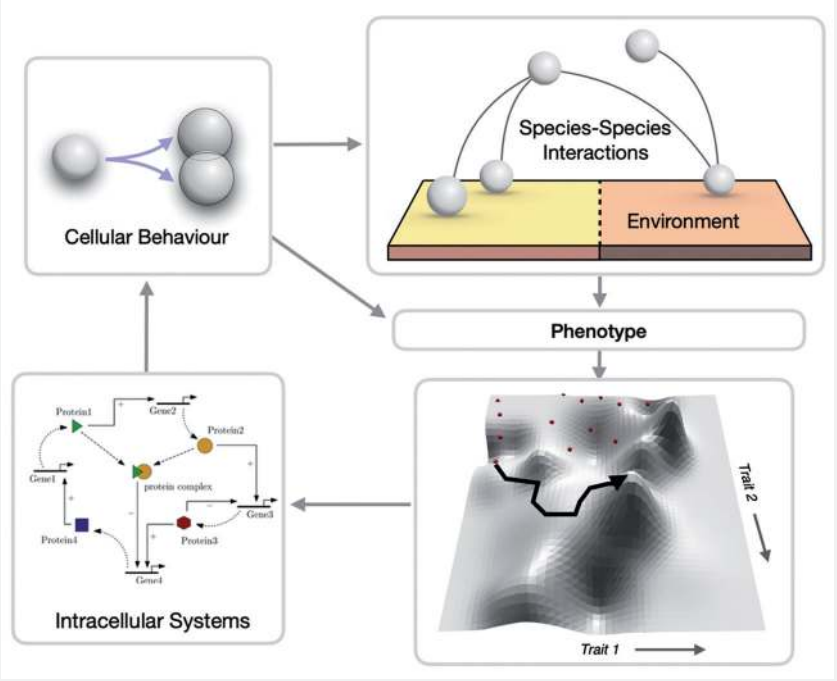
contingent or stochastic phenomena in their generation. Evolution itself can be viewed as a tool for Synthetic Biology, since directed evolutionary selection is a way to attain desired functions. A major goal is to integrate design with a selection-driven exploration of phenotypic spaces. Advances will be conditional on the construction of large evolutionary platforms where experimental evolution informed by design and theory can proceed *en masse*. The exploration of the possible phenotypes of engineered organisms at a large scale, as opposed to simple addition or deletion of a gene, calls for an urgent understanding of the plasticity and adaptability of new organisms and their traits of interest. An eventual commercial use of these organisms demands more research at the frontiers between evolutionary science, climate, and ecology: how will the ecosystem be affected by these organisms, but also, how will the organisms be affected by the ecosystem in a rapidly changing environment?

1. INTRODUCTION AND GENERAL DESCRIPTION

This question is at the core of Evolutionary Systems Biology (EvoSysBio), a new interdisciplinary research area that reconstructs fitness landscapes to predict (1) the fitness of individual organisms under state and environmental changes and (2) the evolutionary trajectories of populations across landscapes (Medina, 2005; Soyer and O'Malley, 2013; Loewe, 2016). EvoSysBio belongs to a broader trend in the biological sciences, i.e., the Modern Evolutionary Synthesis, which has been constructing a coherent view of evolution since the 1920s (Mayr and Provine, 1998). EvoSysBio is an emergent field that takes knowledge currently dispersed over many research fields, from population genetics and biochemistry to ecology. EvoSysBio recognises that evolution is a complex, multilevel process operating at long temporal scales, which can only be understood from a truly systemic perspective. Previous efforts attempted to explain evolutionary processes in the narrower context of individual genes and protein structures. However, complex organisms cannot be reduced to the workings of their components in isolation. EvoSysBio addresses this goal by modelling phenotypes as the outcome of evolving intracellular subsystems, e.g., signalling, regulation and metabolism, which are interacting with each other. EvoSysBio aims to synthesize and ultimately, predict, the complex multi-scale interactions between evolutionary processes and systemic properties (Fig. 1).

EvoSysBio has also been fuelled by the widespread adoption of quantitative and computational methods in the biological sciences. A clear exponent is

FIGURE. 1–The goal of EvoSysBio is to understand and predict genotype-phenotype maps in biological systems. Evolution is a multilevel process operating at a wide range of temporal scales. SysBio addressed this question by focusing on intracellular sub-systems (*e.g.* gene regulatory networks) while overlooking organism evolution. EvoSysBio extends the SysBio vision by including the ecological and evolutionary drivers of organismal complexity. Cellular networks determine species' interactions with their environment and other species. Ecological interactions among species are responsible for the fitness of organisms. Evolutionary processes (*e.g.* neutral drift and adaptation) move populations of organisms on dynamic fitness landscapes by changing the features of intracellular networks.



Systems Biology (SysBio), or the study of how organisms are organized by combining experimental data with mathematical modelling and computer-aided analysis techniques (Ideker et al., 2001; Kitano, 2002). SysBio has traditionally oriented towards the modelling aspects of biological systems ("how" systems are implemented at the molecular level, see Boogerd et al., 2007). Inevitably, overlooking the evolutionary component ("why" biological systems have those features) yields partial explanations of biological functions. The combination of evolutionary and systems biology leads to a better understanding of complex biological features. The evolutionary aspect is what unifies the high diversity of methods employed by EvoSysBio researchers. By

integrating the "how" and "why" questions under the same framework, we can better understand how biological systems work and why they function in that particular way.

EvoSysBio encompasses an integration of theoretical, empirical and computational approaches. A key component is the synthesis of universal "design principles" in biology (Poyatos, 2012; Katsnelson et al., 2018). Achieving this long-term goal crucially depends on identifying (and characterising) the universal principles that hold for large domains of life. For example, we have developed sophisticated empirical and computational tools that enable the detailed study of biological functions. It is however unknown to what extent the observed dynamics and organisation of any particular species, *e.g.*, how signalling networks enable chemotaxis *in Escherichia coli* or the dynamics of osmoregulation in *Saccharomyces cerevisiae* (Alon et al., 1999; Klipp et al., 2005), can also explain the physiological responses in other species. Ignoring intermediate states of population-level variation cannot fully explain the evolution of biological complexity (Lynch, 2007). Complex features in biological systems can be adaptive (Kashtan and Alon, 2005), neutral (Wagner, 2003; Solé and Valverde, 2006) or a mix between functional and non-adaptive processes (Wagner, 2008). In this context, evolutionary methods can support generalization of system properties, e.g., network patterns, beyond any specific biological model.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Evolution itself can be viewed as a tool for Synthetic Biology (SynBio) since directed evolutionary selection is a way to generate and optimize desired functions (Arnold, 2018). A major goal is to integrate design with a selection-driven exploration of phenotypic spaces. Advances will be conditional on the construction of large evolutionary platforms where experimental evolution informed by design and theory can proceed *en masse*. The exploration of the possible phenotypes of engineered organisms at a large scale, as opposed to simple addition or deletion of a gene, calls for a deep and reliable understanding of the plasticity and adaptability of new organisms and their traits of interest. An eventual commercial use of these organisms demands more research at the frontiers between evolutionary science, climatology, agriculture, and ecology: how will the ecosystem be affected by these organisms, but also, the analysis and prediction of the emergence of new pathogens and their potential epidemic spread.

2.1 Modelling and Computer-Aided Simulation of Biological Systems

The interplay of mathematical modelling with experiments is one of the central elements in SysBio. Model-building is therefore a key step, involving the reverse engineering (i.e., inference and identification) of equations that describe the biosystem, and their calibration to existing data (Klipp et al., 2005). Typically, these models can be either mechanistic (Stelling, 2004) or data-driven (as in machine learning and statistics; Kell & Oliver, 2004). Although the latter can be useful in many applications, mechanistic models generally provide a better framework to distil knowledge and understanding from data. Despite the many advances in model building in SysBio during the last two decades, the evolutionary perspective is absent in most of these models. Thus, a general theoretical and computational framework for multiscale modelling in EvoSysBio remains as a core objective.

The real power of mathematical models is unleashed when exploited via methods for their computer aided simulation, analysis, optimization and control (Wolkenhauser and Mesarović, 2005; Sontag, 2005). Although these tools have been widely used in areas like biosystems and bioprocess engineering (Park et al., 2008), we are still missing the evolutionary component integrated across different scales. Thus, another core objective would be to exploit EvoSysBio models as the kernel of process systems engineering in the bio-industries (e.g., agri-food and industrial biotechnology), particularly in areas like metabolic engineering. A possible route towards such objective could be through the integration of evolutionary game-theoretic approaches with multiscale modelling. The same approach can be adapted and extended to key problems in areas like environmental engineering (e.g., dynamics of microbial communities in bioremediation) and medicine (e.g., cancer evolution, or multiple microbial infections).

2.2 A first step towards understanding ecological complexity: modelling complex microbial communities

The last decades have witnessed the development of modelling approaches to describe cellular communities: from continuous to individual-based models of tissues and biofilms; from phenomenological ecological models to more mechanistic genome-scale approximations (Zomorrodi and Segrè, 2016). Still, their scope mostly restricts to "simple" systems (single or co-cultures), under steady-state or controlled extracellular environmental conditions. Also, evolutionary aspects are barely considered. The integration of multi-species multi-scale dynamic models incorporating both ecological and evolutionary

mechanisms (Valverde et al., 2020) are required to fully realize the potential of multicellular systems in answering biological questions and applications.

Model-based optimally designed microbial mixed-cultures will enable metabolically complex tasks by the division of labour and/or experimentally evolved cooperation (Harcombe, 2010; Hays et al., 2015; Thommes et al., 2019). As a result, bioprocesses will be more efficient, productive and stable than the current single-species bioprocesses and will allow for a broader range of products.

Similar approaches could be used to engineer the microbiome to generate potential therapies against metabolic, inflammatory, and immunological diseases, among others. The evolutionary dimension becomes fundamental to understand the dynamics of communities in which microbes and viruses coexist. In such systems, rapid co-evolution of viruses and microbial hosts (e.g. via acquisition of spacers in CRISPR arrays) cannot be disentangled from ecological processes, which has direct implications for the development of phage therapies and other virus-based microbiome engineering strategies.

2.3 Synthetic Biology: human design of non-biological constructs

In SynBio, we aim at using engineering principles of rational design to modify organisms, or to build new bio-artifacts (Purnick & Weiss, 2009; Cameron et al., 2014; Schwille et al., 2018). As in SysBio, SynBio can also exploit model-based approaches to guide the design, analysis, optimization and control of genetic systems (Marchisio et al., 2009). In addition to recent progress in genetic parts standardization and characterization (McLaughlin et al., 2018), in recent years we have also witnessed significant advances in the application of microfluidics, machine learning and automation to SynBio (Melin and Quake, 2007; Nielsen et al., 2016; Aoki et al., 2019; Carbonell et al., 2019). We are also already close to having programming languages to design computational circuits in living cells (Nielsen et al., 2016). These approaches open up new avenues for important applications, including biosensors (Gupta et al., 2019), biotherapeutics (Ozdemir et al., 2018), metabolic engineering (Keasling, 2012), biomanufacturing (Chen et al., 2020), and bioremediation (de Lorenzo et al., 2018).

However, with the exception of efforts in the field of directed evolution (Arnold, 2018), we are still lacking a truly evolutionary approach to SynBio. In particular, although computer-aided methods can help us in the design of synthetic parts in a similar way, as done in e.g. electronics, the unpredictability

and complexity created by the variability and evolvability of cell behaviour (Kwok, 2010) needs to be taken into consideration in order to achieve the envisioned engineering of biology. Finally, cell-free approaches (Hodgman and Jewett, 2012) will help us to improve our understanding of the design of evolved natural biosystems, and also to enable a novel kind of biomanufacturing with more freedom of design and improved control.

3. KEY CHALLENGING POINTS

3.1 Multiscale theoretical framework

Though we have a considerable wealth of experimental data, a main challenge is to develop models and theoretical frameworks explaining these results and pose better-focused experimental questions. We currently lack a multi-scale modelling framework that captures the essential features of biological systems, from genome to metabolism. At the same time, we need models that are simple enough to provide useful answers and insights. An essential part of future developments in EvoSysBio is the construction of a hierarchy of *in silico* tools and computational models that can guide experimental studies.

3.2 Evolution of Novelties

Ecosystems are highly nonlinear, complex dynamical systems (May and Leonard, 1975; Clark and Luis, 2020). Competition, cooperation, or victim-exploiter dynamics, are density-dependent interactions that induce non-linear effects in population dynamics. This is particularly relevant when dealing with ecosystem's responses to external perturbations, in particular, of anthropogenic origins (Lade et al., 2020). It has been conjectured these nonlinearities give rise to sharp shifts in the ecosystem composition, also known as "tipping points", which are becoming an important subject or research in ecology (Berdugo et al., 2020). Moreover, in SysBio, a tipping point driving population to extinction has been reported in yeast (Dai et al., 2012).

Sudden changes could be associated with large modifications. Interestingly, studies suggest that smooth alterations to the environment might also be responsible for such drastic shifts. Theoretical analyses of dynamical systems have interpreted these transitions as jumps along evolutionary optimization, where long periods without changes evidence the presence of high fitness barriers that the population cannot easily overcome (Huynen et al., 1996). This research program will have to investigate the causes of these sudden shifts in the genetic

composition of populations. Explaining the mechanistic origin of evolutionary transitions, and the role played by external forcing versus contingent or stochastic phenomena in their generation, will help understanding the effects of (and hopefully predict) evolutionary novelties (Wagner and Lynch, 2010).

3.3 Modelling and simulation of biological networks

Biological systems are composed of genes encoding the molecular machinery that provides the basic functions of life. Biological functions can be rarely reduced to specific components in isolation. Instead, they are the outcome of multiple interactions between different components. For example, networks of regulatory interactions specify how genes are expressed, with both operating on multiple, hierarchical levels of organization. Quantifying the structural features of biological and artificial systems has been the target of Network Science (a branch of Statistical Physics and Complex Systems developed during the last 25 years). SysBio has relied on the results of Network Science but we still do not understand the origin of structural regularities in biological networks, and how they shape function and evolution. This is particularly relevant in one of the main open problems in SysBio, namely, how to define a robust approach to reverse engineering and systems identification in biological systems (Villaverde and Banga, 2014). What is needed is the cooperation of network scientists with EvoSysBio researchers for developing a rigorous, biologically-realistic, evolutionary theory of biological complexity.

Genome-scale models and optimality principles have shown their potential for application in generating mappings genome-phenotype in EvoSysBio, for example in the prediction of phenotypic outcomes of short-term adaptive evolution or in the analysis of viability of mutant strains (Palsson, 2015). However, developing its full potential for EvoSysBio requires the predictive capabilities of these models to be improved in different ways. Clearly, the assumption of optimal growth in genome-scale metabolic models is not suitable for mutants and not valid in many (time-varying) environmental conditions. Thus, alternative evolutionary objectives or trade-offs or game-theoretical approaches are to be explored. Expansions including protein structures, integrated models of metabolism and protein expression (O'Brien et al., 2015) as well as hybrid modelling frameworks that incorporate explicit information on reaction rates are yet in progress. GSMs have also been expanded from single populations of cells to simple microbial communities (Harcombe et al., 2014), further developments are required to describe complex multi-species populations and changing environments (in time and space).

3.4. Structure of fitness landscapes

To predict the evolutionary path from one species to another, we need first to understand the underlying configuration space, *i.e.*, we need a 'map' where we can locate every possible intermediate species, for each environmental condition (see Appendix C). Unfortunately, fitness landscapes are huge, they live in spaces of very high dimensionality, and are difficult to visualize. The traditional metaphor of the fitness landscape is based on a gradualistic perspective about organismal change and evolution. Evolution has been described as the diffusion of populations on a relatively smooth landscape, always climbing towards regions of high fitness, which are eventually trapped in mountain peaks possibly separated by deep valleys of lower fitness (Wright, 1931). Due to conceptual advances in our understanding about the molecular structure of populations, we now know this picture is clearly incomplete and misses important ingredients. The structure of fitness landscapes is not a smooth and continuous surface, but rugged (Kauffman and Levin, 1987). Instead, the structure of the space of genotypes is a network of networks (or multilayer network) whose nodes (genomes) are mutually accessible through mutations. In the landscape, we can find regions with sharp discontinuities (signalling the presence of lethal and deleterious mutations) and long-range connections between distant regions of the landscape. EvoSysBio will help us understand (and characterise) landscape properties to obtain realistic models of adaptive fitness landscapes (in particular, through the development of models of genotype-phenotype maps at different levels of the biological hierarchy, see below), and validate these models with the reconstruction of empirical landscapes using network tools.

3.5. New constructions of genotype-phenotype (GP) maps

The mapping function between the instructions encoded in the genotypes and the structures and functions of the phenotypes is fundamental to every aspect of Biology. Given its importance, GP maps have attracted a lot of attention both from theoreticians and experimentalists (de Visser and Krug, 2014; Ahnert, 2017). As a result of these exercises, a number of universal properties shared by most GP maps have been derived. These properties are structural in the sense that they depend on the distribution of phenotypes across the network of genotypes. These properties include redundancy of genotypes (many encode for the same phenotype), a highly non-uniform distribution of the number of genotypes per phenotype resulting in a high phenotypic robustness, and the capacity to explore the landscape efficiently, reaching very distant phenotypes throughout a quite limited number of genotypic changes.

Despite these advances, we still miss a coherent theoretical description of the GP maps that explain why all these properties emerge (Manrubia et al., 2020). Novel approaches such as modelling GP maps as a network of networks in which different nodes in the genotypic network level correspond to networks of neutral genotypes and these nodes map into similar networks in the phenotypic space (Aguirre et al., 2018), or the seascape metaphor, in which the landscape topology fluctuates as a result of changes in the biotic and abiotic composition (Mustonen and Lässig, 2009) seems promising. Still open questions exist. To name two of the most relevant ones: a) whether the GP map as an object evolves (Manrubia et al., 2020) and whether it does so by natural selection or by neutral processes; b) how the GP map accommodates changes in genome size; in other words what are the consequences of making the genotypic network (of networks) growing or shrinking with evolutionary time? (see also Challenge 4).

CHALLENGE 5 REFERENCES

Aguirre, J. et al. (2018). On the networked architecture of genotype spaces and its critical effects on molecular evolution. *Open Biology 8*, 180069.

Alon, U. et al. (1999). Robustness in bacterial chemotaxis. *Nature 397* 168–171.

Aoki, S.K. et al. (2019). A universal biomolecular integral feedback controller for robust perfect adaptation. *Nature 570* 533–537.

