Theme 3 – GENOME & EPIGENETICS

Coordinator: **Lluís Montoliu José** (CNB, Madrid)

Deputy coordinator: **Álvaro Rada Iglesias** (IBBTEC-CSIC/UNICAN, Santander)

Index Pages

Los autores de este volumen sobre Genoma y Epigenética dedicamos esta publicación a la memoria de José Luis Gómez Skarmeta (1966-2020), Profesor de Investigación del CSIC en el Centro Andaluz de Biología del Desarrollo en Sevilla

3.1 METHODS TO ANALYSE AND MODIFY THE GENOME

ABSTRACT

Our capacity to understand biological systems is restricted by our ability to identify, manipulate and control the genetic information. However, in the last years, great technical advances have extended our capabilities to predict, influence, and to "comprehend" genomic information in virtually every living organism. The ability to identify disease and deleterious genetic and epigenetic traits will reshape biological and biomedical research, and will have implications in how this technical and scientific revolution and the information it will generate may be used in decision-making, about how diseases are diagnosed and treated, and in the decision-making process in reproductive biology

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS

- María Domínguez (IN, Alicante, *Coordinator*)
- Pablo Huertas (CABIMER, Sevilla, *Deputy Coordinator*)
- Lluís Montoliu (CNB, Madrid
- José Pascual López Atalaya Martínez (IN, Alicante)
- Marian Ros (IBBTEC, Santander)
- Alberto M. Pendás (CIC, Salamanca)
- Marta Casado (IBV, Valencia)

Executive Summary

In the last two decades, we have witnessed major technological advances that have announced the advent of the genomic era. Great international efforts have provided us with the genomic information not only of many different species, but even now from different individuals within the same species. The development of the -omics techniques, genomic, transcriptomic, epigenomic, proteomic, metabolomic, etc., also provide us with multiple layers of information to analyse. And finally, a major breakthrough has been made when the possibility to alter virtually any genome of any species at will has been made possible with the identification and popularisation of genomic editing tools. Thus, currently we are in the verge of being able, for the first time, to really understand genomic information and manipulate it for basic research or specific applications.

In this chapter, we will revise the current state-of-the-art of the available methods to analyse and modify the genome. Moreover, we will identify the major challenges that this field is facing and should be tackled in the next years. Among them, we will highlight the need to implement and improve tools to extract and read the information from -omic data as well as to modify the genome.

Ultimately, we analyse CSIC strengths and weakness for research in this area. CSIC is in a unique position to contribute to it, thanks to its broad expertise in many different sciences, from physics to biology, from informatics to chemistry and in a wide variety of experimental models. Such multidisciplinary setup will benefit the development of these tools. Also, we recognize the existence of specific institutes and laboratories within CSIC that are experts and world leaders in specific topics required for understanding and further improve these technologies.

Introduction and general description

This chapter will cover our current understanding and future challenges of different research tools, mainly:

- *Computational strategies and bioinformatic pipelines*
- *Genetic and epigenetic editing tools*
- *New methods: single-cell technologies, etc.*

One of the key challenges of whole genome sequencing lies in the accurate assembly of large and complex genes, repetitive regions, and the annotation and assembly of all gene variants. These steps are all essential in order to comprehensively analyse complete genomes, transcriptomes, epigenomes and 3D genome organization. Computational methods must be optimized in order to allow the immediate and user-friendly access to all this -*omic* data. Defining functionally important sequences and epigenetic marks from genomic and epigenomic raw data is central to our understanding of biology, health and disease.

High-throughput functional genomic analyses, single-cell -omic technologies, and genome editing methodologies necessitate powerful and radical advances in bioinformatic and computational pipelines to be able to extract relevant information and make sense out of such large-scale datasets. Improved and more accurate genome editing, single-cell genetic and epigenetic editing, super-resolution imaging at single-cell, organ, and organism levels provide opportunities to accomplish the goal of significantly advancing genetic research in humans and other (model and non-model) organisms.

Advance in genome editing tools are revolutionizing genetic research and have notably advanced human and medical genetics. Human genome editing will soon become a reality for clinical therapeutics, particularly for ex-vivo therapies. Moreover, these editing tools are also improving genomic functional annotations. Functional annotation requires knowledge on the molecular, cellular, and organismal functions of each and every gene in a genome, and an understanding of the 'context' within which genes function and respond to environmental challenges, and the relationship between genetic and epigenetic, and ultimately phenotypic variation. Annotation of the human genome and those of model and non-model organisms and micro-organisms (e.g. microbiota) is providing unprecedented opportunities for biological interpretation of genomic function, evolution, diversity etc. Until recent years, the rate-limiting step in functional gene annotation has been the availability of mutants. The rapid and efficient genome editing technologies are boosting our ability to interrogate genomes in any

organism, allowing functional studies with an evolutionary perspective. These studies may include now species that could not be genetically manipulated previously. It can also, in some cases, allow us to do such gene editing experiments in a high throughput scale. In the past, genetic analysis of non-model organism has been highly complex, time-consuming, and limited to a few species. However, as the genomes from new species are sequenced (and annotated and assembled), the new genome editing tools will be immediately suitable for functional genetic analysis. The plethora of species that will be analysed in the short-term future will be useful to understand several biological processes, including pattern formation (diatoms, *Stentor*), branching morphogenesis, (*Physcomitrella*, *Ashbya*), regeneration (*Axolotl* and hydra), multicellularity (*volvox*), human development and disease (iPS-derived organoids), aging (killifish, naked mole-rat), cancer (dog, cat, naked mole-rat), complexity of parasite life cycles (malaria), coronavirus infections (organoids), hibernation (bats), hypoxic and cold adaption (groundhog), salinity stress (crops) to the most complex control of plagues by the most innovative gene drive strategies.

During the ongoing new era of biological research new tools for genome manipulation likely guided by Watson and Crick base-paring will emerge from metagenomic projects or will be synthetically designed/improved through the intervention of artificial intelligence (AI). Therefore, given the multitude of processes that will be analysed *de novo* for the first time, it is likely that we will witness an explosion of biological breakthroughs and techno-scientific advances with no historical precedent, whose transformational potential is difficult to foresee.

Genome Project-write (GP-write) is one such advances. GP-write will generate whole genome engineering of human cell lines and other organisms of agricultural and public health significance. The Human GP-write (HGP-write) will focus on synthesizing human genomes in whole or in part and will work in cell and the organoids derived from them. The main goal of GP-write is to expand the genetic engineering tools available and to generate information connecting the sequence of nucleotide bases in DNA with their physiological properties and functional behaviours for applications in healthcare, energy, agriculture, or bioremediation.

The ability to decipher the information encoded in the genome and epigenome and to manipulate this information with precision needs single cell and high content technologies. These technologies will revolutionize how genome information can be translated into precise and personalized medicine and healthcare and how the genome content can be manipulated to define and understand gene function, genetic networks and gene-environment interactions. One important challenge will be to use novel genomic editing tools, such as CRISPR-Cas9,

TALENs and the zinc finger nucleases (ZFNs), for the manipulation of single- cell genomes, as this can provide major insights into how a single cell can influence its neighbouring cells within the same tissues, in other tissues or in the context of the whole organism.