Ahnert, S.E. (2017). Structural properties of genotype-phenotype maps. *Journal of the Royal Society Interface 14*, 20170275.

Arnold, F.H. (2018). Directed evolution: bringing new chemistry to life. *Angewandte Chemie International Edition 57*, 4143–4148.

Berdugo, M. et al. (2020). Global ecosystem thresholds driven by aridity. *Science 367*, 787–790.

Boogerd, F.C. et al. (eds.) (2007). *Systems Biology: Philosophical Foundations*. Amsterdam: Elsevier

Cameron, D.E., Bashor, C.J., Collins, J.J. (2014). A brief history of synthetic biology. *Nature Reviews Microbiology 12*, 381–390.

Carbonell, P., Radivojevic, T., García Martín, H. (2019). Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synthetic Biology 8*, 1474–1477.

Chen, Y. et al. (2020). Systems and synthetic biology tools for advanced bioproduction hosts". *Current Opinion in Biotechnology 64*, 101–109.

Clark, T.J., Luis, A.D. (2020). Nonlinear population dynamics are ubiquitous in animals. *Nature Ecology and Evolution 4*, 75–81.

Dai, L. et al. (2012). Generic indicators for loss of resilience before a tip- ping point leading to population collapse. *Science 336*, 1175–1177.

De Lorenzo, V. et al. (2018). The power of synthetic biology for bioproduction, remediation and pollution control. *EMBO Reports 19*, e45658.

De Visser, J.A.G.M., Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Review Genetics 15*, 480–490.

Gorter, F.A., Manhart, M., Ackermannm, M. (2020). Understanding the evolution of interspecies interactions in microbial communities. *Philosophical Transactions of the Royal Society B: Biological Sciences 375*, 20190256.

Gupta, N., et al. (2019). Cell-based biosensors: Recent trends, challenges and future perspectives. *Biosensors and Bioelectronics 141*, 111435.

Harcombe, W.R. (2010). Novel cooperation experimentally evolved between species. *Evolution 64*, 2166–2172.

Harcombe, W.R. et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports 7*, 1104–1115.

Hays, S.G. et al. (2015). Better together: engineering and application of microbial symbioses. *Current Opinion in Biotechnology 36*, 40–49.

Hodgman, C.E., Jewett, M.C. (2012). Cell-free synthetic biology: thinking outside the cell. *Metabolic Engineering 14*, 261–269.

Huynen, M.A., Stadler, P.F., Fontana, W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences of the USA 93*, 397–401.

Ideker, T., Galitski, T., Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Reviews in Genomics and Human Genetics 2*, 343–372.

Kashtan, N., Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the USA 102*, 13773–13778.

Katsnelson, M.I., Wolf, Y.I., Koonin, E.V. (2018). Towards physical principles of biological evolution. *Physica Scripta 93*, 043001.

Kauffman, S., Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology 128*, 11–45.

Keasling, J.D. (2012). Synthetic biology and the development of tools for metabolic engineering. *Metabolic Engineering 14*, 189–195.

Kell, D.B., Oliver, S.G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays 26*, 99–105.

Kitano, H. (2002). Systems Biology: A Brief Overview. *Science 295*, 1662–1664.

Klipp, E. et al. (2005). *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons.

Klipp, E. et al. (2005). Integrative model of the response of yeast to osmotic shock. *Nature Biotechnology 23*, 975–982.

Kwok, R. (2010). Five hard truths for synthetic biology: can engineering approaches tame the complexity of living systems? *Nature 463*, 288–291.

Lade, S.J. et al. (2020). Human impacts on planetary boundaries amplified by Earth system interactions. *Nature Sustainability 3*, 119–128.

Loewe, L. (2016). Systems in evolutionary systems biology. *Encyclopedia of Evolutionary Biology 4*, 297–318.

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics 8*, 803–813.

Manrubia, S. et al. (2020). From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *arXiv*, 2002.00363v1.

Marchisio, M.A., Stelling, J. (2009). Computational design tools for synthetic biology. *Current Opinion in Biotechnology 20*, 479–485.

May, R.M, Leonard, W.J. (1975). Nonlinear aspects of competition between three species. *SIAM Journal of Applied Mathematics* 29, 243–253.

Mayr, E., Provine, W.B. (eds) (1998). *The evolutionary synthesis: perspectives on the unification of Biology*. Harvard University Press

McLaughlin, J.A. et al. (2018). SynBioHub: a standards-enabled design repository for synthetic biology. *ACS Synthetic Biology 7* 682–688.

Medina, M. (2005). Genomes, phylogeny, and evolutionary systems biology. *Proceedings of the National Academy of Sciences of the USA 102*, 6630–6635.

Melin, J., Quake, S.R. (2007). Microfluidic large-scale integration: the evolution of design rules for biological automation. *Annual Reviews Biophysics and Biomolecular Structure 36*, 213–231.

Mustonen, V., Lässig, M. (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics 25*, 111–119.

Nielsen, A.A. et al. (2016). Genetic circuit design automation. *Science 352*, aac7341.

O'Brien. E.J., Monk, J.M., Palsson, B.O. (2015). Using genome-scale models to predict biological capabilities. *Cell 161*, 971–987.

Ozdemir, T. et al. (2018). Synthetic biology and engineered live biotherapeutics: toward increasing system complexity. *Cell Systems 7*, 5–16.

Park, J.H., et al. (2008). Application of systems biology for bioprocess development. *Trends in Biotechnology 26*, 404–412.

Poyatos J.F. (2012). On the search for design principles in biological systems. *Advances in experimental medicine and biology 751*, 183–193. https://doi.org/10.1007/978-1-4614-3567-9_9.

Purnick, P.E., Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology 10*, 410–422.

Schwille, P., et al. (2018). MaxSynBio: avenues towards creating cells from the bottom up. *Angewandte Chemie International Edition 57*, 13382–13392.

Solé, R.V., Valverde, S. (2006). Are network motifs the spandrels of cellular complexity? *Trends in Ecology and Evolution 21*, 419–422.

Sontag, E.D. (2005). Molecular systems biology and control. *European Journal of Control 11*, 396–435.

Soyer, S.O., O'Malley, M.A. (2013). Evolutionary systems biology: what it is and why it matters. *BioEssays 35*, 696–705.

Stelling, J. (2004). Mathematical models in microbial systems biology. *Current Opinion in Microbiology 7*, 513–518.

Thommes, M., et al. (2019). Designing metabolic division of labor in microbial communities. *mSystems 4*, e00263-18.

Valverde, S., Vidiella, B., Montañez, R., Sacristan, S., Fraile, A., and García-Arenal, F. (2020). Coexistence of nestedness and modularity in host-pathogen infection networks. *Nature Ecology and Evolution 4*, 568–577.

Villaverde, A.F., Banga, J.R. (2014). Reverse engineering and identification in systems biology: Strategies, perspectives and challenges. *Journal of the Royal Society Interface 11*, 20130505.

Wagner, A. (2003). Does selection mold molecular networks? *Science Signaling 202*, pe41.

Wagner A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature Review Genetics 9*, 965–974.

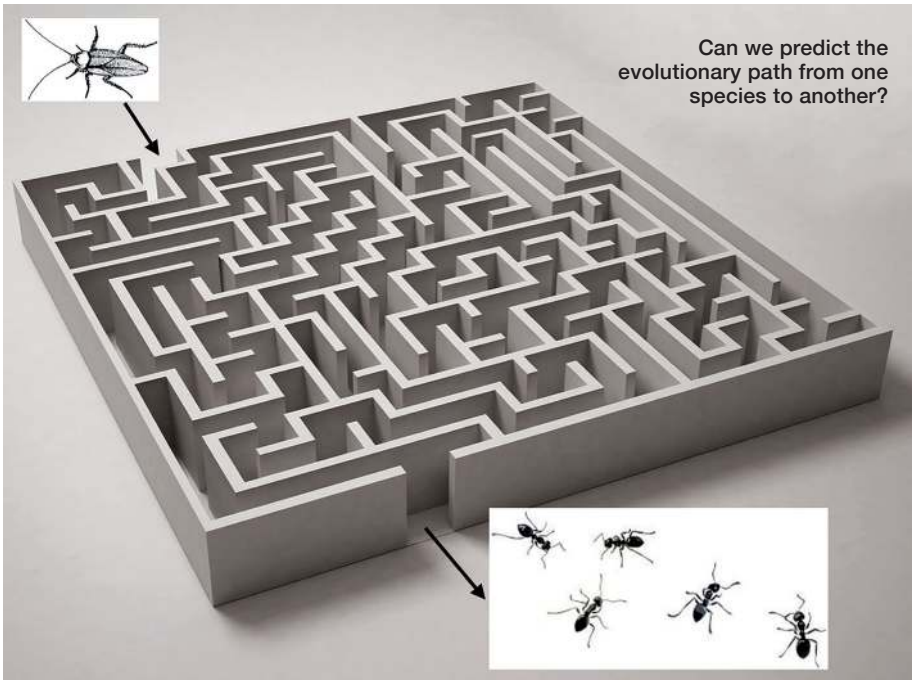
Wagner, G. P. and Lynch, J. L. (2010). *Evolutionary novelties 20*, R48–R52.

Wolkenhauer, O., Mesarović, M. (2005). Feedback dynamics and cell function: why systems biology is called systems biology. *Molecular BioSystems 1*, 14–16.

Wright, S. (1931). Evolution in mendelian populations. *Genetics 16*, 97–159.

Zomorrodi, A.R., Segrè, D. (2016). Synthetic ecology of microbes: mathematical models and applications. *Journal of Molecular Biology 428*, 837–861.

SUMMARY FOR EXPERTS AND GENERAL PUBLIC



CHALLENGE 6

ABSTRACT

One of the greatest challenges in the study of Human Evolution is to understand the social and biological adaptive processes that took place along our evolutionary history. This challenge has molecular, genetic, behavioural, social and anatomic morphological dimensions and requires new multidisciplinary and technological approaches.

KEYWORDS

paleogenomics | paleoproteomics | paleoepigenomics | genome/phenome | big data | organoids | brain | digital anatomy | 3D imaging | virtual databases | paleoanthropology | biotic/abiotic factors | taphonomy | archaeology of human evolution

CHALLENGE 6

SOCIAL AND HUMAN EVOLUTION

Coordinators

Carles Lalueza-Fox
(IBE)

Ignacio de la Torre
(IH-CCHS)

Participating researchers and research centers

Markus Bastir
(MNCN)

Victor Borrell
(INA)

Pedro Díaz del Río
(IH-CCHS)

Arcadi Navarro
(IBE)

Tomàs Marquès-Bonet
(IBE)

Juan Manuel Vicent
(IH-CCHS)

EXECUTIVE SUMMARY

Understanding the social and biological evolution of the human lineage is a complex enterprise that will require the integration of different scientific fields and of their most innovative and state-of-the-art approaches. In a deeper timeframe, the availability of full-genomes from multiple species that provides information about the phylogeny of our lineage allows for detailed studies on when and how some of the human-defining traits appeared during evolution. The application of -omics techniques, both using extant and extinct species (paleogenomics, paleoepigenomics, the emerging paleoproteomics, together with comparative genomics and evolutionary genomics data) to the study of the past now enables researchers to tackle a range of questions that previously were almost exclusively studied by disciplines in the Humanities, such as History and Archaeology when investigating ancient migrations. The multidisciplinary efforts involved in these challenges will also require the secure management of large amounts of information that are being generated (known as Big Data), ranging from currently expanding bio-banks to the results of particular research projects that need to be made open according to FAIR principles (Findable, Accessible, Interoperable, Reusable). At the same time, new ethical issues will emerge and will need to be addressed.

As for the study of human cultural and behavioural evolution, primarily addressed through archaeological studies, disciplinary challenges should

attempt to redefine the role of technology in shaping past societies but also in transforming the environment, as well as the nature of interactions between biotic and abiotic agents throughout the evolution of our Genus. Such studies must embrace big-data approaches and be international in scope, in order to supersede local perspectives. They should also aim at the identification of patterns, rather than events, in the course of the behavioural and cultural evolution of our species. Thus, any characterization of human culture will necessarily include technology as a defining element and a key concept to explain social evolution in our ancestral lineage and historical trajectory. Despite its central role in explanatory mechanisms of human biological evolution and social change, the mere definition of technology, its potential uniqueness in our Genus and its part as engine for cultural transformations are all concepts currently questioned and constitute pivotal challenges in disciplines concerned with the reconstruction of past social evolution.

One of the top questions in the study of past human behaviour is to what extent biotic versus abiotic evolutionary factors dictate divergent trajectories in the human past. From the role of climate in early human adaptations to its relevance in the emergence of food production, the secular interest in the influence of abiotic causes in social evolution is now leading to a shift in perspective. The new grand challenges in social evolution should highlight, for example, the importance of mammal community structure in the shaping of early human behaviour, or the impact of human actions over fauna and flora and, in recent times, even over the climate. Primary datasets and metadata sources should be integrated in large bodies of information such as Spatial Data Infrastructures amenable for spatial analysis and accessible to inter-connected collaborators from multiple disciplinary fields and institutions, to which continental- scale perspectives can be applied.

1. INTRODUCTION AND GENERAL DESCRIPTION

The study of human evolution since the divergence from the lineage of our closest living relative, the chimpanzee, has been mainly based, until recently, on the analysis of archaeological and paleontological evidences. However, after the sequencing of the human and the chimpanzee genomes (2001-2006), little progress has been made in understanding the functional bases of genetic differences found between both species. This can be attributed in part to the six million years elapsed on each evolutionary lineage as well as to the complexity of the processes involved and the inherent limitations of the

fossil record. It seems clear that a new angle on this scientific problem, coupled with new methodological and technical approaches, is needed.

Fast-evolving digital techniques have provided a momentum for disciplines such as paleontology and archaeology, and novel spatial analysis perspectives and big data initiatives enable large-scale studies of human evolutionary patterns under new theoretical premises. These advances in well-established disciplines involved in human evolutionary studies, alongside the revolutionary perspectives introduced by disciplines derived from genetics, present a unique opportunity to promote a true multidisciplinary view of the study of the past, for which the CSIC should take a leading role in the international arena.

2. IMPACT IN BASIC SCIENCE AND POTENTIAL APPLICATIONS

Results obtained through characterising the genetics of archaic hominins -including Neanderthals and Denisovans- for reconstructing past migratory and admixture events also include evidence about past demography or social structure that can be correlated with morphological and archaeological observations drawn from the fossil record. Of course, this approach requires a better integration with other disciplines and a true multidisciplinary view of the study of the past. The application of different "omic" approaches, some of them still in early developing phases such as the retrieval of ancient peptides ("paleoproteomics"), bears promises of boosting significantly the human evolutionary field in the next years.

In parallel, these emerging ancient genetic datasets will allow the direct screening of adaptive genetic variants that have increased along time to current frequencies. Thus, selective processes that could have biomedical implications could be studied in detail in the near future, directly in space and time. These genomes however are indicative of selective events that took place in the last hundreds of thousands of years of human evolution.

It has been proved that some evolutionarily relevant traits (e.g., sense of equity or problem-solving abilities) are not exclusive of our species but may have undergone convergent evolution in several occasions. Importantly, some traits of biomedical and ecological relevance, such as brain size and function, lifespans, diet or fertility, are particularly well suited for a phylogenomic approach. Humans, for instance, differ markedly from most mammals by a number of features, such as having the largest brain in nature (relative to body size)

with remarkably increased area due to its specific folding. How this complexity and diversity is achieved remains largely unknown and can only be thoroughly approached if deep phylogenetic studies on encephalization are carried out across the whole mammalian lineage, including the study of any relevant adaptive processes. These questions, again, generate the opportunity of evolutionary studies that both illuminate human evolution and have relevant biomedical impact.

In parallel to the advances in the "omic" approaches and its multidisciplinary requirements, Evolutionary Archaeology is at the forefront of disciplinary advances in Social Sciences due to its continuous interaction with Natural Sciences, and its reliance on the progress of techniques, from Physics and Chemistry, to develop new analytical approaches in dating and the characterization of archaeological assemblages.

A compendium of the 21st century grand challenges for Archaeology (Kintigh et al., 2014) recognises that key questions of the discipline are no longer concerned with the earliest, the longest, or the otherwise unique. If anything, the ever-increasing age depth for the emergence and dispersal of Prehistoric humans only exacerbates the complexities of the early archaeological record. As the gaps between known sites become wider in time and space, it becomes crucial to go beyond first-appearance discussions and focus on the biogeographic and cultural patterns involved. Substantial fieldwork efforts in the last few decades have produced extensive archaeological datasets, and the time is now for large-scale synthetic research in cultural evolution (Brewer et al., 2017).

As for transfer of results and potential applications: 1) advances in our understanding of the nature of mobility and migrations in Prehistory shall provide precious data to challenge simplistic interpretations of migrations in the past that are sometimes (mis) used to construct artificial narratives of national identities. 2) Spatial Data Infrastructures are the future of data sharing in social sciences -as it is now common in biological sciences- and will serve to both boost international scientific collaboration, and promote outreach of academic results to the public. 3) A dedicated field unit may promote collaboration with the private sector through its participation in civil projects involving archaeological sites relevant to this Challenge.

3. KEY CHALLENGING POINTS

3.1 Application of "multi-omics" to the recent past

Paleogenomics. The emergence of human paleogenomics or archaeogenomics (the retrieval and analysis of past human genomes), triggered by the development of next-generation sequencing (NGS) technologies after 2005, has provided a powerful, new source of information that can be used to test previous hypotheses regarding human evolution and past migrations. In the field of human history, paleogenomics has provided answers to long-standing debates about the evolutionary relationships of our species and some ancient lineages such as Neandertals and Denisovans, as well as to inform on other important facets of the human experience, such as past migrations, social structure or sex-biased population movements. The interaction between archaeologists, anthropologists, and researchers from the emerging field of paleogenomics has traditionally been plagued by a lack of collaborative efforts. These disagreements between different fields can be partially explained by some limitations associated with the genetic markers typically employed some years ago (e.g., mitochondrial DNA) prior to the advent of NGS technologies, but also by the way geneticists have approached the study of the archaeological material. Therefore, the possibility of working with ancient genomes is an opportunity for building a more integrative approach to the study of the past involving different research institutes focussed on different disciplines (Racimo et al., 2020).