The single-cell technologies are becoming an essential tool in biological studies. These techniques are giving us the opportunity to both study cellular heterogeneity and to unmask previously obscured cellular populations. Single-cell technologies coupled with high-resolution imaging such as super-resolution imaging, mass spectrometry, and deep sequencing will enable to analyse, identify, and reveal cellular subtypes and rare cell types to study in depth cell, organ biology and pathology. These technologies hold the potential to unfold the continuous dynamic changes in cell type/state along biological processes such as differentiation, immune response, or cancer expansion. Single cell and high content approaches also offer some challenges and limitations. For example, some single-cell epigenomics methodologies are emerging but the full potential of them is still unclear. Single-cell epigenomic profiling of cancer cells combined with other gene editing and genomic analysis (single-cell-ATAC, CUT&Tag, or scTrio-seq) will revolutionize epigenetic analysis and may be also useful to modify rare cells, such as cancer stem cells, and metastatic cells. Validated animal models need to be developed side by side with in vitro model platforms providing the opportunity to investigate multicellular interactions and dynamic multistep processes. In that sense organ-on-a-chip (OOC) technology has evolved from a combination of various engineering platforms to address the difficulties of conventional drug testing models. Organoid cultures were a major breakthrough in the *in vitro* culture of tumour cells from patients, and are becoming now the most attractive tool to be used as an *in vitro* screening platform

Here is a summary of the main challenges we can foresee, which will be developed further in the following sections:

- *Challenges in genome editing tools*: i) Methods have to be robust, with a high efficiency and limited off targets effects. ii) Beyond CRISPR Gene Editing: efforts are being devoted to develop newer and safer genome-editing systems. iii) Methodologies need to enable to modify any locus, regardless of the position, structure and neighbouring sequences and in isolated cells or in the organism in somatic and germline cells.

- *Implementation of new methods*: , i) precise manipulation of the genome content of any cell, *in vivo* and in cell culture and ii) powerful bioinformatics to make sense of information stored in genomes, and the effect of manipulation to advance research addressing animal and human biology, behaviour, diseases such as cancer, metabolic disorders, degenerative disease, and longevity; iii) organoid cultures combined with novel developments in live imaging, genetic engineering and biomaterials represent a tour de force that will influence, in the very near future, how we will study human development and how we will treat human disease.

Impact in basic science panorama and potential applications

The opportunity to precisely read, analyse, interpret and finally control at will the genome and the epigenome of any organism seems feasible in the near future. The acquisition and further development of those abilities will result in a major technical and scientific revolution that will impact equally the basic and applied sciences. It will also bring about ethical concerns about how this genetic information and its manipulation might be useful in decision making about therapies and to identify, and eliminate, deleterious genetic traits at early embryonic stages. The latter will only be possible after improving current methods and deciding whether the laws currently prohibiting, in Spain and many other countries, the irreversible alteration of the human embryo genome should be modified.

Basic researchers will benefit enormously from these advances in genomic and epigenomic to gather new information and clues that will help to comprehend biological processes in a deeper and more thoroughly manner. We will be able to: interrogate full genomes and to search and find *in silico* the genes or genetic sequences that are more likely involved in controlling specific processes or pathological conditions; extract meaningful information of natural existing variants; analyse and predict how the genetic information is controlled and influenced by the environment and experiences, both at the level of DNA sequences and also at the epigenetic and 3D chromatin organization levels. Knowing and understanding this genetic information will help us to elucidate how the genetic and epigenetic variants are translated into changes in function, performance, and behaviour through specific proteins, noncoding RNAs, or their interplay. Importantly, increasingly and inexpensively sequencing of an animal or a person's entire genome could be combined with precise genomic or epigenomic editing, which will be instrumental in the development of precision and personalized medicine. The ability to demonstrate cause-and-effect will also revolutionise and guide personalized treatments. Single-cell and high-resolution techniques

will also advance the current understanding of multiple biological processes or diseases, helping us to understand natural and pathological biological heterogeneity and dynamics. Therefore, these technological advances will let us observe processes that were virtually invisible to us just a few years ago.

The improvement of genomic and epigenomic technologies, and the analysis at single-cell resolution, will extend these abilities to virtually any organism, not only model organisms. Thus, new and old biological questions could be reformulated and addressed with the new technologies. CRISPR/Cas9 mediated genomic editing, for example, has been already successfully applied in many different animal taxa, fungi and plants, and microorganisms (El-Mounadi et al., 2020; Lee et al., 2020; Schuster and Kahmann, 2019; Sun et al., 2017; Zhang et al., 2018). Moreover, new developments allow now to manipulate genome regulatory regions, the epigenome, the non-coding genome, to identify chromatin interactions, and tag specific loci for live-cell imaging and analyses of the genome (Pickar-Oliver and Gersbach, 2019). The information of physical interactions between loci and distal located genomic region through long-range chromatin interactions is also essential to understand the dynamics of 3D chromatin organization during differentiation and in response to growth, differentiation factors and hormones (Gutierrez et al., 2019). Systematic mapping of protein and RNA-chromatin interactions (Sridhart et al., 2017) and the ease to adapt the technologies to genome (or epigenome) wide screens will be powerful tool for the research of complex phenotypes.

The widespread use of these novel methodologies will also create a huge library of research tools available for the community. As an example, the aforementioned Genome Project-write (GP-write) will represent one of the many advances in that direction. This international research aims to minimize the cost of large-scale genome engineering using both genomic editing but also large-scale synthesis in cell lines from multiple organisms. Both, the technology developed under the GP-write project and the cell lines will be a very valuable resource for the community. Similar international approaches will provide a solid reservoir of genetic variants for research.

Regarding the potential applications of those tools, the limitation is strictly in the imagination of the researchers. Improved methodologies to analyse the genome will help us to understand human health and disease, how plants impact and adapt to the environment, the interaction between host and pathogens, the biology of the microbiota in the soil and the gut, the emergence and evolution of novel diseases such as the SARS, MERS and COVID19-pandemia. The ability to predict and control the genomic and epigenomic information will

depend on our ability to precisely modulate the genome, holding promise to be able to repair or restore gene function associated with certain pathologies (Doudna, 2020; Pickar-Oliver and Gersbach, 2019), editing the microbiome (Menchaca et al., 2020), creating genetically altered improved crops (Doudna, 2020), bettering veterinary treatments thus positively impacting the welfare and health of pets and livestock (Menchaca et al., 2020), etc. They will also facilitate the generation of animal mutants in order to understand biological processes and model human disease. These capabilities will also require international agreements to update the ethical guidelines regulating genomic research, which should take into consideration the possible risks and benefits to the human society and the environment. Overall, we are witnessing the emergence of new tools that will drastically increase our ability to decipher and modify genetic and epigenetic information. This will certainly revolutionize several research fields that will be covered in more detail in following Chapters.