Improvements in bioinformatics and population genetic inference have served to extract invaluable information from these genomes, including patterns of population growth and contraction, interbreeding between distantly related groups and evidence for natural selection operating on phenotypically important loci. Some of the traditional ancient DNA limitations have largely been solved by the application of NGS technologies, while others have emerged. Most importantly, NGS does not rely on targeted polymerase chain reaction (PCR) amplification of the short ancient molecules using primers. Endogenous DNA tends to degrade to shorter fragments with time and environmental conditions (mainly temperature, but also humidity and soil pH), this possibility means that there is now access to a much larger fraction of endogenous DNA than with the use of PCR. Another key advantage of NGS is that it allows the use of degradation patterns at the ends of the reads (an increase of C to T substitution at the 5'end and of G to A substitutions at the 3' end) to discriminate between modern DNA contaminants and endogenous DNA. With enough

coverage and with imputation tools, it is now possible to attribute specific sequences to either damage or contamination. Thus, even if contamination is not controlled *a priori* in a particular site, it can be monitored *a posteriori*, obtaining estimates from the high coverage mtDNA, or from heterogeneities in the sexual chromosomes. Contamination estimates below 1-2% would be acceptable in these ancient genomes.

However, a critical feature now is the efficiency of a particular sample in the shotgun sequencing (that is, the ratio of endogenous DNA vs. the environmental contaminant DNA). This was not important with the PCR based approach, because specific primers were designed and used to retrieve DNA sequences of interest, irrespective of the environmental content of the sample. Summarising data from different studies, shows that conventional bone or teeth samples from a temperate European environment usually have efficiency below or rarely over 5%-10%. In southern Europe, this figure usually goes between 0.1 and <2%. Samples with an endogenous content higher than 50%, such as the Tyrolean Ice Man derived from unusual archaeological contexts and unique taphonomic conditions but likely represent a rarity in prehistoric Europe. Therefore, an adequate sampling strategy is crucial for these new paleogenomic approaches. A further improvement has been the recent discovery that petrous bones (a very hard region from the temporal bone) are the best DNA-containers.

Despite the aforementioned methodological problems, whole genome sequences are not required for most population genetic analyses and an array of several hundreds of thousands of informative single nucleotide positions or SNPs can allow high-resolution analyses. It is increasingly clear that most of the ancient samples cannot be shotgun-sequenced, but nevertheless have enough DNA content to be captured for SNP-genotyping. Although it is difficult to work out precise figures, it might be that around one sample in ten or twenty has the potentiality of being shotgun-sequenced, while about one in two or three could be SNP genotyped at the same time frame.

Despite technical limitations, paleogenomic studies are unravelling the complex evolutionary patterns of the human lineages, showing multiple admixture events in different moments and regions, as well as providing information on adaptations to environmental conditions, past migrations, demographic trends and social structures. It is likely that these studies will become mainstream in maybe 10 more years but right now they still have a great scientific potentiality.

Paleoproteomics. Paleogenomics represents a valuable resource of information from extinct species and past populations. Nevertheless, there is a clear time frame beyond which no DNA is expected to survive. This limit is currently unknown –partly because it is very dependent on environmental thermal conditions- but it does not seem to go beyond half million years in temperate conditions or maybe up to one million years in extremely favourable –polar-like- conditions. Right now, the absolute record in temperate conditions corresponds to the Sima de los Huesos site in Atapuerca (dated to ca 430,000 years ago) whereas that in favourable conditions to a horse bone found in Alaska and dated to ca 700,000 years ago. In warm environments, including tropical or subtropical, DNA is only expected to survive few tens of thousands years.

In contrast, ancient proteins represent a more durable source of molecular information. Some recent studies have reported survival of proteins up to the range of 2 million years, especially in dental enamel that seems to preserve around half dozen proteins involved precisely in the formation of this tissue. Although these proteins are, by definition, very stable in evolution, these very ancient peptide sequences can be reliably used to build molecular trees that reconstruct the evolutionary relationships between extant and extinct species.

Ultimately, paleoprotein sequencing can push molecular-based evolutionary reconstructions much further back in time than ancient DNA analysis. This is especially interesting because it can provide phylogenetic information in the critical time frame around 2-3 million yeas (at least) that saw the emergence of our own genus, *Homo*, and its first dispersal out of Africa, as well as the appearance in the fossil record of some Australopithecus lineages, including the so-called "robust" forms. Right now it has been successfully applied to an extinct rhino (*Stephanorhinus)* from Dmanisi site in Georgia (dated to 1.77 million years ago), to the extinct *Gigantopithecus* from China (dated to 1.9 million years) (Welker et al., 2019) and to the paradigmatic remains of *Homo antecessor* from Atapuerca (dated to 800,000 years ago) (Welker et al., 2020). If the technique can be further pushed back (the temporal limits are currently unknown), it is possible that ancient hominins -and maybe Miocene hominids- could also be analysed.

It is therefore likely that in the next 10-15 years the field of the human evolution (and in general, animal paleontology) will be greatly impacted by paleoproteomics developments, which will also create the need for qualified

professionals with training combining computational and phylogenetic analysis, mass spectrometry expertise, as well as paleoantropological background.

Paleoepigenomics. Besides traditional human-ape or human-Neandertal (or Denisovan) comparative genomics, an additional source of information to understand what may constitute the unique human features is the study of the paleoepigenomes. It can be hypothesized that, besides the genomic sequence comparison, many human-specific genomic traits with phenotypic effect can be identified by comparing gene regulatory programs, as differences in gene regulation are widely considered to be a major evolutionary force explaining phenotypic differences between closely related species. Consequently, in order to understand the phenotypic differences between various human groups, there is a critical need to understand the ways they differ in their regulatory programs.

DNA methylation occurs almost exclusively in cytosines in the context of CpG dinucleotides. Enhancer DNA methylation is a key hallmark of gene activity in mammals and is inversely correlated with expression level. This information can be extracted from high-quality ancient genomes -that is, high coverage genomes, although it is likely that future algorithms will allow to work also with low-coverage genomes- and could help understand how these genomes were regulated. The approach is based on natural degradation processes of ancient DNA, in which methylated and unmethylated cytosines deaminate with time into thymines and uracils, respectively.

One way to study differences in the epigenomic patterns is to identify regions that are differentially methylated in humans, archaic humans, and great apes. The potential differences found in methylation patters could be subsequently tested in human cell lines and organoids. CRISPR/Cas9 genome editing technology can be used to generate archaic versions of the genes, as well as dCas9 fused with DNA (de)methylation enzymes to specifically alter DNA methylation patterns, and test their effect in the context of human organoids.

Paleoepigenomics is an emerging field that likely will provide important advances in human evolution and diversity in the next 10-15 years, especially as more high-quality ancient genomes are available. A recent study has been able to predict with surprising accuracy the morphology of a Denisovan skull only from the epigenomic analysis of the high-coverage Denisovan genome

(Gokhman et al., 2019). It is worth mentioning also that right now there are only five epigenomes from ancient *H. sapiens* genomes analysed, one of which is from Spain (the La Braña 1 Mesolithic genome, in León).

3.2 Comparative phylogenetic approach from distant species to study human evolution

Phylogenetic Genome-Phenome analysis

To understand polygenic and complex biological processes such as those that may be exclusive to humans it is possible to adopt a comparative phylogenetic strategy by leveraging "multi-omics" data. Some genomic changes (for instance, amino-acid changes, rates of protein evolution or gene multicopy states, to name but a few) can be associated to changes in ecological, anthropometric or biomedical traits of interest (including, but not limited to, increased or shortened lifespan and brain size, among others). These changes can be better understood by analysing a large evolutionary context; that is, by studying a wide selection of increasingly available, vertebrate, mammalian and primate species, covering highly informative phylogenetic nodes, instead of focusing only in humans and its closest living relatives. With this comparative phylogenetic approach, it is possible to identify molecular evolutionary correlates (for instance, parallel mutations) between genetic variation in a large context and specific human traits.

For this objective, current techniques such as Phylogenetic Generalized Least Squares (PGLS), Phylogenetic Path Analysis, Phylogenetic GWAS or Bayesian methods, are already available and it only remains to fine-tune them, adapting them to issues such as incomplete lineage sorting (so one can use the right tree for each genome segment) or including, on top of nucleotide variation, regulatory changes or large changes such as segmental duplications.

Evolution of the human brain. The human brain is the result of intricate developmental processes that unfold over decades, during which distinct cell types mature as they change their molecular, morphological and functional identities via the precise spatiotemporal regulation of their transcriptome. Identifying human-specific features of neurodevelopment, and how these appeared during evolution, is critical to understanding the evolution of the distinct characteristics and capabilities of the human brain. These accentuated -by not exclusive- traits include its one-thousand fold increase in cortical size compared to mouse, or three-fold increase compared to our closest relatives, chimpanzees, as well as our higher-level cognitive abilities (such as abstract thinking,

syntactical-grammatical language, or episodic memory), specific structural and hodological properties, and specific brain disorders.

There is increasing evidence that deregulation of the transcriptional and regulatory events key in neurodevelopment have direct and profound consequences on brain function, or strongly favour the risk of neuropsychiatric disorders. Revealing the molecular events underlying the evolution of human brain features, and their role in disease, necessarily requires the identification of specific variants occurring in functional elements of the genome, followed by experimental testing. However, finding the relevant variants that underlie phenotypic changes in the developing and mature brain represents a challenging task, as it requires the identification of relatively few meaningful variants from thousands of neutral variants, resulting in a "needle in a haystack" scenario. An initial approach for reducing the search space might consist on following an evo-devo approach, comparing the genomic regulation of brain development across mammalian phylogeny to identify signals of positive or negative selection that have arisen during brain evolution leading to humans. Work on species with particular strategic phylogenetic value, including mouse, ferret and macaque, has already proven to be extremely powerful to identify and understand genomic determinants of human brain evolution. Focusing on genes expressed in the brain that show signals of association to neuropsychiatric traits (GWAS/rare variants) and/or signals of positive selection that have arisen since the split from our closest relatives, may be of unique value for the identification of human-specific features of brain development and evolution.

A primary obstacle for studying brain development in humans and our closest relatives (chimpanzees) is the lack of samples. The discovery of induced pluripotent stem cells (iPSC) and their capacity to be differentiated into any cell type holds great potential for investigating development under laboratory conditions in an, in principle, unlimited amount of samples and at a fine time scale. On the other hand, understanding the evolutionary mechanisms influencing brain development and leading to the acquisition of human features can be successfully achieved by studying non-primate species with strategic positions in phylogeny and displaying unique brain features. These may include mouse (rodent with a small and smooth brain), ferret (carnivore with a medium-sized and folded brain), marmoset (primate with a smooth and small brain), macaque (primate with a large and folded brain) and human. Comparative genomic and developmental studies across these species have

already demonstrated the outstanding potential of evo-devo studies for the identification of evolutionary trajectories of mammalian brain development leading to human. Most importantly, since the advent of CRISPR-mediated genome editing, genetic manipulations of brain development are now being performed in some of model species – notably mouse – *in vivo*, in the developing embryo, which is of unrivalled value compared to any *in vitro* approach.

A second major obstacle is the challenge of comparing cell types and their fine scale organization across species. Recent advances in massively parallel single-cell profiling (transcriptomics, epigenomics, proteomics) are enabling the matching and comparison of homologous cell types between species. The first comparative single-cell analyses in developing brains indicate a conservation of broad cellular populations between humans and mouse, but dramatic differences in the proportions of cell types, their laminar distribution and morphologies. This rapidly evolving technology will soon allow the simultaneous analysis of multiple features in single cells with much finer resolution, in greater numbers and at greater speed, which when combined with the rapidly improving analytical methods holds great promise for studies of the evolution of cellular mechanisms in brain development. Therefore, the fields of evolutionary biology and neurodevelopment (with new and refined experimental animal models and genome editing tools) are ripe for being combined to test hypotheses about the molecular changes that took place in the evolution of the developing human brain, leading to its increased size and architectural complexity, and thereon to explain malfunction observed in disease states.

A pressing scientific challenge is to elucidate the genetic basis of the evolution of lineage-specific features of neurodevelopment and disease in humans, with respect to both the closest relative non-human primates and other more distant mammalian species in key steps of the evolutionary process. Effective testing of this question will require using in vitro systems proximal to human foetal brain development, and non-conventional animal models displaying features of relevance in this issue. Development of induced pluripotent stem cell lines and cerebral organoids from humans and other great apes are promising *in vitro* approaches to unravel functional processes of early brain development leading to human specific traits. Genome editing of these iPSC cell lines to replace endogenous sequences with those of strategically relevant species or groups, and then produce cerebral organoids for functional analysis of brain development, holds great promise for the next generation of

discoveries in this field. For example, editing human iPSCs with Neanderthal and Denisovan sequences, or editing chimpanzee iPSCs with modern human sequences can help unravel metabolic and physiologic differences between species. These approaches require access to live cells from great-apes, to be transformed into iPSCs, and will also allow unravelling molecular and gene networks that underlie differences between humans and great apes in other tissues. Regarding the elucidation of earlier steps of mammalian brain evolution leading to human, this will involve use of non-conventional species amenable to experimental genetic manipulation in the developing embryo, and of strategic relevance to this question due to their key position through long-term human phylogeny. Species where transient and stable transgenesis for genes affecting brain development has been achieved, and hence that are emerging as the next generation animal models to study brain development and function, include macaque, marmoset and ferret.

In summary, understanding the evolution of the human brain will require i) in silico multi-omics bioinformatics analysis leveraging multiple types of data from multiple sources (single cell vs. bulk tissue; transcriptomic, chromatin organization and proteomic data; selection scans and association studies) to prioritize evolutionarily relevant genetic variants affecting candidate genes related to lineage-specific neurodevelopmental phenotypes. ii) Utilization of genome editing tools onto non-conventional but strategic animal models, and induced pluripotent stem cells (iPSC) differentiated into neuronal lineages and brain organoids, to test alterations of cytological (e.g. cell morphology, cell composition and neurogenesis) and molecular (e.g. transcriptome and epigenome) parameters. For example, a human specific variant could be introduced into chimpanzee iPSC to test whether and to which extent it can switch molecular networks and phenotypes to a more human-like status. This approach is most suited for studying early developmental processes and the only possible one now that research with ape individuals is -rightfully- banned from experimentation.

3.3 Advancements in paleoanthropology for the study of human evolution

While the topics above mainly relate to innovations in molecular-biological and genetic of data extracted from fossil hominin remains, at the phenotypic level important advances in the study of human evolution are also occurring.

Digital anatomy, advanced 3D imaging and virtual databases. Methodological improvements in recent times have taken quantitative studies of hominin fossil morphology to new levels. This has been facilitated by advances in high resolution digital imaging techniques, both at internal (volume) levels by advanced industrial micro-/nano-tomographic scanning and external (surface) levels. Applications of these imaging methods to hominin fossils have enhanced the possibility of quantifying internal and external features of hard tissue morphology, leading to a better understanding of ancient biological process related to growth, morphogenesis and, more generally, ontogeny. For example, the application of high radiation methods to fossilized teeth has produced insights into change of growth rates not only of dentition, but also of the history of the organism and life in recent hominin species, as well as of highly mineralized cranio-dental fossils at the origin of the genus *Homo*, *Australopithecus* and *Paranthropus*.

At a different level, advanced 3D imaging has also led to more easily accessible digital data-bases of digitized fossil remains. In this respect ongoing analyses of the new hominin species, *Homo naledi* (Rising Star Caves, South Africa) has provided a pioneering example in enhancement of paleoanthropological research by facilitating open and free access to 3D digital models for comparative anatomic research, 3D printing etc., online excavations with live distributions via internet, thus fostering democratization of paleoanthropological research which traditionally has been characterized by limited access to privileged institutions.

More and more research institutions are following these trends by participating in EU-wide and global networks of activities of digitalization collection of fossil and extant hominin and primate species relevant to the study of human evolution. Guaranteeing the current and future participation of CSIC in such digital collection networks (DISSCo, SYNTHESYS, etc.) will be crucial to maintain its position in the phenotypic domains of human evolutionary studies.

High-density and high-resolutio 3D morphometrics. Digital fossil specimens available to virtual morphological laboratories increasingly follow new trends in morphometric analyses and modelling. After pioneering development of geometric morphometric methods for the quantification of curve and surface geometries of virtual 3D anatomical objects, current efforts are mainly directed to developing methods for analysing high-density (HD) coverage 3D measurements and methods for full surface recording based on landmark-free techniques for sophisticated shape analyses and modelling.

These new morphometric methods not only enable powerful quantitative and visual analyses, but also hypothesis-driven, quantitative reconstructions of anatomically incomplete (i.e. fragmented) fossil remains. Hypothesis-driven, quantitative reconstructions of fossil hominin body structures are key to analysing function and thus the assessment of its potential adaptive value, which is becoming more easily comparable with genetic evidence in functional contexts. In addition, reconstruction of body structures is also important for modelling of growth processes of structures that allow for the study of interaction between fossil hominin organisms and their ecosystems (eco-evo-devo). For example, HD 3D morphometrics of fossil hominin skull reconstructions have recently been combined with big data on brain imaging and ancient genome sequences to provide a model for the evolution and emergence of modern human brain functions and skull morphology. Another venue for future development is the aforementioned link between epigenenomic data and its potential for morphological reconstruction of skeletal profiles of hominin species such as Denisovans.

Additionally, body functions are likely to be more effectively studied by using HD morphometrics applied to the virtual anatomy of digital extant and extinct hominins. Simulations of body functions such as locomotion or respiration in reconstructed 3D-structures and organ systems based on computer modelling of real-time 4D processes relevant to human evolution, has the potential of not only producing more accurate paleoanthropological knowledge, but also extending it to the usage in applied biomedical research, sport and other living human related sciences. Because of the need to extract as much information as possible from scarce hominin fossil remains, paleoanthropological research has traditionally interacted with biomedical technology to improve analytical methods, in addition to constant efforts in the field to provide a more complete fossil record for multidisciplinary approaches to studying processes in human evolution.

3.4 The role of technology in shaping human evolution

Technology as a driving force in social evolution. Defined as a set of material and informational devices, the goal of technology in the context of human evolution is to optimize exchange of energy with the environment (White, 1946). Human adaptive success is ultimately based on the thermodynamic efficiency of societies and their ability to progress through a cumulative introduction of new technological devices that compensate for demographic increase, and for changes in the environmental and/or subsistence conditions. Therefore,

the interaction between technological and social change is a prime characteristic of the human evolutionary process and is a high- priority theme in Prehistoric Archaeology as a discipline. In the next few years, three specific challenges should be addressed: 1) The nature of technological systems as thermodynamic devices, and how the evolutionary trajectory of societies is shaped by alternative production processes. 2) The impact of technology on social systems: the interaction between technology and social organization, which is bidirectional as technology shapes the thermodynamic outputs of human labour and determines basic aspects of social organization such as the division of labour and the control of means of production, among others. 3) The role of technological transmission and change in the process of social evolution; technical know-how and skills are not genetically determined in humans and are transferred through learning processes that are socially regulated, which should be studied from the perspective of the social systems in which they are embedded.

The uniqueness of human culture and the emergence of technology in our evolutionary past. The classic idea of the first human technologies assumes that early stone tool making starting at circa 3 million years ago focused on getting cutting edges, mainly through flakes obtained from knapping (Harmand et al., 2016). However, some authors (e.g., Haslam et al., 2009) have highlighted the fundamental and broader role of other tool-use activities apart from knapping, which could have been the most important ones in the earliest stages of human technology. This has important evolutionary implications, as some extant non-human primates use tools in a variety of tasks. Parallels between primate percussive technologies and early archaeological sites are producing ground-breaking results in recent years (e.g., Proffitt et al., 2016) and need to be further explored in order to bridge the gaps between fields such as Primatology and Archaeology, encouraging cross-disciplinary collaborations aimed at laying foundations for a multiple-approach understanding of the evolutionary foundations of human technological behaviour. Questions driving this research include the following: Are there ecological reasons driving the variable technological patterns detected in the early archaeological and primatological record? How similar are archaeological tools and non-human-made lithics? Do they represent the same activities? Are there parallelisms that permit comparing primate and early archaeological stone tools?

3.5. Human responses to abiotic and biotic factors and the role of climate in shaping human behavioural evolution

In human evolutionary studies, there is an overwhelming preference for abiotic interpretations of hominin adaptations and evolution, from models that highlight the role of climate in shaping speciation (deMenocal, 2011) to those that see species turnovers in the context of climate change. However, while the influence of abiotic mechanisms in biogeographic distributions of early humans is evident, they do not explain satisfactorily patterns observed in the archaeological and paleontological record. Recent studies of faunal turnovers have emphasized the importance of continuous (and likely biotic) factors in modulating faunal change (Bibi and Kiessling, 2015), suggesting that mammal communities might be more robust than expected, and that properties of the foodwebs be uncoupled from Pleistocene climatic changes. There is growing evidence that climatic factors alone do not explain hominin biogeographic dynamics, and pioneering research perspectives are exploring models in which biotic interactions (e.g., competition), which are known to be essential in shaping biodiversity and biogeographic patterns, may have played a fundamental role in shaping early hominin adaptive behaviours. This emerging field presents a number of challenges that includes 1) exploring competing hypotheses on the character and influence of environmental fluctuations in Pleistocene mammal communities through computational modelling; 2) calibrating geochemical data from paleoclimate and paleovegetation studies in archaeological and paleontological sequences to ground truth climate maps and aid interpretation of model results; and 3) creating geo-referenced databases of modern and fossil mammals and hominins that should be combined with climatic layers to refine biogeographic models of species distribution across the Old Word.

3.6 Big Data repositories and spatial data infrastructures (SDI)

Research on the pre-modern human occupation of the Old World can no longer be compartmentalised regionally, and any comprehensive study should be based on a comparative analysis of the alternative evolutionary and cultural pathways observed in each area. The need for drawing similarities and differences between the earliest archaeological sequences has been recognised in recent years, but efforts have focused on general literature reviews of the evidence in each region or comparisons of particular attributes of specific archaeological artefacts. The synthetic research advocated by Kintigh et al. (2014) for Archaeology is now possible and greatly needed to understand properly the ecological and cultural patterns of early humans. Nonetheless, such

research should be based on the systematic analysis of quantitative datasets (thus superseding the current stage of superficial literature review comparisons) and include computational modelling. The substantial datasets produced in recent years present a unique opportunity to understand the patterns that structured the behavioural ecology of pre-modern humans during the earliest colonization of the Old World, but also a challenge due to the disparity of archives, research traditions and the lack of systematic intra and inter-regional comparisons. Such exceptional challenges and opportunities should embrace big-picture objectives to understand the alternative evolutionary trajectories adopted by hominins that shared an overarching biological and cultural background (i.e., they were premodern humans using Early Stone Age technological solutions), and address key research questions such as: 1) How can migration waves during the Pleistocene be identified and how many were? 2) How continuous is human occupation in the presumed evolutionary centres as opposed to the discontinuous record of other regions of the Old World? 3) Can risks and challenges faced by humans in each region explain variability of hominin adaptations?

CHALLENGE 6 REFERENCES

Bibi, F. and Kiessling, W. (2015). Continuous evolutionary change in Plio-Pleistocene mammals of eastern Africa. *Proceedings of the National Academy of Sciences 112*, 10623–10628.

Brewer, J. et al. (2017). Grand challenges for the study of cultural evolution. *Nature Ecology & Evolution 1*, 0070.

De Menocal, P.B. (2011). Climate and Human Evolution. *Science 331*, 540–542.

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., Marques-Bonet, T., Rak, Y., Carmel, L. (2019). Reconstructing Denisovan anatomy using DNA methylation maps. *Cell 179*, 182–192.

Harmand, S. et al. (2015). 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya. *Nature 521*, 310–315.

Haslam, M. et al. (2009). Primate archaeology. *Nature 460*, 339–344.

Kintigh, K. W. et al. (2014). Grand challenges for archaeology. *Proceedings of the National Academy of Sciences USA 111*, 879–880.

Proffitt, T. et al. (2016). Wild monkeys flake stone tools. *Nature 539*, 85–88.

Racimo, F., Sikora, M., Vander Linden, M., Schroeder, H., Lalueza-Fox. (2020). Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nature Review Genetics 21*, 355–366. doi: 10.1038/s41575-020-0218-z.

White, Leslie A. (1946). *The Science of Culture. A study on man and civilization*. Grove Pess Books and Farrar, Strauss Giraux. New York.

Welker, F. et al. (2019). Dental enamel proteome shows that *Gigantopithecus* was an early divergent pongine. *Nature 576*, 262–265.

Welker, F. et al. (2020). The dental enamel proteome of *Homo antecessor*. *Nature 580*, 235–238. https://doi.org/10.1038/s41586-020-2153-8.

SUMMARY FOR EXPERTS

D 2.6 – SOCIAL AND HUMAN EVOLUTION

Understanding the social and biological clues in the human evolutionary lineage

Challenges

-To reconstruct the hominin phylogenetic tree and human dispersals

-To unravel human adaptations

-Genetic evolution

-Evolution of the human brain

-Evolution of human sociality, behaviour and technology

-To explore human diversity (genetic and morphological) and disease

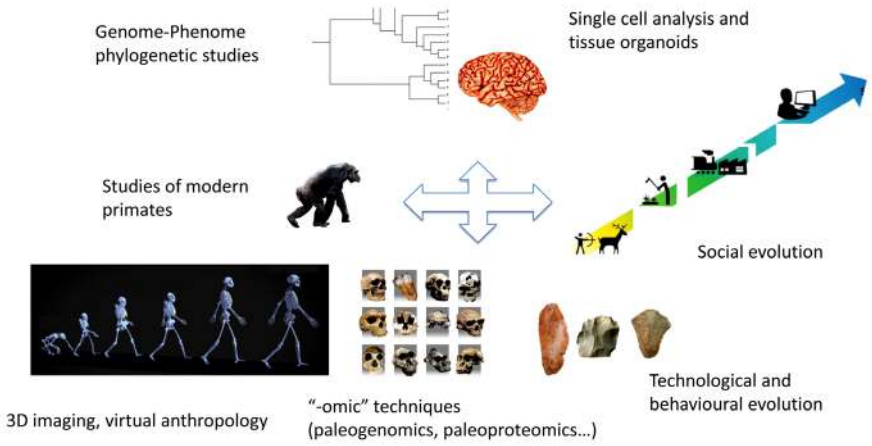
CSIC centres: IBE; IH-CCHS; MNCN; INA; CNB; IMF; IACT; INCIPIT; IAM; EBD

Work packages

-Application and development of "-omics" techniques (paleogenomics, paleoproteomics, paleoepigenomics)
-Genome-Phenome phylogenetic analysis
-Single cell profiling
-Organoids
-3D high-resolution imaging
-Big data repositories and Spatial Data Infrastructures

Technology: Big data management (on genetic, 3D fossil, and archaeology spatial databases); high throughput sequencing; mass spectrometry, peptide and paleoenvironmental reconstructions; Advanced 3D imaging; CRISPR; tissue differentiation

SUMMARY FOR THE GENERAL PUBLIC



CHALLENGE 7

ABSTRACT

Diseases result from the disturbance of the physiological homeostasis of organisms. Disturbances can be endogenous (*e.g.*, heritable diseases, developmental constraints or cancer) or exogenous (*e.g.*, infections or intoxications). In as much as organisms themselves, and the way they interact with their biotic and abiotic environments, are the result of evolutionary forces, their diseases are also the result of complex (co)evolutionary processes. Nowadays, by incorporating an evolutionary perspective, it is possible that a better understand will emerge and help to combat disease.

KEYWORDS

antibiotic resistance

cancer as an evolutionary system

Darwinian medicine emerging pathogens

evolvability | host-pathogen coevolution

One-Health | pathogen evolution

CHALLENGE 7

EVOLUTION OF HEALTH AND DISEASES

Coordinators

Santiago F. Elena
(I2SysBio)

Iñaki Comas
(IBV)

Participating researchers and research centers

Jesús Blázquez
(CNB)

Jordi Gómez
(IPBLN)

Elena Gómez-Díaz
(IPBLN)

Jesús Pérez-Losada
(CIC)

Francisco Sobrino
(CBM-SO)

EXECUTIVE SUMMARY

This chapter describes how incorporating an evolutionary thinking would generate new insights and provide fresh perspectives into human, farm animals and crop diseases. A shift in the reductionist conceptual paradigm is urgently needed to tackle complex problems such as the origin, spillover and spread of new infectious agents, the origin and spread of microbes resistant to antibiotic drugs, the role of microbiota in preventing invasions from pathogenic microbes, the dynamics of cancer cells and tumour growth, the origin of medical conditions such as diabetes, obesity, cardiovascular disease, asthma, allergies and autoimmune diseases, aging, preeclampsia, menopause. All these processes are governed by the same basic and universal forces of evolution: mutation, genetic drift, migration, natural selection, and adaptive tradeoffs. It is becoming increasingly clear that solutions to such complex problems potentially involve every level of organization, from molecules to populations. The Modern Synthesis, or Neo-Darwinian Theory of Evolution, provides the conceptual, theoretical and mathematical frameworks, which together with genomics, bioinformatics, and an integrative systems biology perspective will provide a holistic understanding. Safer and more durable chemotherapeutic treatments, more effective vaccination strategies that consider the entire biosphere, and a more rational use of medicines could be designed by considering the unavoidable evolutionary component of living organisms.

1. INTRODUCTION AND GENERAL DESCRIPTION

The way in which living beings respond to disease, be it infectious or the result of a failure in cellular homeostasis, is the result of a process of (co)evolution. Any rational future development of new therapies should consider this evolutionary component. The most obvious example of the importance of evolutionary biology is that of infectious diseases. Humans, farm animals and vegetable crops, are continuously being exposed to pathogens that place strong selection pressures on the evolution of genomes, evolving complex defence mechanisms. In response, immune pressure or the action of antibiotic drugs on pathogen populations, results in the evolution of counter defences and resistant variants capable of infecting and reproducing. This "arms race" leads to coevolution between both players. Another paradigmatic example of eco-evolutionary process and disease is cancer. In many ways, tumour cells behave like parasites of their host, replicating much faster and generating many more errors than normal cells, displacing them from their environment. Characterizing and understanding genetic diversity for disease susceptibility genes existing in different host populations has become the focus of many studies in recent years. Understanding why certain individuals are more resistant than others cannot be done without understanding the evolutionary history of these populations. Why alleles of disease susceptibility have been maintained at high frequencies cannot be understood without evolutionary concepts such as fitness trade-offs.

Life has demonstrated a great flexibility and can be found in every environment: perpetual ice, boiling water, extreme pH, high salt, or overwhelming pressure. Thus, it should not be unexpected that living beings can develop mechanisms to resist any drug humans use against them. Resistance is thus a natural evolutionary response to drug exposure. On the one hand, maintaining the stability of genetic information is vital for the perpetuation of species. Therefore, cellular organisms evolved a DNA replication machinery in order to minimize mistakes, protect, and repair DNA. Most viruses, by contrast, do not encode for such error correction mechanisms as their living strategy is based in a fast yet low fidelity replication. In both cases, errors are produced during genome replication. On the other hand, adaptive evolution, which occurs through heritable variation and selection, allows organisms to adapt to new environments or to adverse conditions. Without genetic change, there are no new genes or alleles and, hence, no adaptation. Therefore, living beings had to confront the dilemma of changing to adapt or remaining stable and face

the possibility of extinction. Nature has selected those organisms possessing a mutation rate that is a compromise between adaptability and resilience.

Infectious diseases represent 20% of all human diseases (www.who.int/whr/1996/media_centre/press_release/en/). Pathogens have accompanied humans since their origin ~200,000 years ago. Smallpox, bubonic plague, influenza, schistosomiasis, tuberculosis, and malaria have decimated entire populations and shaped human evolution (Domínguez-Andrés and Netea, 2019). While Science is producing outstanding new results every day, still little is known about the molecular basis of human infectious diseases. As the World Health Organization (WHO) expressed in its last report: "in the context of globalization, climate change and population growth, the risk and impact of emerging and re-emerging infectious diseases is extreme, and there is an urgent need to find new strategies for their eradication" (Engebretsen et al., 2017).

Nowadays, cancer has shifted the focus of attention away from infectious diseases. Cancer is a critical problem worldwide. Yet, an appropriate conceptual framework is needed to answer questions such as why cancer occurs, why the risk of cancer throughout life is so high and has increased, why when after introducing a drug there is tumour regression followed, almost inevitably, of recurrence and resistance. Cancer is a Darwinian process, and this must be considered in attempts to understand and control it.

Here, a few cases in which evolutionary thinking provides a new perspective to disease will be highlighted. Evolutionary theory provides a common conceptual framework to understand otherwise unrelated processes such as the evolution of pathogen's virulence, pathogens and cancer resistance to drug therapies, and how human's own evolutionary history has shaped its susceptibility to disease.

Infectious diseases: evolution of virulence and host-pathogen interactions. One of the greatest challenges in infectiology is to understand variation in the damage caused by the parasite (virulence) and its ability to transmit (infectivity) as well as in host susceptibility and resistance to infection. Current evidence shows that intra- and inter-host interactions and dynamics shape the infection phenotype and are key in the evolution of new strategies for virulence and infectivity management and host resistance (Ewald, 1994). Within the host, parasite survival depends mainly on evading its immune system, and there are multiple ways to do this: strategies to penetrate and multiply inside

different cells and tissues, vary surface and modulate proteins, or suppress the immune response. To optimize between-hosts transmission, transmissible parasite stages need to be present at the right time/tissue. This requires a very precise control of the cell cycle. However, parasites and hosts experience an enormous variation in their external and internal environments that impact their fitness, and so adjustments in phenotype must occur quickly. The solution is to endow the system with phenotypic plasticity: the ability of a genotype to express different phenotypes in a heterogeneous environment (Taylor et al., 2006; Gómez-Díaz et al., 2012; Westneat et al., 2019). In this context, a better understanding of the molecular mechanisms underlying phenotypic variation in parasite virulence/transmission and host susceptibility/resistance, is key to prevent the (re-) emergence of infectious diseases and the failure of existing eradication interventions (Louhi et al., 2012).

The role of antigenic variation in pathogens: Viruses, bacteria, fungi, and protozoa all face similar challenges upon infecting a susceptible host. They must avoid mechanical clearance, successfully adhere to their preferred tissue/cell, avoid recognition by the immune system, and avoid being killed by various components of the immune system. In order to survive and trigger infections, and thereby ensure transmission, they have evolved similar strategies. One of these strategies is antigenic variation: the ability of a pathogen to alter surface proteins or carbohydrates avoiding the immune response. The term encompasses both *phase variation*, which is the on/off of a gene, and true *antigenic variation*, which is the expression of alternative forms of a protein.

Varied genetic and epigenetic mechanisms have been described that underlie the ability to produce variable phenotypes. Genetic mechanisms include mutation and recombination and imply a change in the genomic sequence of an antigen-encoding gene, resulting in changes in the protein secondary structure, or of its regulatory elements, that alter the expression level. These changes occur at relatively low rates (except in RNA viruses), are irreversible and heritable. Epigenetic mechanisms involve, by contrast, changes in gene expression that are independent or occur without altering the primary genomic sequence. These changes are stable and potentially inherited. Contrary to genetic events, they can occur within a generation, are reversible and inducible by internal or external stimuli.

Phase variation was first described in bacteria as the phenomenon of switching between two alternative phenotypes among the cells in a clonal population. A common mechanism is through transcriptional regulation, which can

be either genetic or epigenetically controlled. *Borrelia* spp., *Salmonella* spp., *Treponema pallidum* are examples of phase variation that rely on homologous DNA recombination (Vink et al., 2012). Epigenetic transcriptional regulation has been also described, for example, in *Escherichia coli* where the expression of pili, which is under the control of the pap operon, is regulated by environment-dependent DNA methylation (Braaten et al., 1991). Similarly, in *Candida glabrata* phase variation of the gene *Epa6* is induced as a direct response to the level of NAD^+ found in urine (Mundy and Cormack, 2009). Phase variation can also be the result of translational regulation through a variety of mechanisms including slipped-strand mispairing, early ribosome dissociation, and mRNA instability.

Some microbes have evolved a more complex strategy of antigenic variation that relies on large, multi-copy and hyper-variable gene families, in which each individual gene copy encodes an antigenically distinct surface protein. In *Plasmodium* spp. and *Trypanosoma* spp., antigenic variation involves transcriptional switching among members of the *var* (~60 genes) and the *vsg* (~1000 genes) multi-copy gene families, causing parasites with different antigenic and phenotypic characteristics to appear at different times within a population. The switching is under genetic through homologous recombination, and epigenetic control, involving different mechanisms: heterochromatin propagation and histone acetylation, chromatin accessibility and nuclear positioning, and specific transcription factors (Maree and Patterton, 2015; Gómez-Díaz et al., 2017; Ruíz et al., 2018). In bacteria, antigenic variation through homologous recombination has been described for *Borrelia* spp. and *Neisseria* spp., and in fungi for *Candida* spp. (Cahoon and Seifert, 2011).

The genetics and epigenetics of the susceptibility/resistance to infection. The outcome of the interaction between parasites and their hosts does not only depend on the pathogen's variability and evolvability, but also in the spatial and temporal variability in the host susceptibility genes.