Key challenging points

I. Computational strategies and bioinformatic pipelines

1) Towards a more accurate definition of a gene:

Since the term "gene" was first coined, over a hundred years ago, its definition has been evolving to keep up with our knowledge. From the original and abstract representation of a hereditary unit used by Johannsen in the early 20th century, to the "one gene—one mRNA—one polypeptide" of the 60s, to the more molecular concept of a localized nucleotide sequence of DNA that will code for an RNA, regardless of whether is translated or not into one or several polypeptides. Thus, the term "gene" has become looser, and its current molecular definition has become more complex. One of the greatest challenges in terms of bioinformatic analysis will be to acquire the capacity to find and define genes out of a raw genomic sequence, and to search and locate its critical elements (promoter and transcription termination sites, introns and exons and regulatory elements). This is even more complex in the case of nested genes and genes covered by multiple noncoding genes and mobile genetic elements. Those analyses should enable the description of all the RNA variant(s) produced by any given gene, predict their coding sequence(s) (in the case of mRNAs), their structures (in the case of structural RNAs) or their functions (in the case of mRNAs and ncRNAs).

2) Towards the creation of tools that improve the understanding of gene expression control:

Biological system responses rely greatly in how the different genomic loci are regulated intrinsically and in response to environmental cues. Thus, it is essential the development of

new tools that will allow us to extract from the genomic raw data not only the actual sequences of genes, but also to predict how those genes could be regulated. Sequence information has been complemented with full characterization of chromatin proteins and their modifications. The development and combination of genome-wide techniques such as ChIP-seq, ATAC-seq, BiS-seq is helping us to catalogue chromatin proteins and their modifications. The development of novel computational analyses to these data allow segregating this complexity into discrete numbers of chromatin states that in turn, is revealing how chromatin directs functions such as transcription and RNA, processing, and, crucially, how chromatin biology contributes to disease. New methods at single cell resolution could even further discriminate transcriptional states among individual cells (see Chapters 4 and 5 for more details on these topics).

3) Towards genome data analysis in real-time powered by ease-to-use bioinformatic tools:
The use of bioinformatic tools is rapidly extending in biological and biomedical research. However, there are deep limitations on their use. First, -omics experiments produce a staggering amount of information that have to be safely stored, in a format and support that is both easy and rapidly retrievable. With the sharp and steady accumulation of data, these will require hardware and software improvement in the next few years. These are advances that should essentially come from informatic engineers. Second, one of the restrictions of many bioinformatic tools is that they require a steep learning curve. Although some tools are easily available and user-friendly, many of them are continuously developing and require deep informatic knowledge. In the future, we predict those methods and software tools will evolve toward more efficient, more flexible choices, and more user-friendly protocols, therefore spreading the use of bioinformatic tools in research. These topics will be extensively covered in Chapter 2.

II. Genome editing tools

1) Towards robust, efficient and specific methods:
Genome editing tools should be highly efficient and specific, ideally with limited or no off-target effects. These are the major challenges in efficient genomic editing that we are currently facing, and a great effort is being made in the search for solutions. Improvement of those aspects require a deeper understanding of how DNA repair takes place, as genome editing is based on tricking the cell to alter specific genomic sequences during such processes. A myriad of modifications of the CRISPR/Cas technologies have been proposed or are in the

pipeline to modify the system to increase its efficiency and/or robustness. These modifications vary, including modifications of the original CRISPR/Cas9 system itself (Kleinstiver et al., 2016), fusion to specific DNA repair proteins (Charpentier et al., 2018; Jayavaradhan et al., 2019; Rees et al., 2019), and the temporal tampering of the repair processes (Jinek et al., 2013; Wienert et al., 2020; Xu et al., 2020a). Also, currently a great emphasis has been put in the development of new tools, either by finding new natural and more efficient genomic editing systems (see below) or by directly altering the genomic editing enzymes *in vitro*, such as creating CRISPR/Cas9 derivatives that do not break the DNA (Anzalone et al., 2019; Komor et al., 2016).

2) Beyond CRISPR Gene Editing:

Before the emergence of CRISPR/Cas9 technology (Jinek et al., 2012), several other approaches, such as Zinc Finger proteins (ZNF) or TALEN nucleases, were implemented for gene editing purposes (Xu et al., 2020b). All of them have in common the exploitation of molecular biology tools already available in nature or derived from our understanding of how some nuclear proteins operate. Thus, it is foreseeable that we will discover new options in the next years just by looking in different organisms. Already, alternatives to the most widely used *Streptococcus pyogenes* modified SpCas9 enzyme (Jinek et al., 2012; Kleinstiver et al., 2016) are in use. Shorter Cas9 proteins, from *Staphylococcus aureus* (SaCas9), *Neisseria meningitidis* (NmCas9), *Streptococcus thermophilus* (St1Cas9) or *Brevibacillus laterosporus* (BlatCas9), have been found an successfully applied for gene editing (Xu et al., 2020b). Additionally, orthologs of this family of proteins with potential application for gene editing have been isolated from different bacteria. Besides the different Cas9 orthologs, other Cas proteins have been discovered as well, including Cpf1 (CRISPR from *Prevotella* and *Francisella* 1 also known as Cas12a), and the various Cas13 and Cas14 variants (Abudayyeh et al., 2016; Konermann et al., 2018; Strecker et al., 2019; Zetsche et al., 2015), some of them leading to new innovative applications such as genetic diagnostic (Gootenberg et al., 2017). Moreover, other enzymes unrelated with the Cas proteins, such as CasX enzymes, have also been proposed to be able to perform RNA-guided genome editing (Liu et al., 2019). Thus, the incorporation or alternative Cas9-type enzymes, or even the discovery of completely new systems to perform gene editing will render in the near future a plethora of genomic editing tools with distinct capabilities.

3) Towards more flexible and general genome editing tools:

Methodologies should enable the modification of any locus, regardless of the position, structure and neighbouring sequences, and in isolated cells or in the somatic or germline cells of whole organisms. One of the greatest challenges for gene editing tools has been to extend their applicability beyond *in vitro* cell culture, thus making them useful in medical applications and the creation of genetically modified organisms (Doudna, 2020; Lee et al., 2020; Menchaca et al., 2020; Seruggia et al., 2015; Sun et al., 2017; Xu et al., 2019; Zhang et al., 2018). Given the development and advantages of genome-editing technologies, research that uses genome editing to improve horticultural crops has substantially increased in recent years. The combination of rapidly advancing genome-editing technology with breeding will greatly increase horticultural crop production and quality (Xu et al., 2019). In model organisms such as the fruit fly *Drosophila melanogaster*, the ability of researchers to engineer targeted genome modifications for studies of genes and genetic elements has significantly been transformed by the generation of transgenic flies expressing Cas9 and modified Cas9 variants (Gratz et al. 2013; Ewen-Campen et al 2017) and transgenes to express synthetic guide RNAs (sgRNAs) for gene disruption, deletion, or gene activation. In the near future more precise delivery of the genomic editing tools in animals will be available (Lino et al., 2018). However, genome editing tools still suffer from a very heavy sequence-context component that limit the sequences that can be efficiently targeted. Thus, it will be required to be able to relax or break such constrains in order to modify any given sequence, regardless of the genomic context. In this regard, it will be also relevant the development of new bioinformatic methods to improve the design and selection of more efficient and unique RNA guides to be used in combination with CRISPR-Cas technologies (Oliveros et al., 2016; Torres-Perez et al., 2019).