The basis of many infectious diseases is associated with variation in host genomes, and in this case, such disease susceptibility trait can be inherited. An example is the sickle-cell anaemia trait, which is associated with a reduced susceptibility to malaria. Its prevalence in the population is considered a result of the selective pressure that malaria parasites have been exerting on human populations through evolutionary time (Karlsson et al., 2014). Genome-wide association analysis (GWAS) have been widely used to characterize genetic markers associated with disease risk, inform about disease

progression and thereby help to choose the most appropriate treatment (Craig, 2008). As a result of these studies, human leukocyte antigen (HLA) variation has been associated with susceptibility or resistance to malaria, tuberculosis, leprosy, AIDS, and persistence of viral hepatitis. Variation in the tumour-necrosis factor-α gene promoter, as well as allelic variants of the vitamin D receptor, appear to be associated with differential susceptibility to several infectious diseases (Cooke and Hill, 2001). In the context of the current COVID-19 pandemic, investigations are now being undertaken to unveil the genetic basis of susceptibility to SARS-CoV-2 that may explain differences in clinical outcomes and across countries. Variations in the *ACE2*, *TMPRSS2* and the HLA genes have been proposed as likely risk factors (Sungnak et al., 2020). In this and future pandemics/epidemics, GWAS and comparative genomics will be instrumental in the understanding of host-pathogen genetic interactions, and a prerequisite to developing effective vaccines and therapies.

There is also growing evidence that genotype-genotype associations between the host and the pathogen are relevant in different aspects of infection. For example, transmissibility in tuberculosis seems to depend on specific combinations of host and genetic background (Gagneux et al., 2006). The same is true for clinical outcomes in *Helicobacter pylori* where gastric cancer is more likely in certain combinations of human and bacterial genetic backgrounds (Kodaman et al., 2014). In both cases, these differences relate to the disruption of the long-term co-evolution of specific combinations of human and bacteria genotypes, what is called sympatric associations. Some of the loci involved in the pathogen can now be explored using GWAS approaches given the large number of pathogen sequences available (Falush, 2016). However, to identify relevant loci that can inform vaccine development of host-directed-therapies GWAS must evolve towards genome-to-genome comparison where both the host and the bacterial variation are incorporated. This approach has started to be applied in viruses (Fellay and Pedergnana, 2019) and will likely be a cornerstone to decipher the host and pathogen interacting loci for a number of infectious diseases.

In addition to the genetic component, environmental factors also influence the aetiology of the disease, the so-called disease triangle (Scholthof, 2007). The interaction of the individual with the environment, and over time, results in a great variability in the forms in which the disease occurs, the phenotype of infection, with an impact in public health interventions. Tuberculosis is a

classic example of poverty disease in which socioeconomic determinants play a major role. Immune senescence in humans is characterized by a decrease in the innate and adaptive response, which increases the infectious disease risk. There is also a seasonal variability in the susceptibility to autoimmune and infectious diseases (Castle, 2000) probably linked to seasonal variation of sun exposure and skin colour with consequences in micronutrients like vitamin D.

When talking about environmentally driven disease variability, an important consideration is the underlying molecular mechanisms, since this variability is, *a priori*, non-genetic. Epigenetic mechanisms have been proposed to link environmental changes and phenotype by modulating gene expression (Klemm et al., 2019). Understanding the role of epigenetics in disease has become a priority of current biomedical research. In fact, changes in the methylation status and histone acetylation/methylation of multiple genes involved in immune signalling and immune attack pathways has been associated with disease susceptibility and progression for viral and bacterial plant pathogens (Wang et al., 2019), malaria, AIDS, hepatitis B, and tuberculosis (Grabiec and Potempa, 2018), though studies are still scarce. The study of how these epigenetic risk factors can be edited or reversed is known as *epigenetic therapy* (Dumitrescu, 2018). Which is the impact of these changes in the long term (heritability), their causality and the clinical applications of epigenetic therapies, is a research challenge in the field.

Emerging infectious diseases. The devastating SARS-CoV-2 pandemic the world is currently facing highlights the tremendous risk and importance that the emerging and re-emerging infectious diseases pose for humankind. The extremely fast dispersion of the SARS-CoV-2, its high contagiousness, and the relatively high number of severe cases of COVID-19 that require intensive care, is stressing health systems and in the short term national and regional industrial activity and economy. This situation is even more threatening for developing countries that lack basic health infrastructures and resources, and where the pandemics will co-occur with other epidemics such as tuberculosis, malaria or HIV-1. The SARS-CoV-2 pandemic is the last and paradigmatic episode of a series of human epidemics caused by emerging viruses and bacteria that have occurred in the past decades. These inlcude Lyme disease, enterohemorrhagic *E. coli* O157:H7, Chikungunya virus, Dengue fever virus, Hantavirus, HIV-1, West Nile virus, Zika virus, and the coronaviruses SARS-CoV and MERS-CoV, among many others, including the still active epidemics of Ebola in central Africa. In most of these outbreaks, the

emerging virus appears as a result of a jump from wild or domestic species to humans that has led to coin the term *One Health* to deal with this new scenario, which is concomitant with the global warming and other human aggressions to nature and wildlife (Zinsstag et al., 2018). Being coherent with this One Health concept, not only virus affecting directly to humans but also any other pathogen affecting animal welfare and production as well as in plant health should be considered. Three examples that illustrate this complex network of interconnections, whose complexity is just being glimpsed, are the observation that SARS-CoV-2 can be transmitted from human to tigers and domestic cats, animals acting as reservoir of antibiotic resistance linked to animal husbandry and the existence of a non-viral clonally transmissible cancer which affects Tasmanian devils.

Evolution of resistances. In 2016 infections with antibiotic resistant bacteria caused 700,000 deaths and devastating economic losses. By 2050, antibiotic resistant infections deaths will amount to 10 million, surpassing cancer as the main cause of mortality worldwide. In addition, viral, fungi and parasitic resistance also accounts to a major number of worsened clinical outcomes and deaths. The incidence and type of resistant infections is markedly different between developed and developing countries. While in high-income countries major problems are associated with the emergence of superbugs in nosocomial settings, in low- and middle-income countries (LMIC) community transmission is a major public health challenge. For example, every year 250,000 persons died of multidrug resistant tuberculosis and three million are predicted by 2050. Importantly, genomic and epidemiological data combined with the power of evolutionary analysis, is allowing to understand, even anticipate, how resistance emerges and spreads with a resolution not available before.

Resistance to any molecule is developed via the basic genetic mechanisms of mutation and recombination (which create and combine resistance determinants). In the case of bacterial pathogens, their capacity of horizontally transfer of resistance determinants (among members of the same or different, even phylogenetically distant, species) must also be taken into account (MacLean and San Millan, 2019). Gaining insights into these mechanisms and how they are affected by drug challenges is fundamental to understanding the basic process of resistance acquisition and the development of novel targets for both new drugs and diagnostic tests. Consequently, it is expected that reducing the impact of these basic mechanisms, development and

dissemination of resistance will be minimized. The generalized use of drugs changed the practice of medicine allowing the development of modern medical practices. However, the emergence and spread of drug resistances in recent decades is threatening global health.

Many resistances are acquired by horizontal transfer of resistant genes. In many cases, including bacteria, fungi and human cells, drug resistance is generated by mutations that increase drug efflux and/or decrease binding to the drug target, drug import or prodrug modification. Resistance mutations or their effects can be ameliorated by recombination as, for instance, widening the spectrum of resistance enzymes. Hence, hyper-mutation and hyper-recombination rates will be beneficial to become resistant and fitter. Hyper-mutator clones have been found in all domains of life, including bacteria, fungi, parasites, and malignant cancer cells (Blázquez, 2003). Viruses have, in general, a high mutation rate, thus effectively behaving as hyper-mutators. In addition, some anti-infective agents act not only as selectors of resistant or hypervariable alleles but can cause pathogen genetic instability and thus influence the evolution and spread of resistance determinants in different ways, including enhancement of mutation, recombination and horizontal gene transfer. Furthermore, most public health policies have relied on the assumption that resistance comes with a fitness cost in absence of the drug, thus most public health actions for a number of infectious diseases are directed at the individual treatment level. However, in *Mycobacterium tuberculosis* there is now clear evidence linking experimental evolution and clinical data that the fitness cost can be compensated rendering resistant strains as transmissible as the susceptible ones (Comas et al., 2011). Molecular epidemiology studies are also characterizing the spread of resistance in the nosocomial and community settings for a number of pathogens and resistance determinants.

Nevertheless, understanding the complex factors and their interactions influencing the level of resistance reached nowadays is still far from understanding. These factors include, among others, pathogen interactions with drugs and hosts, intrinsic and induced genetic variability of the pathogen, co-resistance to unrelated drugs, emergence of successful resistant clones that do not pay a fitness costs (owed to the presence of other epistatic mutations) and transmission rates of pathogens among humans, animals, and the environment. Other public factors such as rates of vaccination, health care systems and population density and mobility are also important. Therefore, resistance

is a multifactorial problem that must be addressed from different disciplines, including Medicine, Genetics, Microbiology, Epidemiology, Chemistry, Ecology, and Sociology, yet always under the perspective of evolution, as finally, resistance is the unavoidable consequence of the evolutionary process of life. Not taking this premise into consideration will lead to a bitter fight against microbial evolution, as, up to now, no drug has escaped from resistance.

There is thus an urgent need to act to avoid returning to the pre-antibiotic era. Among the measures, a better understanding of the evolutionary mechanisms and drivers of resistance, development of new evolution-proof drugs and the implementation of methods to increase treatment efficiency based on evolutionary concepts are key to deal with this worldwide problem. Viruses, bacteria, fungi, parasites, and human cells will always find a solution to survive drug pressures. The problem can be ameliorated or avoid it being worse, but experience shows that it is not possible to find a drug/condition for which microbial life will not find a solution.

Cancer as an evolutionary process. The main problem of cancer is clonal heterogeneity, that is, the presence and continuous appearance of different groups or clones of tumour cells, with new genetic and phenotypic characteristics, including the ability to metastasize and resist treatments. Tumour clonal heterogeneity derives from clonal evolution, in which some cells give rise to others with new phenotypic characteristics (Greaves, 2015).

Evolution is characterized by long periods of stability interrupted by punctuated explosive changes (Gould and Eldridge, 1977). The latter would also be the case in cancer: explosions of clonal expansion secondary to the action of high-impact mutations selected by environmental pressure changes. Meanwhile, there would be multiple clones coexisting and competing with each other in parallel (Greaves, 2015).

The evolutionary selection unit, responsible for tumour initiation and the formation of new clones, are stem cells with self-renewal capability. Thus, they are essential for the maintenance of normal and cancer tissues. Any tumour cell can acquire mutations, but only stem cells could propagate. Stem cells would be responsible for generating new clones, also for metastases, as self-renewal capacity would also be required to keep a clone at a distance, as in the primary tumour, and even the resistance to chemotherapy (Greaves, 2013). Of all these mutations, only a few of them induce essential functions for tumour development (drivers) and generate new clones. Thus, only 3% of the

mutations of a tumour would be drivers, the rest would be passenger mutations not subjected to selection. Driver mutations are context-dependent. They can be advantageous in some tumour environments, but not in others. In addition, mutation advantages may be epistatic with other mutations. Driver mutations may be different in different clones, so targeted therapy may not be optimal for treating the entire tumour.

Cancer complexity goes beyond local tumour cells since cancer is not a cell autonomous process but is a tissue that grows in a systemic context. To grow, cancer cells need to interact with the stroma and with long-distant compartments related to the physiology of the organism, needs nutrients, oxygen, hormones like insulin, etc. Therefore, the selection pressures for the appearance of new tumour clones come not only from the compartment of the tumour parenchyma but also from the other compartments, stromal, and long-distant ones. In these compartments, several structures and functions behave as intermediate phenotypes or sub-phenotypes of cancer at systemic, tissular, cellular, and molecular levels. Lastly, there is the genetic level, with the mutations in tumour cells, and the genetic background of the organism that regulates all the levels of sub-phenotypes. Sub-phenotypes have complex interactions and negative feedbacks that stabilize cancer as a system and make it resistant to new imbalances. All this organization results in multiple connections that can be studied from a System Biology perspective. Thus, cancer constitutes an evolutionary and adaptive system that usually evolves in the direction of more significant robustness, that is, of a greater ability to withstand challenges or stress, including the evasion of therapies.

Human evolution, evolutionary trade-offs and the incidence of cancer. When multicellular organisms appeared 600 billion years ago, mechanisms emerged to maintain tissue integrity. Thus, at that period, the first tumour suppressor genes appeared. DNA repair mechanisms are an adaptive response to prevent the increased risk of cancer. They are an inherited trait and have therefore been generated and improved throughout evolution. The ability to reproduce in humans decreases markedly after 55 years. All physiological adaptations are designed to act efficiently before old age. Achieving old age is a relatively recent event and is a consequence of the medical and public health improvements. However, genes were selected by a pressure that no longer exists in modern day lifestyles. Therefore, there would be a dissociation between genes and the actual environment because genes would be mainly prepared to respond to the Paleolithic environment with considerable physical

activity, intermittent food, resistance to famines, cold, etc. (Tooby and Cosmides, 1990). Among these gene-environment dissociations would also be the DNA repair systems. With the advent of aging, the DNA reparative capacity begins to fail, and there is a progressive increase in the incidence of cancer with age.

Other changes in the incidence of cancer come from changes in lifestyle compared to the Paleolithic one. Importantly, this fact offers the possibility of prevention by strategies that try to imitate those environmental processes. The Darwinian vision of cancer offers a better understanding of the disease, as well as new prevention alternatives and therapeutic approaches that will promote more personalized medicine. This is how Evolutionary Medicine arises and addresses the question of how evolutionary past has shaped the vulnerability to diseases such as cancer, type II diabetes, osteoarthritis, etc., both as a species and regarding the different susceptibility between individuals (Williams and Nesse, 1991).

Maladaptations, evolutionary constraints and disease. There are many examples to illustrate how past evolutionary history of humans determines the prevalence of certain diseases in human populations. Here, for illustrative purposes, three well-studied cases will be discussed. First, women in cultures without contraception and with normal birth intervals followed by ca. three years of breastfeeding had about 100 menstrual cycles along their lifetime. However, in the modern western world, women may have up to 400 cycles. Consequently, they are experiencing a far larger number of cell divisions, which put them into increased risk of breast cancer (Eaton et al., 1994). Women who experience first birth young and that spend most of their reproductive period pregnant or in lactational amenorrhea have significantly lower breast cancer rates (Strassmann, 1999). Menopause was not a phenomenon of relevance for primate and *Homo* ancestors given their shorter life expectancy. Now, it is a real issue: reduced estrogen levels result in development of osteoporosis and weaker bones. Better contraceptive treatments should be designed keeping in mind these evolutionary trade-offs. Indeed, there are several strategies underway that want to mimic the protective effect of early and repeated pregnancy. Regarding the existence of allelic forms that favour the onset of breast cancer, the relatively high prevalence of *BRCA1* mutations has been associated with increased fertility in its carriers, resulting in a phenomenon of antagonistic pleiotropy. Likely, this fact would also occur in other tumours (Smith et al., 2013).

Humans evolved in an environment in which infections were pervasive, and most people carried all sorts of parasites most of the time. Some of these parasites down-regulated host immune responses to enhance their survival. By doing so, they reduced the susceptibility to autoimmune diseases by reducing the overall production of antibodies. The antiseptic environments keeps societies free of many parasites, but autoimmune diseases are much more common now. For instance, Gabonese children with schistosomiasis have fewer allergic reactions to dust mites; likewise, the incidence of asthma is far less in adults infected with nematodes living in LMIC (Wilson and Maizels, 2004). It may take hundreds of generations for selection to bring the screening mechanisms of the immune systems into equilibrium with the cleanliness of modern environments.

Another evolutionarily interesting problem is the well-known mother-offspring conflict. While mothers are equally interested in the success of each of her offspring, the foetus alas, has evolutionary interests that differ from its mother's with respect to its siblings. The foetus manipulates the mother's physiological state via the placenta. Mother's pre-eclampsia and diabetes are by-products of this evolutionary conflict (Trivers, 1974).

A change in the paradigm: the One-Health approach. Humans are probably one of the greatest evolutionary forces on the planet nowadays (Palumbi, 2001). One of the most dramatic pieces of evidence for the influence of humans on evolutionary processes is the rapid selection of resistance in microbial pathogens (Davies, 2007). Anthropogenic effects impact significantly on all components of the biosphere. A significant quantity of anti-infective agents from human and veterinary use are released into the environment. These molecules are not only selecting for resistant microorganisms, but are likely to affect bacterial evolvability, producing unpredictable consequences for the biosphere equilibrium. Because of the necessity to rapidly adapt to anti-infective challenges, microorganisms with high evolvability could have been selected by a second order selection effect (Gillings and Stokes, 2012). In addition, as indicated above, some anti-infective agents can directly increase microbial evolvability (Blázquez et al., 2018). In summary, the presence of anti-infective agents at global scale is probably changing the pace of microbial evolution.

Ethical challenges of the One-Health approach. Most aspects mentioned above have important ethical implications related to three main areas: where is research being done, how is it being done and who is doing it?

Where? Today more than ever, understanding the origin and emergence of organisms is key to identify genetic determinants of phenotypic traits like disease susceptibility in humans or virulence and drug resistance in pathogens. In that regard, there is an emerging body of work at the crossroads between genomics, archaeology and sociology in which the roots of disease are investigated and how they have changed over time. This is possible because of the advent of ancient DNA sequencing technologies. These technologies have allowed scientists access to the complete genomes of hosts and microbes from human ancient remains. For example, ancient human DNA analyses have uncovered a plethora of *Homo* species that coexisted and interbred with *Homo sapiens*. Likewise, it has been possible to recover the genome sequence of pathogens from human remains dating hundreds to thousands of years ago like, *M. tuberculosis*, *Yersinia pestis*, *Vibrio cholerae*, Influenza A virus H1N1, HIV-1, or *Plasmodium falciparum* (Bennett and Baker, 2019). These techniques allow understanding now the causes of well-known plagues. Most of these pathogens are still major contributors to human morbidity and represent major global health issues. Importantly, these extraordinary advances in ancient DNA genomics with potential to inform human health are only possible by multidisciplinary teams of anthropologists and sociologists that can identify likely causes of death. However, the race for ancient DNA material sometimes has left behind the communities and countries where the samples are obtained.

In general, there has always been a bias in the representation of LMIC in research, and especially in global health issues. For example, integration of LMIC is key to understanding the emergence of global pathogens with origin in those regions, as cultural and social issues cannot be ignored to understand the big picture. A reflection of the issue is the minor role that LMIC has had until now in human genomics research. Most analyses of human populations at the genetic level aimed to look for example for pathogenic variants have been based on individuals from the global north and this has also translated into tools, like SNP typing arrays, highly biased by diversity in these populations. This is paradoxical, as it is known that Africa is a hotspot of human genetic diversity not only important to understand human evolution but also to understand genetic determinants of human health. As in the case of ancient DNA, there is a need to transition to more ethical and socially responsible research. Key will be to involve in research and in the communication of results to the communities affected. In the case of human genomics, new initiatives like H3Africa are aimed to correct the bias by empowering African countries

in genomics applied to human health issues. It not only covers research led by African researchers but also builds capacities and technologies on site.

How? One of the most important scientific advances in recent molecular biology is the application of CRISPR/Cas9 gene editing approaches as a new strategy to fight infectious and parasitic diseases (Soppe and Lebbink, 2017). Such a breakthrough opens a major ethical and moral debate when it comes to releasing GMOs into the wild, especially for gene drive systems aiming to suppress natural populations or even species, for example disease-transmitting mosquitoes (Kyrou et al., 2018). The biosafety risks are expected to be low as these systems are very specific and the technology is well developed. Ecological risks, however, do exist. Natural systems are complex and what works under confined laboratory conditions may not work as expected in the wild. It is challenging to predict how the ecosystem will react to such an introduction and what will be the short- and long-term side effects. However, the social risks are very high. There is a worldwide movement against GMOs and a general mistrust in experts and institutions. The public debate and the conflict around this technology is more important in developing countries. In these regions, the potential benefits of curing major diseases such as malaria or Dengue fever are prioritized over the ethical/social and ecological risks. Appropriate scientific community scrutiny on the technical aspects and methodology, conducting pilot field trials, containment and control procedures as well as community engagement programs, become essential.

Who? This is a multi-facet problem that depends on the investment in research by different countries and, more worrisome, by gender inequality issues. In an ideal world, there are equal opportunities for women and men to enter and progress in all scientific disciplines without bias or prejudice. However, the reality is that women represent, at best, less than 30% of researchers worldwide. More importantly, they lack a critical mass of representation in all areas of STEM (Science, Technology, Engineering and Mathematics), including global health and infectious diseases related fields. That is, despite women scientists are leading ground-breaking research across the world, they remain underrepresented in citations, on editorial staff at scientific journals, as invited speakers at scientific conferences, and as members of decision-making committees, and many more women than men will leave science as career levels progress (English et al., 2020). The same situation applies to Spain according to the last report by the Women in Science Commission of the CSIC, and that of the Ministry of Science and Innovation. A clear example is the

composition of the PTI Global Health where women coordinate only 34% of the subareas. Similarly, women lead only 32 % of the challenges in this book. The solution to the glass ceiling and the gender imbalance in science is complex and it will require a reform of the educational, social and scientific systems. Nonetheless, it is about time for this transformative change and institutions must take urgent action.

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

An evolutionary approach to minimizing the impact of emerging pathogens. For emerging infectious disease prevention, it becomes essential to anticipate the rapid adaptation potential of these pathogens by characterizing comprehensively existing genomic/epigenomic diversity, and to unveil the mechanisms of regulation of antigenic variation.

Several areas of research that were already relevant for the identification, understanding and control of emerging diseases turn now crucial for modern societies to realize and prevent not only SARS-CoV-2 but any related or unrelated pathogens that will, soon or later, arise. The main areas of research to be potentiated in this scenario, always under the consideration of promoting multidisciplinary and polyvalent approaches, as well as applied and translational research, are:

- Molecular structure and biology of already known emerging and re-emerging pathogens (viroids, virus, bacteria, parasites, fungi), to gain information relevant for their detection and control. Studies should address mechanisms leading to appearance in the natural population of new or altered human pathogens with enhanced virulence, antigenic variations, drug resistance, or modified transmissibility or infectivity.
- Ecologic and environmental factors influencing disease emergence and distribution, such as the influence of natural, direct man-made, or climate-induced environmental change on the emergence of pathogens; the effects of alterations in host or vector population density and distribution on diseases. Indeed, human cases of Dengue virus have been reported in the Mediterranean coast, and its mosquito vector is present in areas of Spain. More worrisome, a short chain of Crimean-Congo haemorrhagic fever transmission took place in Castilla-León a few years ago, being the tick vector common in wildlife.

- Development of new diagnostic tools based on the detection of both nucleic acids and specific antibodies using, as much as possible, versatile, rapid, automatable and easy-adaptable approaches to ensure pathogen detection and surveillance. Including cheap and easy-to-use point-of-care tests particularly in LMIC. Likewise, high quality epidemiological studies should be supported with special emphasis in optimizing the use of national and international databases as well as encouraging the incorporation of environmental and wildlife data.
- Develop local and global real-time genomic surveillance of pathogens associated with eHealth developments as genomic data simultaneously enables diagnostics, surveillance and epidemiology. Use real-time genomic data coupled with evolutionary analyses to inform interventions in community, nosocomial and epidemic scenarios.
- To identify antivirals and other antimicrobial and anti-parasitic compounds, using screening libraries of already known compounds as well as considering new investigational compounds and monoclonal antibodies. Besides the basic and applied knowledge resulting from these researches, the availability of a wide variety of drug families will become relevant for fighting emerging infectious agents.
- Viral and other pathogen resistances can no longer be considered as bad luck. The O'Neill report (amr-review.org), probably the most comprehensive scientific assessment on the present and future impact of antibiotic resistance, estimates that ten million lives will be lost per year by 2050 due to an antimicrobial resistant infection, surpassing cancer. Even today the number is staggering, 700,000 per year. More than a quarter of those deaths are and will be related to multidrug resistant forms of tuberculosis while the rest are accounted mainly by the rise of superbugs like methicillin-resistant Staphylococcus aureus (MRSA). In addition, anti-treatment resistance in parasites like the case of artemisinin resistance in malaria, are a major contributor to morbidity mainly in LMIC.
- In the case of viruses, the properties of the structure of RNA genomes, in the form of molecular collectives that interact with each other in many different ways (quasi-species), and their evolutionary dynamics, make it possible to predict that mutants resistant to a single drug will appear. One of the properties of the collective is memory, and single-drug treatments can only serve to exhaust it as a future possibility. This knowledge should guide in the future not only how to administer antiviral drugs and follow their effect, but also in the search for new

therapeutic strategies based on this knowledge. For example, considering highly conserved RNA viral structures as potential drug targets or employing drug mixtures leading the virus to extinction by the accumulation of mutations.

- Development of new versatile and easy-adaptable prophylactic and therapeutic vaccines, either based on inactivated viruses, RNA, DNA, virus subunits, or engineered attenuated viruses, as well as new vaccine delivery technology, which are to prevent and treat disease, with a substantial impact on human and animal health. Attention should be paid to the assessment of their efficacy and antigenic spectrum, to the risk of selection of resistant mutants, and to simplify vaccine matching to cope with new variant viruses.

Evolution-proof design of vaccines and other therapies. Vaccination has greatly reduced the burden of infectious diseases. However, current licensed vaccines, almost exclusively antibody-based in their action, are protective against pathogens with low antigenic variability. Antigenically variable pathogens represent indeed the major burden of infectious diseases today and their ability to escape natural host immunity undermine all therapeutic efforts. The reality is that the success of vaccines against these pathogens has been limited, with 40-50% of efficacy in the best of the cases, a protection that tends to diminish with time.

There have been two main strategies: (*i*) live-attenuated pathogens. These vaccines mimic natural infection, but in a weakened non-pathogenic fashion. (*ii*) Derived toxins, subunit preparations, carbohydrate cocktails, or conjugate vaccines. This strategy assumes that conserved sequences derived from the pathogen are promising immunologic targets, however, they have in most cases failed to induce and maintain protective immunity. The assumption that all immune responses are beneficial for the host is being challenged for a number of infectious diseases. Identification of where it resides protective immunity is a major challenge in infectious diseases. A clear counterintuitive example is *M. tuberculosis*, whose antigens are highly conserved and show no hallmark of immune evasion. The emerging view is that some immune responses are needed for the immunopathology required for the transmission of the bacilli (Comas et al., 2010).

An additional problem for vaccination and drug development in rapidly evolving pathogens is the diversity and variability of strains and phenotypes in natural infections, which remain in most cases uncharacterized. This has been

shown to be the cause of temporal and spatial limited efficacy of many vaccines. In addition, most vaccinology studies focus on reference strains that have been maintained and serially passed in the laboratories for decades (using non-standardized media conditions), and that are not representative of the adaptation potential of the pathogen. Another problem is the use of non-appropriate animal models for preclinical vaccine development (Gerdts et al., 2015).

The challenges are to find new strategies specifically targeting genetic/antigenic variability and that are representative of the natural diversity/infection process. One strategy is to design disease, population and patient specific immunogens reflecting and/or incorporating complex and rapidly changing epitope/antigen landscapes of vaccine targets. The rationale is that activation of diverse pools of B and T immune cells requires the vaccine immunogens carrying antigenic diversity that mirrors the natural infections (Servín-Blanco et al., 2016). Another approach could be the development of molecules that interfere with the regulatory proteins/sequences controlling variant expression, to enhance vaccine efficacy. Compared to traditional vaccine targets, these master regulators are probably more robust, less mutagenic and/or recombinogenic.

Environmental pressures and oncotherapeutic opportunities. Cancer, as an adaptive system, acquires several functions that allow surviving. These functions coincide with the so-called hallmarks of cancer (Hanahan and Weinberg, 2011), and all of them favour, directly or indirectly, clonal expansion. Usually, treatment strategies attack some of these hallmarks, as proliferation, angiogenesis, or genomic instability. The vision of cancer as an evolutionary process opens the possibility of trying to modify the environmental selection pressures that determine tumour evolution. Many of these selection pressures are unknown, such as the ones that push cancer cells to metastasize. It has been linked to tumour hypoxia, but more studies are needed to clarify this point. Some therapeutic opportunities would be: (*i*) The selective pressure that determines chemotherapy resistance is chemotherapy itself. It is not easy to eliminate this selection pressure. However, it should be accompanied by a better knowledge of the signalling networks and genes that interact in a context that turns mutation into drivers. Knowing the other mutations that participate with the driver mutation behaviour (epistasis) and targeting them, could be an excellent therapeutic opportunity. This is the genetic concept of synthetic lethality, in which the combination of mutations in multiple genes results in cell death, and provides a framework to design novel therapeutic

approaches to cancer. (*ii*) The dynamics of tumour stem cells that come out of the cell cycle under hypoxia conditions increase their resistance to chemotherapy. It would be feasible to alter this dynamic and reintroduce them into the cell cycle in a specific manner, to avoid their resistance to chemotherapy. (*iii*) The therapy paradigm that modifies the pressure selection of the tumour environment is carried out with the new immunotherapy strategies with a high impact on cancer therapy. The immune system infiltrates tumours, such in scarring tissues, activating the stroma, and allows the new angiogenesis and thus the arrival of oxygen and nutrients and the growth of tumour cells (Coussens & Werb, 2002). The immune system also removes tumour cells that are antigenically very different from normal ones (immunoediting). Thus, there is a selective pressure to be tumour cells antigenically tolerated by the immune system. This fact helps explain why cancer may arise in immunocompetence conditions. The new immunotherapy treatments, and especially with anti-PD1/PDL1 antibodies, decrease the immune system's tolerance to tumour tissues and are rejected (Havel et al., 2019).

3. KEY CHALLENGING POINTS

- Identify pathogen determinants of virulence and spread and the trade-off between both.
- Integrate experimental research on the evolution of pathogens with clinical and epidemiological observations.
- Identification and characterization of the basic genetic mechanisms and pathways that govern the evolution and dissemination of resistance. Increase the knowledge of the roles of anti-infective molecules in the biosphere and their effect on microbial evolution and resistance development.
- Search for mechanisms to avoid the dumping of anti-infective agents into the environment and/or inactivating them to prevent their effects on evolution of resistance.
- Development of evolution-proof strategies against microbial pathogens, including search for drugs that interfere with genetic variation and horizontal transfer.
- Consolidating what is already known as Evolutionary Medicine or Darwinian Medicine.
- Environmental metagenomics and metaviromics to get a realistic picture of what pathogens are present in diverse ecosystems (e.g., urban, fresh and coastal waters, agroecosystems, wild areas).

- Impulse research agendas that integrate basic, applied and clinical research.
- Incorporate the most recent advances in "omic" techniques (in particular, single cell genomics) to understand the interaction between pathogens and host cells at different levels of biological complexity (from cells to populations).
- Integrate epidemiological modelling and evolutionary analysis tools, including real-time genomic epidemiology, to characterize basic epidemic parameters of pathogens.
- Development new mathematical and statistical methods to integrate multi-level information into predictive models.
- Incorporating systems biology computational approaches to better understand host-pathogen and cancer evolution.
- Attract clinical immunologists to this research area.
- Incorporate entomologists that work in the transmission vectors of the zoonotic pathogens.

CHALLENGE 7 REFERENCES

Bennett, R.J., Baker, K.S. (2019). Looking backward to move forward: The utility of sequencing historical bacterial genomes. *Journal of Clinical Microbiology 57*, e00100-19.

Blázquez, J. (2003). Hypermutation as a factor contributing to the acquisition of antimicrobial resistance. *Clinical Infectious Diseases 37*, 1201–1209.

Blázquez, J., Rodríguez-Beltrán, J., Matić, I. (2018). Antibiotic-induced genetic variation: how it arises and how it can be prevented. *Annual Reviews Microbiology 8*, 209–230.

Braaten, B.A. et al. (1991). Evidence for methylation-blocking factor (mbf) locus involved in pap pilus expression and phase variation in *Escherichia coli*. *J. Bacteriology 173*, 1789–1800.

Cahoon, L.A., Seifert, H.S. (2011). Focusing homologous recombination: pilin antigenic variation in the pathogenic *Neisseria*. *Molecular Microbiology 81*, 1136–1143.

Castle, S.C. (2000). Clinical relevance of age-related immune dysfunction. *Clinical Infectious Diseases 31*, 578–585.

Cooke, G.S., Hill, A.V.S. (2001). Genetics of susceptibility to human infectious disease. *Nature Reviews Genetics 2*, 967–977.

Coussens, L.M., Werb, Z. (2002). Inflammation and cancer. *Nature 420* 860–867.

Craig, J. (2008). Complex diseases: research and applications. *Nature Education 1*, 184.

Comas, I. et al. (2012). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics 44*, 106–110.

Davies, J. (2007). Microbes have the last word. *EMBO Reports 8*, 616–621.

Domínguez-Andrés, A., Netea, M.G. (2019). Impact of historic migrations and evolutionary processes on human immunity. *Trends in Immunology 40*, 1105–1119.

Dumitrescu, R.G. (2018). Early epigenetic markers for precision medicine. *Methods in Molecular Biology 1856*, 3–17.

Eaton, S.B. et al. (1994). Women's reproductive cancers in an evolutionary context. *Quarterly Reviews in Biology 69*, 353–367.

Engebretsen, E., Heggen, K., Ottersen, O.P. (2017). The sustainable development goals: ambiguities of accountability. *Lancet 389*, 365.

English, E.D., Power, B.J., Gómez-Díaz, E. (2020). Building parasitology communities to promote gender equality. *Trends in Parasitology*, In press.

Falush, D. (2016). Bacterial genomics: microbial GWAS coming of age. *Nature Microbiology 26*, 16059.

Ewald, P.W. (1994). *Evolution of infectious diseases*. Oxford, Oxford University Press,

Gagneux, S. et al. (2006). Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the USA 103*, 2869–2873.

Gerdts, V. et al. (2015). Large animal models for vaccine development and testing. *ILAR Journal 56*, 53–62.

Gillings, M.R., Stokes, H.W. (2012). Are humans increasing bacterial evolvability? *Trends in Ecology and Evolution 27*, 346–352.

Gómez-Díaz, E. et al. (2012). Epigenetics of host–pathogen interactions: the road ahead and the road behind. *PLoS Pathogens 8*, e1003007.

Gómez-Díaz, E. et al. (2017). Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. *Scientific Reports 7*, 40655.

Gould, S.J., Eldridge N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology 3*, 115–151.

Grabiec, A.M., Potempa, J. (2018). Epigenetic regulation in bacterial infections: targeting histone deacetylases. *Critical Reviews in Microbiology 44*, 336–250.

Greaves, M. (2013). Cancer stem cells as 'units of selection. *Evolutionary Applicat 6* 102–108.

Greaves, M. (2015). Evolutionary determinants of cancer. *Cancer Discovery 5*, 806–820.

Hanahan D. Weinberg, RA. (2011). Hallmarks of cancer: the next generation. *Cell 144*, 646–674.

Havel, J.J., Chowell, D., Chan, T.A. (2019). The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature Reviews Cancer 19*, 133–150.

Karlsson, E.K., Kwiatkowski, D.P., Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics 15*, 379–393.

Klemm, S.L., Shipony, Z., Greenleaf, W.F. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics 20*, 207–220.

Kodaman, N. et al. (2014). Human and Helicobacter pylori coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences of the USA 111*, 1455–1460.

Kyrou, K. et al. (2018). A CRISPR/Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nature Biotechnology 36*, 1062–1066.

Louhi, K.R., et al. (2013). Genotypic and phenotypic variation in transmission traits of a complex life cycle parasite. *Ecology and Evolution 3*, 2116–2127.

MacLean, RC, San Millan, A. (2019). The evolution of antibiotic resistance. *Science 365* 1082–1083.

Maare, J.P., Patterton, H.G. (2014). The epigenome of *Trypanosoma brucei:* a regulatory interface to an unconventional transcriptional machine. *Biochimica et Biophysica Acta 1839*, 743–750.

Wilson, M.S., Maizels, R.M. (2004). Regulation of allergy and autoimmunity in helminth infection. *Clinical Reviews in Allergy and Immunology 26*, 35–50.

Mundy, R.D., Cormack, B. (2009). Expression of *Candida glabrata* adhesins following exposure to chemical preservatives. *Journal of Infectious Diseases 199*, 1891–1898.

Palumbi, S.R. (2001). Humans as the world's greatest evolutionary force. *Science 293*, 1786–1179.

Fellay, J., Pedergnana, V. (2020). Exploring the interactions between the human and viral genomes. *Human Genetics 139*, 777–781

Ruíz, J.L. et al. (2018). Characterization of the accessible genome in the human malaria parasite Plasmodium falciparum. *Nucleic Acids Research 18*, 9414–9431.