III. New methods to analyze and visualize the genome

1) Organoid and 3D cultures:

An organoid is a 3D structure derived from stem cells of organ-specific cell types that self-organizes and resembles the physiological characteristic of that organ (Clevers, 2016; Yin et al., 2016). The importance of these organ-in-a-dish cultures is that they can recapitulate and eventually, upon improvement, will mimic the natural microenvironment of an organ, allowing researchers to pose more complex questions regarding the function of the human (epi)genome during development or in response to different stimuli. Moreover, in biomedical research, they have the potential to more accurately model human diseases through the use of patient-derived pluripotent stem cells (Yin et al., 2016) or, in the future, be a source of organs

for transplantation in regenerative medicine. Human organoids can also overcome the genetic and gene dosage differences that some times exist between humans and model organisms and that can complicate the study or certain human disorders. The actual challenge, however, is to create the organoids themselves. Although organoids for multiple organs have been successfully created in the lab, they still lack the complexity and organization of real organs. The improvement of organoids is, therefore, a major challenge for the near future. Moreover, organoids are still difficult to work with, time- and labour-consuming, more expensive that traditional 2D cultures and difficult to study with standard techniques (for example, attempting to apply imaging devices becomes complicated) (Jensen and Teng, 2020).

2) Novel developments in live imaging:

Although the relationship between imaging techniques and genomic tools is not obvious, there are interesting synergies between them that would have to be explored in the future. As mentioned, genome editing can provide of new tools for live imaging by helping to tag and visualize specific DNA regions, which, for example can improve the current understanding of 3D genome organization and overcome some of the limitations of bulk genomic approaches (*e.g.* heterogeneity, dynamics) (revised in Pickar-Oliver and Gersbach, 2019). On the other hand, the development of high-content high-throughput microscopy could complement nicely the data obtained by single-cell -omics. Also, it will be relevant to streamline pipelines that connect genome wide screens with microscopy-based outcomes.

3) Single cell –omics:

The great advantage of performing OMICs in single cells is allowing researchers to account for the natural heterogeneity of the biological systems, as instead of averaging the signal of multiple cells into a single output, the information of individual cells is kept. We currently possess the ability to interrogate single cells at different -omic layers, either the genome, the transcriptome, the epigenome or, the 3D genome (Chappell et al., 2018). However, there are still challenges to be solved in the future. Some -omics technologies are still difficult to perform at the single cell resolution due to limited biological material, so there is still room for improvement on that direction. Newly developed methods are being implemented to allow high throughput gene expression mapping at the single-cell level within tissues (spatial transcriptomics). Also, they are still expensive, and improvement in the computational side for better analysis will be required. Finally, the combination of several of those technologies

in single-cell multiomics will see further development in the near future (Chappell et al., 2018).

CSIC advantage position and multi/interdisciplinarity

The challenge posed in this section represent an ambitious prediction and a major tour-de-force in the novel methodologies that will dominate biological and biomedical research in the next decades. These technological aspect means that, differently to other challenges posed in this White Paper, the strength of CSIC does not rely only in the presence of specific research groups or institutes, but in its appealing to a great majority of them. The advances presented here will benefit, and will stem from the needs, of almost any biology or biomedicine laboratory, but will also attract others in disciplines like chemistry, physics, computational sciences, etc. Thus, CSIC, with its broad approach to science, and its transversal research is in a great position to foster the interactions and provide the environment where this technological research will flourish. In addition to the multidisciplinary approach, biological and biomedical research of CSIC research centres and institutes covers a wide field, with groups working in many areas of biology using many biological systems and disciplines. Thus, CSIC accommodate a huge variety of prospective applications and interests that cater specifically to this type of major of technological breakthrough. So, advances in the genomics and epigenomics tools can be expected in crop development, livestock welfare, biomedical research, microbiology, and on curiosity-driven basic research in a broad range of living organisms. This is a strength that only major research institutions like CSIC can provide. Furthermore, in addition to the widespread interest that the development of these tools will have, CSIC also host specific laboratories that focus in aspects of great relevance in the development of these technologies. For example, simply considering the authors of this Chapter, we can already identify labs devoted to the study of epigenetics (María Domínguez (Gutierrez,et at 2019); Angel Barco (Fernandez-Albert et al., 2019)), transcriptional regulation (Marta Casado (Inserte et al., 2009; Moncayo-Arlandi et al., 2016; Motiño et al., 2019; Remesal et al., 2020); José Lopez-Atalaya (Lipinski et al., 2020); Marian Ros (Bastida et al., 2020; Saiz-Lopez et al., 2015); (Lluís Montoliu (Seruggia et al., 2015); Jaime Carvajal (Vicente-García et al., 2017)), genome 3D structure (see chapter D3.3 for details), DNA repair (Pablo Huertas (Jimeno et al., 2019; López-Saavedra et al., 2016; Soria-Bretones et al., 2017), Alberto M. Pendás (Hellmuth et al., 2018)), genetics of inheritance (Alberto M. Pendás (Caburet et al., 2014; Gómez-H et al., 2016, 2019)). Moreover, many of our labs routinely use

OMICs technologies and bioinformatic tools (María Dominguez (Vallejo et al. 2015; Villegas et al., 2018; Oswald et al. 2016), José Lopez-Atalaya (Lipinski et al., 2020)). The ethics of genome editing methods is also one of the future challenges in this area where CSIC is also well positioned (Lluis Montoliu (Chneiweiss et al., 2017; Hirsch et al., 2019; Montoliu et al., 2018))

Plan and resources

In order to maintain and enhance the international impact of the CSIC in this area in the next years, a series of steps might be implemented:

1) CSIC should build and develop a hub for Next Generation Sequencing infrastructures in order to offer easy access to state-of-the-art, well maintained sequencing facilities to ICUs. It is of utmost importance for CSIC to prioritize large and sustained investments in sequencing facilities to foster the incorporation of genomics to its research laboratories and groups. CSIC should be involved in international initiatives such as EASI Genomics (https://www.easi-genomics.eu/). CSIC must prioritize participation and national leadership in global initiatives such as Human Cell Atlas (https://www.humancellatlas.org/), ENCODE (https://www.encodeproject.org/), 4D nucleome (https://www.4dnucleome.org). These efforts are instrumental to help secure CSIC's position at the forefront of genomics and biomedical research in the coming decade.

2) The creation of a multidisciplinary platform for the development of genomic tools. This will require:

a. To identify and bring together laboratories with expertise in this area to discuss and coordinate research lines and platforms in strategic institutes of CSIC. This will also foster and support collaborations across CSIC institutions.

b. To provide financial support for such projects after external peer revision.

c. The promotion of those projects that show more promise in national and international/European calls.

3) The collection and curation of the resources, both future and already available, within CSIC labs in genome editing tools. This can be done with the simple implementation of a web-based platform at CSIC intranet that lists all those resources. Which such an easy and cheap approach, all CSIC research labs could immediately identify prospective collaborators or find neighbouring labs that can help with the implementation of those technologies.

4) The creation of one or several genomic/epigenomic/computational biology support unit(s), that enables the safe storage of raw data, compiles major bioinformatic tools, but also support

personnel to help CSIC researchers. This unit will also support deployment of project-associated web applications from dedicated servers. This can be implemented as a central support unit, or within a specific or several research institute(s) but available to all CSIC members.

5) New recruitments and stabilization of researchers with the demonstrated expertise and/or potential to implement and develop these tools. For example, positions at the level of "Cientifico Titular", "Investigador Cientifico" and "Profesor de Investigacion" to recruit national and international researchers to lead new research groups and research units in this area. The candidate should have strong background in the fields of epigenetic, transcriptional regulation, genome 3D structure, DNA repair, gene editing, or working experience in computational biology, bioinformatics, biomedical engineer, or related fields.

References

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

Anzalone, A. V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature *576*, 149–157.

Bastida, M.F., Pérez-Gómez, R., Trofka, A., Zhu, J., Rada-Iglesias, A., Sheth, R., Stadler, H.S., Mackem, S., and Ros, M.A. (2020). The formation of the thumb requires direct modulation of Gli3 transcription by Hoxa13. Proc. Natl. Acad. Sci. U. S. A. *117*, 1090–1096.

Caburet, S., Arboleda, V.A., Llano, E., Overbeek, P.A., Barbero, J.L., Oka, K., Harrison, W., Vaiman, D., Ben-Neriah, Z., García-Tuñón, I., et al. (2014). Mutant cohesin in premature ovarian failure. N. Engl. J. Med. *370*, 943–949.

Chappell, L., Russell, A.J.C., and Voet, T. (2018). Single-Cell (Multi) omics Technologies.

Charpentier, M., Khedher, A.H.Y., Menoret, S., Brion, A., Lamribet, K., Dardillac, E., Boix, C., Perrouault, L., Tesson, L., Geny, S., et al. (2018). CtIP fusion to Cas9 enhances transgene integration by homology-dependent repair. Nat. Commun. *9*, 1–11.

Chneiweiss, H., Hirsch, F., Montoliu, L., Müller, A.M., Fenet, S., Abecassis, M., Merchant, J., Baertschi, B., Botbol-Baum, M., Houghton, J.A., et al. (2017). Fostering responsible research with genome editing technologies: a European perspective. Transgenic Res. *26*, 709–713.

Clevers, H. (2016). Modeling Development and Disease with Organoids. Cell *165*, 1586–1597.

Doudna, J.A. (2020). The promise and challenge of therapeutic genome editing. Nature *578*, 229–236.

El-Mounadi, K., Morales-Floriano, M.L., and Garcia-Ruiz, H. (2020). Principles, Applications, and Biosafety of Plant Genome Editing Using CRISPR-Cas9. Front. Plant Sci. *11*, 1–16.

Ewen-Campen B, Yang-Zhou D, Fernandes VR, et al. (2017). Optimized strategy for in vivo Cas9-activation in *Drosophila*. Proc Natl Acad Sci U S A. 114(35):9409-9414.

Gratz SJ, Wildonger J, Harrison MM, O'Connor-Giles KM.(2013). CRISPR/Cas9-mediated genome engineering and the promise of designer flies on demand. Fly (Austin). 2013 Oct-Dec;7(4):249-55. doi: 10.4161/fly.26566. Epub 2013 Oct 2.

Fernandez-Albert, J., Lipinski, M., Lopez-Cascales, M.T., Rowley, M.J., Martin-Gonzalez, A.M., Del Blanco, B., Corces, V.G., and Barco, A. (2019). Immediate and deferred epigenomic signatures of in vivo neuronal activation in mouse hippocampus. Nat. Neurosci. *22*, 1718–1730.

Gómez-H, L., Felipe-Medina, N., Sánchez-Martín, M., Davies, O.R., Ramos, I., García-Tuñón, I., de Rooij, D.G., Dereli, I., Tóth, A., Barbero, J.L., et al. (2016). C14ORF39/SIX6OS1 is a constituent of the synaptonemal complex and is essential for mouse fertility. Nat. Commun. *7*, 13298.

Gómez-H, L., Felipe-Medina, N., Condezo, Y.B., Garcia-Valiente, R., Ramos, I., Suja, J.A., Barbero, J.L., Roig, I., Sánchez-Martín, M., de Rooij, D.G., et al. (2019). The PSMA8 subunit of the spermatoproteasome is essential for proper meiotic exit and mouse fertility. PLoS Genet. *15*, e1008316.