Scholthof, K.G. (2007). The disease triangle: pathogens, the environment and society. *Nature Reviews Microbiology 5*, 152–156.

Servín-Blanco, R. et al. (2016). Antigenic variability: obstacles on the road to vaccines against traditionally difficult targets. *Human Vaccines and Immunotherapeutics 12*, 2640–2648.

Smith, K.R., Hanson, H.A., Hollingshaus, M.S. (2013). BRCA1, and BRCA2 mutations and female fertility. *Current Opinion in Obstetrics and Gynecology 25*, 207–213.

Soppe, J.A., Lebbink, R.J. (2017). Antiviral goes viral: harnessing CRISPR/Cas9 to combat viruses in humans. *Trends in Microbiology 25*, 833–850.

Strassmann, B.I. (1999). Mensual cycling and breast cancer: an evolutionary perspective. *Journal of Women's Health 8*, 193–202.

Sungnak, W., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature Medicine 26*, 681–687.

Taylor, P.D., et al. (2006). The evolutionary consequences or plasticity in host-pathogen interactions. *Theoretical Population Biology 69*, 323–331.

Tooby, J., Cosmides, L. (1990). The past explains the present. Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology 11*, 375–424.

Trivers, R.L. (1974). Parent-offspring conflict. *American Zoologist 14*, 249–264.

Vink, C., Rudenko, G., Seifert, H.S. (2012). Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiology Reviews 36*, 917–948.

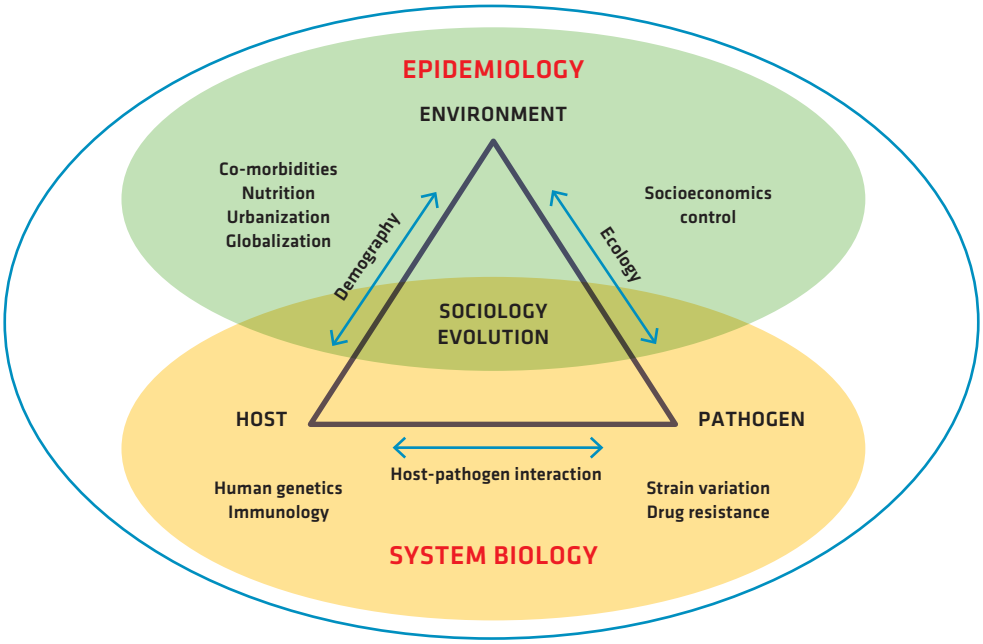
Wang, C. et al. (2019). Epigenetics in the plant-virus interaction. *Plant Cell Reports 38*, 1031–1038.

Westneat, D.F. et al. (2019). Causes and consequences of phenotypic plasticity in complex environments. *Trends in Ecology and Evolution 34*, 555–568.

Williams, G.C., Nesse, R.M. (1991). The dawn of Darwinian medicine. *Quarterly Reviews in Biology 66*, 1–22.

Zinsstag, J. et al. (2018). Climate change and One Health. *FEMS Microbiology Letters 365*, fny085.

SUMMARY FOR EXPERTS AND THE GENERAL PUBLIC



CHALLENGE 8

ABSTRACT

One of the main aims of synthetic biology is to assemble a minimal living unit with programmable functionality. Achieving this grand challenge – building a synthetic cell from scratch – will contribute to our understanding of the basic principles of life and its emergence from lifeless components; it will also provide the tools for novel solutions to outstanding environmental and health-related problems.

KEYWORDS

Synthetic biology (bottom-up, top-down, *in silico*)
minimal cell | proto-cell
biological self-organization | bioengineering
reconstructing cellular machines
cell-free systems | minimal metabolism

CHALLENGE 8

SYNTHETIC LIFE

Coordinators
Germán Rivas, (CIB-MS)
Eva García, (ICP)

Participating researchers and research centers
Jorge Barriuso, (CIB-MS)
Fernando de la Cruz, (IBBTEC CSIC-UNICAN)
Rafael Giraldo, (CNB)
Sonsoles Martín-Santamaría, (CIB-MS)
Miguel A. Peñalva, (CIB-MS)
Manuel Porcar, (I2SysBio CSIC-UV)
Jesús Rey, (IFS-CCHS)

EXECUTIVE SUMMARY

Is it possible to build a synthetic cell from scratch? Modern science has devoted significant efforts to unveiling the basic principles of life. This research is contributing to a deep understanding of the parts that make up the macromolecular machines that operate in the cell. However, despite these advances, we still do not understand how these pieces interact in a coordinated manner to develop cellular functions. Thus, the generation of life from the molecular components that existed on the primitive earth is one of the great unresolved enigmas, and thus a major scientific challenge. Synthetic biology offers new strategies for its resolution. From a fundamental perspective, the integration of molecular modules that will give rise to functional synthetic cells will help to reveal the limits of life. In this regard, we can envision that the merging of synthetic biology with molecular and cellular evolution may end up in the synthesis of living cells from scratch, the functions of which will be tuned by controlled evolutionary mechanisms. Besides answering questions about the basic operating principles of life, the realization of synthetic cells will lead to a new and unprecedented technological revolution. Understanding how the living cell works by reconstructing it from its essential components will open up new horizons of application in medicine and biotechnology. The design of optimized test systems for new drug discovery and biodegradable materials are just some examples. Therefore, knowledge and technologies generated during the process of building synthetic cells will

contribute to a healthier and more sustainable world. Synthetic life research needs nevertheless to address societal challenges, including the ethical and philosophical aspects of these investigations. It also requires the training of future scientists in novel ways of exploring living systems with the application of engineering to understand and master biological complexity.

The quest for synthetic cells is a world-wide effort, in which Europe is playing a major role. Recently, a European Synthetic Cell initiative was launched to engineer a minimal cell from its molecular building blocks, with the aim of achieving this challenge within the next two decades. The CSIC hosts a significant number of top researchers from various disciplines, who are currently working independently on different aspects of synthetic biology and related areas. Integrating their efforts in an intramural program would position the CSIC among the top European hubs in synthetic life research, maintaining a prominent position at the forefront of this grand challenge.

1. INTRODUCTION AND GENERAL DESCRIPTION

1.1. The quest of synthetic life

One of the grand scientific and intellectual challenges of this century is the construction of a synthetic cell from its constituent molecular components, with the inclusion of new functional modules. Despite our extensive knowledge about the molecular building blocks that comprise present-day cells, we do not understand how these building blocks collective operate to make the transition to living systems. Mastering how to build a synthetic cell from lifeless components will help answer the fundamental question of how life works. It will also provide deep insight into how life may spontaneously emerge from its non-living constituents and will enable to interface biological systems with non-living materials (Bayley, 2019; Beales et al., 2018; Porcar and Peretó, 2016)

Accomplishing this challenge will also allow designing artificial cell platforms and reprogrammed cells towards next-generation bio-factories to open unique solutions for environmental, energy, and health-related problems. The profound understanding – and control – of cellular life to build synthetic cells from scratch will also raise important philosophical, ethical and social questions on the impact of these technologies, opening relevant prospects for the dialogue between science and humanities, and between science and society (Schwille et al., 2018)

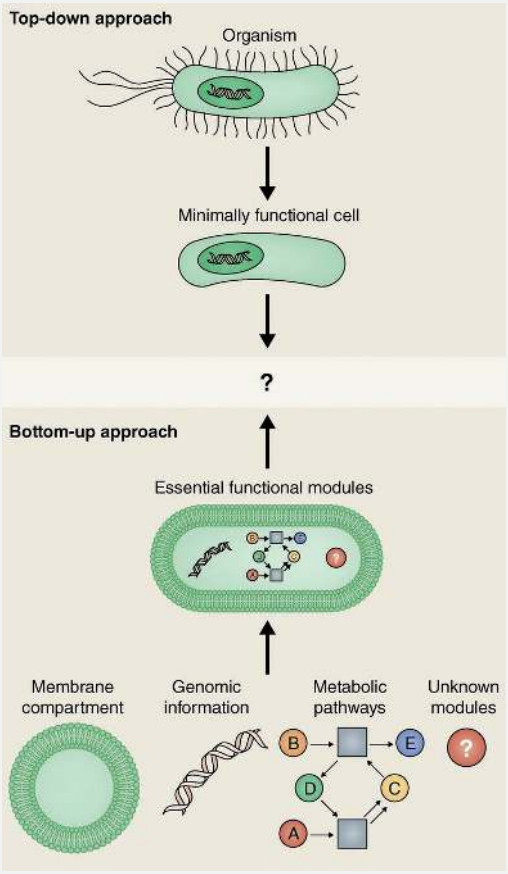


FIGURE 1–Top-down and bottom-up synthetic biology approaches towards building a minimal living cell. Scheme reproduced with permission from Schwille (2015) J. Cell Biol. 210, 687

1.2. Synthetic biology

The design and synthesis of novel biological (or bio-inspired) systems that display useful functions, even those ones that do not exist in nature, has accelerated over the past decades and matured into synthetic biology. This discipline introduces the engineering perspective to study biological processes (Ausländer et al., 2017). In this optics, the cell can be considered as an intricate factory composed of devices and machines that carry out multiple tasks and that can be engineered to produce systems with programmable functionality.

Many achievements in synthetic biology have resulted from top-down approaches in which pre-existing cells have been modified (Ausländer et al.,

2017). These developments have allowed engineering genetic circuits, biological modules, and synthetic pathways to be used for re-programming organisms and producing pharmaceuticals, environmentally sustainable products, etc. However, in many cases, the functional properties of the re-engineered cells and their control are poorly understood, limiting further progress. These fundamental limitations in our understanding have fuelled a complementary bottom-up approach to synthetic biology. This approach aims at redesigning and reconstructing biological parts, devices, and systems with increasing levels of complexity toward a minimal cell-like scaffold (Fig. 1; Bottom-up biology. Nature. 2018; 563, 171).

1.3 Minimal cells and proto-cells

The possibility of producing minimal or synthetic cells has been part of research programs for many years (e.g. obtaining artificial blood for medical purposes). In this line, one of the main goals of synthetic biology is to identify the minimal configuration that sustains a biological cell. What is the minimum set of proteins and genes that a living cell needs to achieve its essential tasks, such as self-replication, metabolism, and response to environmental cues?

The building of such synthetic cells will provide valuable insights into the self-organization processes that led the first cells to evolve from non-living elements (Beales et al., 2018). Today, the scientific community considers that one of the essential requisites for the onset of life was the compartmentalization into defined spaces of the primordial biochemical components. Therefore, one of the objectives of the research programs in synthetic life is the optimization of technologies for the production of semi-permeable compartments, either in the form of micro-droplets, lipid vesicles, or protein or nucleic acid-based capsules.

Minimal genomes containing a reduced number of genes can lead to viable cells, as demonstrated by Craig Venter's group, and the Synthetic Yeast Genome Project Sc 2.0 (Hutchison et al., 2016; Shao et al., 2018). In these top-down approaches, to creation of a minimal cell is based on the selective removal of components from their natural genomes, with the important limitation that the functional properties of a significant number of the genes are unknown. Therefore, this approach leaves open many essential questions about how these cells work. Hence and owing to these complexities, it has not yet been possible to design and build a simplified form of life using bottom up

strategies, starting from a limited number of chemical or biochemical building blocks. Although our fundamental understanding of the individual building blocks of living systems has dramatically improved, assembling a minimal set of components from which life-like properties can emerge remains a formidable challenge.

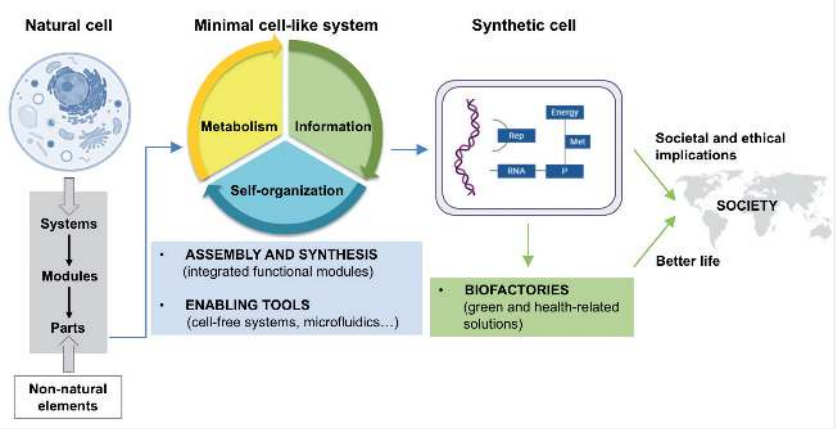
Over the last decades, directed evolution, a method that reproduces in the test tube the main processes of natural evolution, has been instrumental in designing *ad hoc* proteins for biotechnological applications (Arnold, 2018). The implementation of such evolutionary mechanisms should fine-tune entire life-like systems and building blocks in the same way as natural evolution fine-tuned natural organisms. Merging synthetic biology with controlled evolution will eventually generate synthetic cells with unprecedented functional capabilities and provide fundamental knowledge on the molecular basis of evolutionary adaptation (de Lorenzo, 2018)

1.4 Bottom-up biology

In recent years, advances in the biochemical and biophysical understanding of individual building blocks and their interactions at different levels of complexity, including membrane systems, have fuelled a considerable progress in the bottom-up reconstitution of essential cellular machines (e.g., DNA processing, cytoskeletal assembly, cell division, etc.) from simple unicellular organisms and mammalian systems (Bayley, 2019; Göpfrich et al., 2018). Many of these achievements resulted from fundamental biology studies based on reverse-engineering a particular biological function using a minimal set of molecular components, with the purpose of 'understanding by building' (Fletcher, 2016). These efforts also showed the limitations and challenges for reconstructing (and understanding) more-complex biological functions, as those in which individual molecular machines or cells act collectively (Nature 563,188–189 (2018)).

There have been also substantial advances in systems chemistry and nanotechnology, enabling the design and synthesis of complex molecular systems with life-like properties to build bio-inspired scaffolds and containers for the precise functionalization of proteins (Yeates, 2017). On the other hand, progress in biomolecular engineering allowed the design of specific mutants or chimeras of proteins with diverse origin, structure, and function. These include for example the reconstructed ancient variants of enzymes that can evolve by directed evolution (Alcalde, 2017). These developments have been

FIGURE 2–Roadmap in synthetic cell research. *Information*: DNA replication, gene circuits, etc. *Self-organization*: cytoskeleton assembly, cell division, shape and growth, etc. *Metabolism*: enzyme networks, energy production, etc. (see 3.0 and 4.1 for details)



parallel by complementary technological breakthroughs in high-resolution cryo-electron tomography, ultra-efficient genome engineering (e.g., CRISPR technology), and sophisticated molecular imaging tools.

Together, these advances will facilitate the integration of the individual molecular systems into functional modules towards the building of a synthetic cell from bottom-up. These approaches may also lead to generate artificial, bio-inspired devices that could be used to solve outstanding environmental and health-related problems (Fig. 2).

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

Modern science has devoted significant efforts to unveil the basic principles of life. This research is contributing to a deep understanding of the parts that make up the molecular machines that operate in the cell (such as those responsible for DNA replication, protein synthesis, or cell division itself). However, despite these advances, we still do not understand how these pieces are coordinated to develop most of the cellular functions. Building a synthetic cell will greatly impact society because of its potential to answer outstanding

questions such as: What defines life and how does life work? How did life originally start on Earth? Is the current form of life the only possible one?

Besides answering questions about the basic operating principles of life, the design and building of a synthetic cell will also provide novel tools to help facing global environmental and health challenges. A large number of sectors in the areas of green biotechnology, alternative energies, health, food, and biomaterials will likely benefit from the technologies developed in research programs related to the generation of synthetic cells. The design of optimized systems for the production of environment-friendly materials in the high-tech industry, new biofuels, and biodegradable polymers, and the directed improvements in bioremediation and drug discovery are some examples.

Notably, since synthetic cell systems do not need to rely exclusively on natural biomolecules or natural life processes, the use of alternative approaches will provide possibilities for applications that may go beyond currently foreseen synthetic biology technology. They include information carriers (genetic codes), synthetic amino acids and enzymes, and the potential development of orthogonal living systems (not interfering with natural living systems), among others.

3. KEY CHALLENGING POINTS

Defining the level of complexity required for life-like properties to emerge, and integrating modules in time and space to build autonomous systems are the main challenges that need to be overcome in order to develop robust synthetic cell technologies. Achieving these challenges requires an integrative and collaborative approach involving scientists from a wide range of disciplines (from life and physic-chemical sciences to engineering, and social sciences and humanities). In this regard, engineering sciences are capable of dissecting complex systems into a manageable set of fundamental functional modules. The application of engineering approaches (including the standardization of protocols and parts) to living cells will help to build a toolkit of crucial elements (e.g., proteins, lipids) to recapitulate critical functions of life, such as information processing, energy generation and metabolism, compartmentalization, growth and division, etc. These strategies will first allow understanding and controlling the functional properties of simple forms of life, to be then applied to master the features of multi-cellular organisms.