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., et al. (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science *356*, 438–442.
Gutierrez-Perez I, Jordan-Rowley J, Lyu X, Valadez-Graham V, Vallejo DM, Ballesta-Illan E, Lopez-Atalaya JP, Kremsky I, Caparros E, Corce VG, Dominguez M. (2019). Ecdysone-Induced 3D Chromatin Reorganization Involves Active Enhancers Bound by Pipsqueak and Polycomb. Cell Rep. *28*;2715-2727.e5.
Hellmuth, S., Gutiérrez-Caballero, C., Llano, E., Pendás, A.M., and Stemmann, O. (2018). Local activation of mammalian separase in interphase promotes double-strand break repair and prevents oncogenic transformation. EMBO J. *37*.
Hirsch, F., Lemaitre, C., Chneiweiss, H., and Montoliu, L. (2019). Genome Editing: Promoting Responsible Research. Pharmaceut. Med. *33*, 187–191.
Inserte, J., Molla, B., Aguilar, R., Través, P.G., Barba, I., Martín-Sanz, P., Boscá, L., Casado, M., and Garcia-Dorado, D. (2009). Constitutive COX-2 activity in cardiomyocytes confers permanent cardioprotection Constitutive COX-2 expression and cardioprotection. J. Mol. Cell. Cardiol. *46*, 160–168.
Jayavaradhan, R., Pillis, D.M., Goodman, M., Zhang, F., Zhang, Y., Andreassen, P.R., and Malik, P. (2019). CRISPR-Cas9 fusion to dominant-negative 53BP1 enhances HDR and inhibits NHEJ specifically at Cas9 target sites. Nat. Commun. *10*, 1–13.
Jensen, C., and Teng, Y. (2020). Is It Time to Start Transitioning From 2D to 3D Cell Culture? Front. Mol. Biosci. *7*, 1–15.
Jimeno, S., Mejías-Navarro, F., Prados-Carvajal, R., and Huertas, P. (2019). Controlling the balance between chromosome break repair pathways (Elsevier Inc.).
Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-.). *337*, 816–821.
Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2013*, 1–9.
Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.
Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.
Konermann, S., Lotfy, P., Brideau, N.J., Oki, J., Shokhirev, M.N., and Hsu, P.D. (2018). Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell *173*, 665-676.e14.
Lee, H., Yoon, D.E., and Kim, K. (2020). Genome editing methods in animal models. Animal Cells Syst. (Seoul). *24*, 8–16.
Lino, C.A., Harper, J.C., Carney, J.P., and Timlin, J.A. (2018). Delivering crispr: A review of the challenges and approaches. Drug Deliv. *25*, 1234–1257.
Lipinski, M., Muñoz-Viana, R., Del Blanco, B., Marquez-Galera, A., Medrano-Relinque, J., Caramés, J.M., Szczepankiewicz, A.A., Fernandez-Albert, J., Navarrón, C.M., Olivares, R., et al. (2020). KAT3-dependent acetylation of cell type-specific genes maintains neuronal identity in the adult mouse brain. Nat. Commun. *11*, 2588.
Liu, J.J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019). CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature *566*, 218–223.
López-Saavedra, A., Gómez-Cabello, D., Domínguez-Sánchez, M.S., Mejías-Navarro, F., Fernández-Ávila, M.J., Dinant, C., Martínez-Macías, M.I., Bartek, J., and Huertas, P. (2016). A genome-wide screening uncovers the role of CCAR2 as an antagonist of DNA end resection. Nat. Commun. *7*.
Menchaca, A., Santos-neto, P.C., Mulet, A.P., and Crispo, M. (2020). CRISPR in livestock : From editing to printing. Theriogenology.
Moncayo-Arlandi, J., Guasch, E., Sanz-de la Garza, M., Casado, M., Garcia, N.A., Mont, L., Sitges, M., Knöll, R., Buyandelger, B., Campuzano, O., et al. (2016). Molecular disturbance underlies to arrhythmogenic cardiomyopathy induced by transgene content, age and exercise in a truncated PKP2 mouse model. Hum. Mol. Genet. *25*, 3676–3688.
Montoliu, L., Merchant, J., Hirsch, F., Abecassis, M., Jouannet, P., Baertschi, B., Sarrauste de Menthière, C., and Chneiweiss, H. (2018). ARRIGE Arrives: Toward the Responsible Use of Genome Editing. Cris. J. *1*, 128–129.
Motiño, O., Francés, D.E., Casanova, N., Fuertes-Agudo, M., Cucarella, C., Flores, J.M., Vallejo-Cremades, M.T., Olmedilla, L., Pérez Peña, J., Bañares, R., et al. (2019). Protective Role of Hepatocyte Cyclooxygenase-2 Expression Against Liver Ischemia-Reperfusion Injury in Mice. Hepatology *70*, 650–665.
Oliveros, J.C., Franch, M., Tabas-Madrid, D., San-León, D., Montoliu, L., Cubas, P., and Pazos, F. (2016). Breaking-Cas-interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes. Nucleic Acids Res. *44*, W267-71.
Pickar-Oliver, A., and Gersbach, C.A. (2019). The next generation of CRISPR–Cas technologies and

applications. Nat. Rev. Mol. Cell Biol. *20*, 490–507.
Rees, H.A., Yeh, W.H., and Liu, D.R. (2019). Development of hRad51–Cas9 nickase fusions that mediate HDR without double-stranded breaks. Nat. Commun. *10*.
Remesal, L., Roger-Baynat, I., Chirivella, L., Maicas, M., Brocal-Ruiz, R., Pérez-Villalba, A., Cucarella, C., Casado, M., and Flames, N. (2020). PBX1 acts as terminal selector for olfactory bulb dopaminergic neurons. Development *147*, dev186841.
Saiz-Lopez, P., Chinnaiya, K., Campa, V.M., Delgado, I., Ros, M.A., and Towers, M. (2015). An intrinsic timer specifies distal structures of the vertebrate limb. Nat. Commun. *6*, 8108.
Schuster, M., and Kahmann, R. (2019). CRISPR-Cas9 genome editing approaches in filamentous fungi and oomycetes. Fungal Genet. Biol. *130*, 43–53.
Seruggia, D., Fernández, A., Cantero, M., Pelczar, P., and Montoliu, L. (2015). Functional validation of mouse tyrosinase non-coding regulatory DNA elements by CRISPR-Cas9-mediated mutagenesis. Nucleic Acids Res. *43*, 4855–4867.
Soria-Bretones, I., Cepeda-García, C., Checa-Rodriguez, C., Heyer, V., Reina-San-Martin, B., Soutoglou, E., and Huertas, P. (2017). DNA end resection requires constitutive sumoylation of CtIP by CBX4. Nat. Commun. *8*.
Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong Sheng (2017). Systematic Mapping of RNA-Chromatin Interactions In Vivo. Curr Biol, *27 (4)*, 610-612.
Sun, D., Guo, Z., Liu, Y., and Zhang, Y. (2017). Progress and prospects of CRISPR/Cas systems in insects and other arthropods. Front. Physiol. *8*, 1–22.
Strecker, J., Jones, S., Koopal, B., Schmid-Burgk, J., Zetsche, B., Gao, L., Makarova, K.S., Koonin, E. V., and Zhang, F. (2019). Engineering of CRISPR-Cas12b for human genome editing. Nat. Commun. *10*.
Sun, D., Guo, Z., Liu, Y., and Zhang, Y. (2017). Progress and prospects of CRISPR/Cas systems in insects and other arthropods. Front. Physiol. *8*, 1–22.
Torres-Perez, R., Garcia-Martin, J.A., Montoliu, L., Oliveros, J.C., and Pazos, F. (2019). CRISPR Tools-Live Repository of Computational Tools for Assisting CRISPR/Cas Experiments. Bioeng. (Basel, Switzerland) *6*.
Vicente-García, C., Villarejo-Balcells, B., Irastorza-Azcárate, I., Naranjo, S., Acemel, R.D., Tena, J.J., Rigby, P.W.J., Devos, D.P., Gómez-Skarmeta, J.L., and Carvajal, J.J. (2017). Regulatory landscape fusion in rhabdomyosarcoma through interactions between the PAX3 promoter and FOXO1 regulatory elements. Genome Biol. *18*, 106.
Wienert, B., Nguyen, D.N., Guenther, A., Feng, S.J., Locke, M.N., Wyman, S.K., Shin, J., Kazane, K.R., Gregory, G.L., Carter, M.A.M., et al. (2020). Timed inhibition of CDC7 increases CRISPR-Cas9 mediated templated repair. Nat. Commun. *11*, 2109.
Xu, J., Hua, K., and Lang, Z. (2019). Genome editing for horticultural crop improvement. Hortic. Res. *6*.
Xu, S., Kim, J., Tang, Q., Chen, Q., Liu, J., Xu, Y., and Fu, X. (2020a). CAS9 is a genome mutator by directly disrupting DNA-PK dependent DNA repair pathway. Protein Cell *11*, 352–365.
Xu, X., Hulshoff, M.S., Tan, X., Zeisberg, M., and Zeisberg, E.M. (2020b). Crispr/cas derivatives as novel gene modulating tools: Possibilities and in vivo applications. Int. J. Mol. Sci. *21*.
Yin, X., Mead, B.E., Safaee, H., Langer, R., Karp, J.M., and Levy, O. (2016). Engineering Stem Cell Organoids. Cell Stem Cell *18*, 25–38.
Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., Van Der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.
Zhang, Z.-T., Jiménez-Bonilla, P., Seo, S.-O., Lu, T., Jin, Y.-S., Blaschek, H.P., and Wang, Y. (2018). Bacterial Genome Editing with CRISPR-Cas9: Taking Clostridium beijerinckii as an Example. In Synthetic Biology: Methods and Protocols, J.C. Braman, ed. (New York, NY: Springer New York), pp. 297–325.