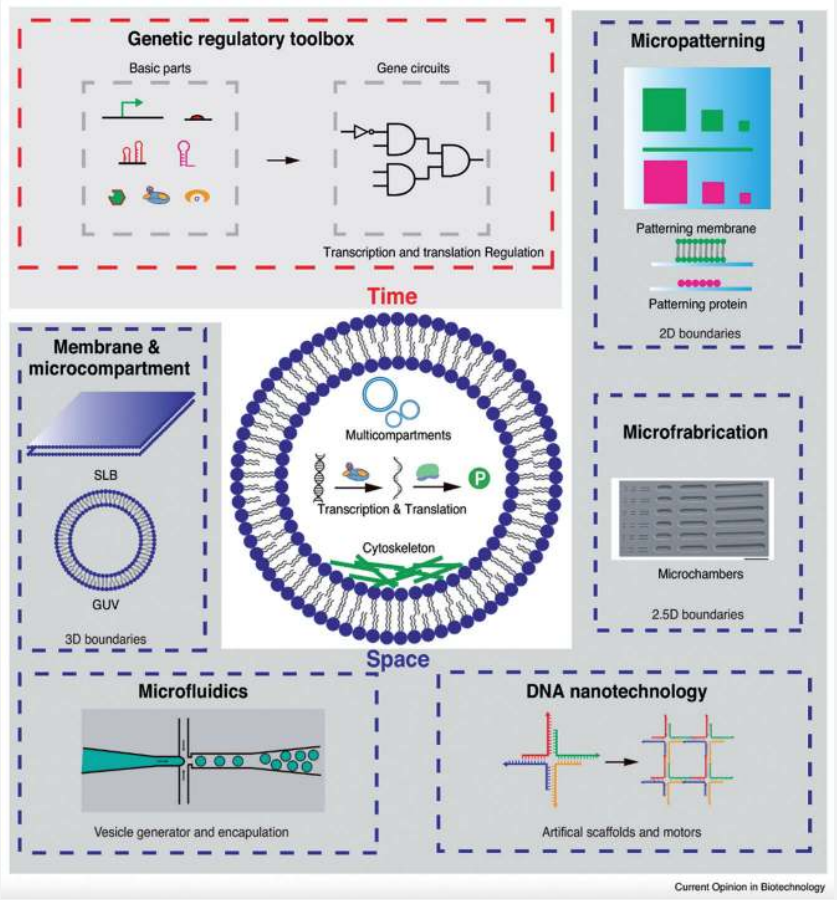
3.1. Assembly: reconstitution of biomimetic functional molecular modules

One of the early phases of the quest for synthetic cells focuses on the systematic dissection and bottom-up reconstitution of the essential features that characterized living systems as opposed to non-living ones (e.g. genetic information processing, metabolism, growth and division, and functional evolution). This assembly stage is a major experimental challenge because of the highly dynamic nature of most of these machines, which, in many cases, are associated with membrane systems (Rivas et al., 2014; Fletcher, 2016). The functional macromolecular systems developed at this stage will render the necessary modules for further integration towards artificial cells (see 3.2). They will also lead the way to optimize technologies and approaches for environmental, biotechnological, and health-related applications (see 3.4; Fig. 2). To achieve this challenge, experts in biology, chemistry, physics, and engineering should work together, exploring multiple strategies. In particular, quantitative genetic and cell biology studies of natural cells will be very relevant as they will help to optimize the design of these systems. Similarly, it will be crucial to integrate into the experimental pipeline insights derived from the research on the origin of life and the identified potential conditions enabling the emergence of life.

3.2. Synthesis: integration of functional modules in artificial and natural cells

One of the most significant challenges towards the successful development of synthetic cells will be the integration, using a modular approach, of the essential macromolecular machines previously mentioned into systems that will eventually lead to autonomously replicating artificial cells. Due to its intrinsic complexity, this goal is still in its infancy (Bayley, 2019; Beales et al., 2018). The optimization of the experimental design requires the combination of physic-chemical and bottom-up approaches, with complementary top-down synthetic biology approaches driven by modern genome engineering technologies, as well as front-line modelling and engineering tools. The quantitative understanding of the basic rules that govern the functional organization of the most straightforward forms of cellular life will then allow to explore different features of the cell, relevant to the formation of multi-cellular tissues and organisms (e.g., cell communication and differentiation, among others; Göpfrich et al., 2018).

FIGURE 3—Molecular aspects, tools and technologies enabling synthetic cell research. Scheme reproduced with permission from Jia and Schwille (2019) *Curr Opin Biotechnol* 60:179



3.3. Tools: enabling technologies in synthetic cell research

The design and production of synthetic cells will require the development of innovative and integrative experimental tools (Fig. 3).

3.3a. Generation and control of cell-like compartments by microfluidics. Biomimetic compartments are essential for reproducing cellular life processes in confined systems. Microfluidics provides a powerful technology to efficiently fabricate a large number of homogeneous cell-sized compartments,

such as water-in-oil droplets and giant vesicles. Many research programs related to microfluidics and synthetic cells are concentrated in the fabrication of 1) compartments and reactors on a chip, as novel biotechnological and biomedical devices, and 2) multi-cellular compartments containing artificial organelles that will be instrumental for the construction of artificial organs and tissues (Götfried et al., 2018; Sato and Takinoue, 2019). Research efforts will focus on solving the following key technological challenges: the optimization of assays to precisely measure the encapsulation efficiency of biomolecules within the containers, the control of the shape of the droplets/vesicles, and the reduction of the compartment size down to the micrometer scale.

3.3b. Cell-free systems for protein production and discovery. The purification of proteins is one of the bottlenecks for synthesizing functional parts to be used in modules of minimal living systems. Cell-free protein synthesis (CFPS) is an attractive alternative strategy for protein production; especially for those that are more difficult to purify, such as integral membrane proteins or soluble proteins containing functionally relevant co- and post-translational modifications (Schwille et al., 2018). Recent developments in microfluidics allow the encapsulation of CFPS in a variety of micro-compartments and this has triggered the use of these systems to develop cellular mimics. However, most commercially relevant CFPS rely on cell extracts, the composition of which is not known in the needed details. Thus, progress in synthetic cell research programs will enable a much more rational design of these cell-free systems, so that they can be individually optimized or designed for specific purposes (e.g. de novo protein design, either driven by *in silico* modelling or by *in vitro* evolution).

3.3c. Engineered natural building blocks. Top-down biology relies on the synthetic biology toolbox (tools for gene expression, genome-wide editing/writing, protein engineering, pathway optimization, metabolic engineering, etc) to provide a robust set of instruments for efficient engineering of natural building blocks. These engineered blocks will eventually be transferred to heterologous chassis, generating functional orthogonal elements, paving the way for the development of advanced chassis as bio-factories. These designed parts will add novel functionalities not directly coupled to cellular processes. Taken together, these engineered natural building blocks represent a modular tuneable platform to program synthetic cells with unique features that behave in a controlled and well-defined manner.

3.3d. DNA nanotechnology. In recent years, DNA nanotechnology has developed several modules for synthetic cells. They consist of folded DNA architectures and devices (e.g. DNA origami) fabricated via computer-aided design in combination with the large toolbox for the site-selective functionalization of DNA (Göpfrich et al., 2018). These DNA structures will be versatile tools to assemble components in a programmable manner or to construct parts of an artificial system. In this regard, designing DNA functional modules based on plasmids capable of replication and evolution inside minimal cells will be instrumental for further progress in synthetic life research.

3.4. Bio-factories: exploitation of synthetic cells for better life

The challenges posed in synthetic cell research must also address those that allow delivering technological breakthroughs, especially to provide solutions to outstanding environmental, industrial and health-related issues, as follows.

3.4a. Synthetic bio-reactors for green solutions: Current biotechnology is already very successful in using simple unicellular organisms to produce compounds with added-value for the environment (e.g., biofuel, green chemicals) and to bio-remediate soil contamination. However, our limited understanding of the cellular principles is hampering progresses. Synthetic cell development will set the basis for the design of novel bioreactors that can be engineered and controlled to improve green solutions (Ausländer et al., 2017). For example, specialized synthetic cells will allow optimization of their metabolic pathways and the encapsulation of bio-catalytic reactions to improve the production of biofuel and green chemicals, thereby facilitating the optimization of the working conditions and production yield in a contained controllable environment. These advances will eventually result in the development of the next-generation bio-refineries. On a related matter, artificial chloroplasts based on synthetic cells will optimize light conversion and energy storage (Miller et al., 2020). Finally, synthetic cells that combine the advantages of each one of the "natural" cells used for bioremediation will offer an efficient and economically viable solution for soil decontamination.

3.4b. Synthetic cell technology for health. Synthetic cells will be useful to test new drugs, to study how a specific drug interacts with a particular disease pathway, or to optimize the effectiveness of existing drugs. The use of cell-like containers for the targeted and controlled delivery of drugs or compounds in the human body has been a technological goal for a long time. There are,

however, fundamental barriers in targeting these containers to the desired human tissues and triggering the release of the active compound in a controlled way. Ideally, one may envision the design of self-supporting artificial cells that are controlled to produce and release a specific compound only after reaching the desired tissue in the body. In the long-term perspective, synthetic cell technology could be instrumental for human health applications. Functional bio-inspired molecular systems, as those made through chemical synthesis routes, will inspire new classes of active materials such as active (polymeric) materials (Budding and van Hest, 2017) or even artificial tissues that could be applied, for example, in regenerative medicine. In principle, these optimized "new materials" should be capable of active self-organization, growth and functional interaction with the human body. Envisioned applications could be the replacement of diseased cells in the body for artificial ones, the generation of artificial organs and synthetic tissues. These research programs are intimately connected to molecular robotics (Sato and Takinoue, 2019).

3.5. Society: implications of synthetic life research

The involvement of social sciences and humanities in synthetic cell research is of utmost importance. The generation of "artificial life" implies a constant and concurrent ethical and philosophical assessment, to accommodate potential paradigm shifts in our conception of living matter, and to ensure that the potential technological implementation will provide benefits rather than threat society. Ethical and societal debate on living versus non-living matter and the safety and security problems connected to emerging biotechnologies are some examples (Schwille et al., 2018; Bauer and Bogner, 2020).

Historical perspective and benchmarking approach will also have a significant role in understanding the implications of synthetic cell life research for society. For instance, the Asilomar Conference on recombinant DNA in 1975 was the first occasion in which the scientific community addressed the bio-safety concerns about recombinant DNA technology and opened scientific research to public debates. Transgenic organisms are also paradigmatic examples of the critical role that assessment and communication of the value of biological research and biotechnology applications may have on the society. Transgenics, and in particular genetically engineered foods, have generated a profound social debate, alarm and social mobilization. These reactions often occurred without a rigorous analysis and detailed knowledge of the biological foundations, interests, values (beliefs), ethical aspects, benefits, and risks.

Synthetic cell research should learn from these and other (e.g internet technology) past experiences, generating the necessary framework to expand in an ethically responsible manner. Thus, science and technology development should occur in concert with the education of the next generation of students, technical support and society at large. This effort will ensure an accurate understanding of the scientific advances resulting from the development and use of synthetic cells. Thus, dissemination activities will be instrumental in bringing synthetic biology closer to the general public.

CHALLENGE 8 REFERENCES

Alcalde, M. (2017). When directed evolution met ancestral enzyme resurrection. *Microbial Biotechnology 10*, 22–24.

Arnold, F.H. (2018). Directed Evolution: Bringing New Chemistry to Life. *Angew Chem. Int. Ed. Engl. 57*, 4143–4148.

Ausländer, S., Ausländer, D. and Fussenegger, M. (2017). Synthetic Biology-The Synthesis of Biology. *Angew Chem. Int. Ed. Engl. 56*, 6396–6419.

Bauer, A. and Bogner, A. (2020). Let's (not) talk about synthetic biology: Framing an emerging technology in public and stakeholder dialogues. *Pub. Underst. Sci. 29*(1), 492–507. doi.org/10.1177/0963662520907255.

Bayley, H. (2019). Building blocks for cells and tissues. *Emerg. Top. Life Sci. 3*, 433–667

Beales, P.A., Ciani, B. and Mann, S. (2018). The artificial cell: biology-inspired compartmentalization of chemical function. *Roy. Soc. Interface Focus 8*, 20180046. doi.org/10.1098/rsfs.2018.0046

Buddingh, B.C. and van Hest, J.C.M. (2017). Artificial cells: Synthetic compartments with life-like functionality and adaptivity. *Acc. Chem. Res. 50*, 769–777.

De Lorenzo, V. (2018). Evolutionary tinkering vs. rational engineering in the times of synthetic biology. *Life Sci. Soc. Policy 14*, 18.

Fletcher, D.A. (2016). Bottom-up biology: Harnessing engineering to understand *Nature. Dev Cell. 38*, 587–589.

Göpfrich, K., Platzman, I. and Spatz, J.P. (2018). Mastering complexity: towards bottom-up construction of multifunctional eukaryotic synthetic cells. *Trends Biotechnol. 36*, 938–951.

Hutchison, C.A. 3rd, Chuang, R.Y., Noskov, V.N. et al. (2016). Design and synthesis of a minimal bacterial genome. *Science 351*(6280), aad6253.

Miller, T.E., Beneyton, T., Schwander, T. et al. (2020). Light-powered CO_2 fixation in a chloroplast mimic with natural and synthetic parts. *Science 368*, 649–654.

Porcar, M. and Peretó, J. (2016). Nature versus design: synthetic biology or how to build a biological non-machine. *Integr. Biol. (Camb) 8*, 451–455.

Rivas, G., Vogel, S.K. and Schwille, P. (2014). Reconstitution of cytoskeletal protein assemblies for large-scale membrane transformation. *Curr. Opin. Chem. Biol. 22*, 18–26.

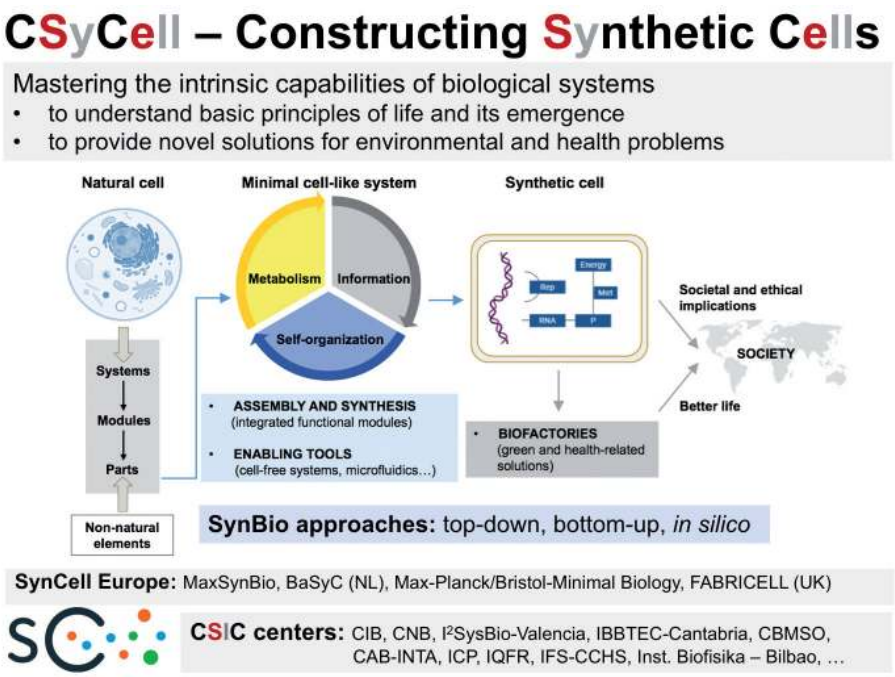
Sato, Y., Takinoue, M. (2019). Creation of artificial cell-like structures promoted by microfluidics technologies. *Micromachines 10*, 216; doi:10.3390/mi10040216

Shao, Y., Lu, N., Wu, Z. et al. (2018). Creating a functional single-chromosome yeast. *Nature 560*, 331–335.

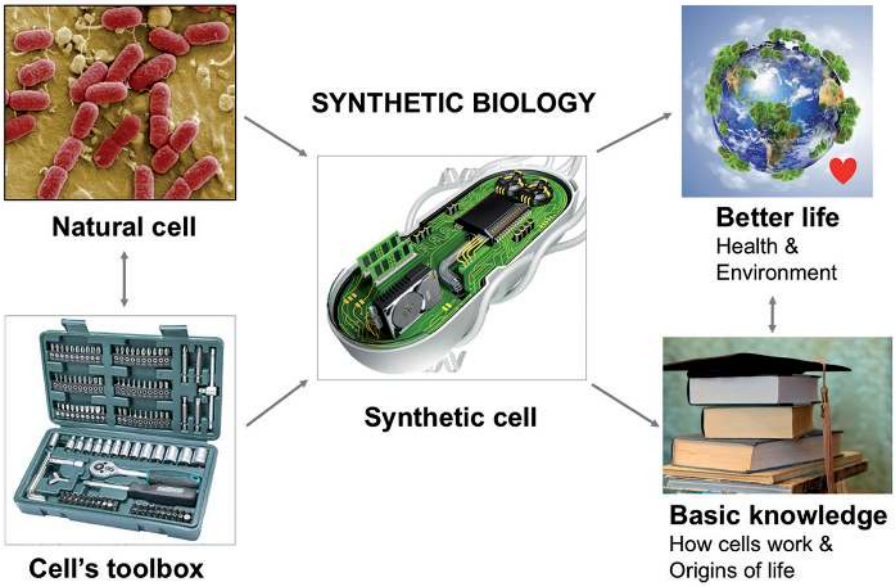
Schwille, P., Spatz, J., Landfester, K. et al. (2018). MaxSynBio: Avenues towards creating cells from the bottom up. *Angew Chem. Int. Ed. Engl. 57*, 13382–13392.

Yeates, T.O. (2017). Geometric principles for designing highly symmetric self-assembling protein nanomaterials. *Annu. Rev. Biophys. 46*, 23–42.

SUMMARY FOR EXPERTS



SUMMARY FOR THE GENERAL PUBLIC



How life appeared on Earth and how then it diversified into the different and currently existing forms of life are the unanswered questions that will be discussed this volume. These questions delve into the deep past of our planet, where biology intermingles with geology and chemistry, to explore the origin of life and understand its evolution, since “nothing makes sense in biology except in the light of evolution” (Dobzhansky, 1964). The eight challenges that compose this volume summarize our current knowledge and future research directions touching different aspects of the study of evolution, which can be considered a fundamental discipline of Life Science. The volume discusses recent theories on how the first molecules arouse, became organized and acquired their structure, enabling the first forms of life. It also attempts to explain how this life has changed over time, giving rise, from very similar molecular bases, to an immense biological diversity, and to understand what is the phylogenetic relationship among all the different life forms. The volume further analyzes human evolution, its relationship with the environment and its implications on human health and society. Closing the circle, the volume discusses the possibility of designing new biological machines, thus creating a cell prototype from its components and whether this knowledge can be applied to improve our ecosystem. With an effective coordination among its three main areas of knowledge, the CSIC can become an international benchmark for research in this field.

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
CSIC
EDITORIAL CSIC