3.2 OMICS TECHNOLOGIES AND PRECISION MEDICINE

ABSTRACT

High-throughput omics technologies are called to revolutionize medical practice by the general implementation of precision and personalized medicine (PPM) approaches to tailor diagnostic, therapeutic and monitoring strategies for individual patients based on their genetic and molecular signatures.

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS (in alphabetical order)

- Javier de las Rivas (CIC-IBMCC, Salamanca, *Coordinator*)
- Leticia Mora (IATA, Valencia)
- Lluís Montoliu (CNB, Madrid)
- Florencio Pazos (CNB, Madrid)
- Ana Rojas (CABD, Sevilla, *Deputy Coordinator*)

Executive Summary

Currently, genomics, focused on the study of entire genomes and the first omics discipline to appear, is applied with quite success to stratify disease subtypes (for example, several subtypes of tumors in cancer) and provide a more customized treatment for each individual. Other recent genome-wide omics-technologies providing massive analysis of DNA methylation or histone acetylation (epigenomics), mRNA (transcriptomics), proteins (proteomics), metabolites (metabolomics), etc., are still poorly applied into clinics, and importantly, for the most part, they are normally studied individually with distinct approaches generating fragmented biological information rather than integrated knowledge. This poor integration of genomic information with functional large-scale downstream biological information constitutes a main limitation to expand precision medicine to complex diseases that result from changes in gene regulation or other important modulators of cell phenotypes and outcomes, rather than from specific genetic mutations. Indeed, the relationships among them are frequently neglected in the literature. They are not independent and they are not accounted usually in the analyses. For instance, gene expression is greatly regulated by DNA methylation levels but also by the levels of proteins and metabolites, so measuring each variable independently is misleading and prone to wrong observations. This poses an example of how relevant is to incorporate data on all fronts from genomes to phenomes, and analyze how relationships are formalised, to fully redeem the promise of precision medicine. The integrative analysis of all these omics data in the clinic aims not only to find the right treatment at the right time for each patient, but also that it can be used in individual-based plans to treat high-risk populations. This global challenge requires to enroll and coordinate resources in multiple fields (biology, bioinformatics, data analysis, medicine) in order to overcome four main limiting steps: (1) data acquisition and integration; (2) big data analysis including machine learning computational techniques; (3) diagnosis and prognosis methods applied to clinical omics data; and (4) ethical aspects associated with the handling of omics data.

Introduction and general description

Genomics was the first omics discipline to appear and has currently been applied quite successfully in the clinic. Genome sequencing using broad global approaches (such as whole genome sequencing, WGS, or, whole exome sequencing, WES) or more specific targeted approaches (such as gene panel sequencing, or, chromosomal sequencing) have already been introduced as key technologies in many biomedical studies with clinical applications, entering medical practice (Biesecker et al., 2014). It is well known, for example in the case of oncology, that these technologies have been applied to identify specific gene signatures, which were then used to stratify cancer patients based on the identification of various biomolecular tumor subtypes, providing genomic-driven profiles and personalized treatments for each individual (Shen et al., 2015).

Along with genomics, we can distinguish a main cast of the other omic technologies that complement essential aspects of the activity and regulation of the genes, the gene-products and the biomolecules acting in a living system. These genome-wide omic technologies (GWOT) are the following: (i) the described technology that measures whole genomes sequences at different depths (genomics); (ii) the technologies that measure the state of DNA methylation or the histone epigenetic modifications (methylation, acetylation, etc) (epigenomics); (iii) technologies that measure genome conformation and 3D architecture (3D-genomics); (iv) technologies that identify regulatory regions and transcriptional regulation (regulomics); (v) technologies measuring the global expression of all the genes, either coding mRNAs or ncRNAs (transcriptomics); (vi) technologies directly associated with the proteome that measure the presence and activity of proteins and peptides and their post-translational modifications (proteomics); (vii) technologies that measure the metabolites at global scale in the cells (metabolomics); (viii) technologies that measure the interactions between the molecules in cells, in different organs, in an entire organism, or between different organisms like a pathogen and its host (interactomics).

Besides genomics, many of these other genome-wide techniques are still poorly applied into the clinic, and importantly, for most part, they are normally studied individually rather than in an integrated omics manner. The limited integration of multiplex data derived from different omics technologies constitutes a main limitation and a key challenge to expand personalized precision medicine and, thus, to efficiently treat complex and currently incurable diseases. Indeed, the functional associations and relational links among the data generated by different omics technologies are not very frequently endeavor in the scientific literature.

However, it is clear that the different omics technologies, centered or rooted in the genome (i.e., on the human genome in the case of biomedical studies), are not independent and should be considered as a coherent compendium of high-performance experimental methodologies that provide complementary data to study the biological systems at global scale. Therefore, although the genome is the basis of all omic studies, and genomics and epigenetics should be at the beginning of any comprehensive biomolecular approach that aims to study a biological system, the integrative analysis of multiple types of omics data is essential to develop adequate diagnosis and prognosis methods, find personalized treatments and implement individual-based plans to treat rare cases or high-risk subpopulations.

In conclusion, the translational power of the different omics technologies will only be fully exploited if integrative, multi-disciplinary and inter-disciplinary approaches are used, which requires enrolling and coordinating resources and expertise from the fields of general biology, medicine, bioinformatics, computer sciences, instrumental technological engineering, big data management and analysis, deontology and law. We will now focused on four main work areas that often constitute limiting steps in the omics field: (I) Data acquisition and integration; (II) Big Data management and analysis; (III) Diagnosis and Prognosis methods applied to Precision Medicine; (IV) Ethical aspects associated with the handling of omics data. Although these work areas will be mostly covered from a human and disease perspective, the discussed topics are certainly applicable to the use of omics technologies in the study of all other living organisms and of many biological processes (as described in the following chapters of this theme).

I. Data acquisition and integration

As genomics is a data-intensive discipline, the computational treatment of the data is as important as the experimental methods themselves. While in other scientific disciplines data handling can be seen as a helping hand to the experimental setup, in genomics (and in any omics technology) it is an intrinsic part, and these approaches would be unfeasible without it. Handling the large amounts of data generated by these approaches poses some important technological challenges. Similarly, it is increasingly common to obtain data from different omics approaches for the same sample/patient (multi-omics). While each omics provide a different point of view on the biological problem under study, only through their integration it is possible to get insight into the characteristic and intrinsic complexity of living systems.

Genomics poses some of the most severe computational challenges that we will have to face in the next decade. The data handling requirements of genomics is on par with other

data-hungry scientific disciplines such as astronomy and particle physics, and modern information and communications technologies (ITCs, e.g. Youtube, Twitter). As in these disciplines, the pace at which new data is generated in genomics is increasing exponentially, what imposes a number of computational changes (Stephens et al., 2015; Wong et al, 2019).

Five years ago, there were more than 2,500 high-throughput sequencers in 55 countries all over the world. The total amount of sequence data produced is doubling approximately every seven months (http://omicsmaps.com/). By 2025, it is estimated that 25% of the world population will have their genomes sequenced, as many countries are carrying out massive sequencing projects due to the promises of genomics for personalized medicine. Even many persons could have more than one sample sequenced (e.g. cancer, different tissues), for different time periods, as well as other omics datasets ultimately based on nucleic-acid sequencing (e.g. epigenomics, transcriptomics). So, even for a single person the amount of genomic data required for personalized medicine could be really huge (Chen et al., 2012). On the other hand, emerging sequencing technologies such as nanopore could decrease these figures a little as they require less over-sample. It is important to take into account that all these estimations are based on a scenario of extensive data acquisition by application of current technologies, i.e. current technologies will be applied in the future to more people and samples, and do not consider eventual new technologies that could change the panorama drastically and increase these figures even more. Consequently, a large data storage capacity is required for handling raw genomic information. Although algorithms for data compression can alleviate this to some extent, it was shown that compression/ decompression time can be an issue in certain scenarios (Loh et al., 2012). In certain circumstances, it is possible to disregard the original raw data or even the assembled genomes and store, for example, only a list of variants relative to a reference genome. But this is not general as, for example, cancer genomes and other complex samples may present large rearrangements that cannot be coded or stored in this way. A promising step in that direction is to use genome graphs for representing collections of genomes (Rakocevic et al., 2019). Certainly, the raw data will be increasingly disregarded in the future as methodologies for inferring higher-level data from them improve: for example, storing only expression values in transcriptomics, instead of the original reads. In any case, even with these alleviations, the efficient storage of these increasing amounts of genomic data will certainly pose a challenge in the future.

Regarding the data distribution, compared with other disciplines and technologies the distinctive feature of genomics data is that they are requested and transmitted in units

spanning a wide range of sizes: from a few bases or genes (e.g., to compare against a database of motifs or to perform a sequence search) to bulk downloads from central repositories (e.g., to perform massive analyses locally). Cloud computing can help alleviating the data transmission requirements as it allows running the analyses in the same (remote) machine where the data are (Marx et al., 2013), consequently requiring the transmission of only the code to perform the analyses and the distilled results. Another issue that has to be addressed when dealing with transmission of genomic data is privacy, as in many cases medically sensitive genomic data could have to be sent to third party machines for its analysis. In this regard, a promising approach is to use encryption methods that allow us to manipulate and perform certain queries on the encrypted data without requiring its decryption. For example, it would be possible to retrieve the mutations or variants in a particular site of an encrypted genome without having access to its whole sequence. Although computationally very expensive, these approaches could facilitate the widespread adoption of genomic medicine by alleviating those problems associated to data privacy (Erlich et al, 2014).

II. Big Data management and analysis

Big data is a long-known concept inherited from the classical "hard" sciences (physics, quantum chemistry, etc.), and refer to the amount of data that a particular experiment generates. Big data requires the rule of the four V's to comply: **V**olume of data, **V**elocity of processing the data, **V**ariability of data sources, and **V**eracity of the data quality. Since biology has become quantitative very recently, it is now when we are starting to reach a critical mass in data availability (Stephens et al., 2015). The humongous quantity of genomic data due to next generation sequencing, are in part the cause of this. In 2019 we have at least 91K species sequences, and projections for the 2025 indicate that even at the more conservative estimates of doubling every 12 months or every 18 months (equivalent to Moore's law), we should reach exabase-scale genomics well within the next decade (Stephens et al., 2015). In 2015, the Sequence Read Archive (SRA) already contained more than 3.6 petabases of raw sequence data which reflected the ~32,000 microbial genomes, ~5,000 plant and animal genomes, and ~250,000 individual human genomes that had been sequenced or were in progress thus far. As sequencing capacities have expanded considerably, if it continues at the current rate by doubling every seven months, then we should reach more than one exabase of sequence per year in the next five years and approach one zettabase of sequence per year by 2025.

As mentioned above, we should also consider that there is a large possibility that a significant fraction of the world's human population will have their genomes sequenced, so estimations lie between 100 million and as many as 2 billion human sequenced genomes by 2025 (four to five orders of magnitude growth in ten years) (Stephens et al., 2015), necessitating generating new sequencing data multiple times per person to monitor molecular activity. Indeed, this number could grow even larger, especially since new single-cell genome sequencing technologies are starting to reveal previously unimagined levels of variation (Massoni-Badosa et al., 2020; Mereu et al., 2020). Furthermore, the technology used to sequence DNA from the genomes has been creatively extended and deployed for other omic applications, some of them associated with many different types of high-throughput sequencing (such as, transcriptomics, epigenomics or regulomics), other applications based on simultaneous sequencing of multiple genomes (such as, microbiomics and metagenomics), or others based on different types of high-throughput technologies (such as, proteomics, metabolomics, etc).

All of these high-throughput technological applications require precise accurate quantification of massive signal (e.g., billions of sequencing reads) to capture diversity of signal and diversity of abundances, thus requiring millions of data points to accurately estimate underlying distributions as they change over time. These procedures are computationally expensive. For instance, variant calling on 2 billion genomes per year, with 100,000 CPUs in parallel, would require methods that process 2 genomes per CPU-hour. As another illustrative example, whole genome alignment used for a variety of goals (from phylogeny reconstruction to genome annotation via comparative methodologies), is a costly task. Just a single whole genome alignment between human and mouse consumes ~100 CPU hours. So, aligning all pairs of the ~2.5 million species expected to be available by 2025 amounts to 50–100 trillion such whole genome alignments, which would need to be six orders of magnitude faster than possible today. In addition, for medicinal and clinical applications, just having the genome will not be sufficient: for each individual, it will need to be coupled with other relevant 'omics data sets, some collected periodically and from different tissues, to compare healthy and diseased states, in this regard, integration issues are very important.

Regarding the data handling, currently there are plenty of large cloud-based genomic resources using cloud computing paradigms, especially to support the requirements of the largest sequencing centres or to support the needs of large communities and international projects (e.g., the Cancer Genome Atlas, TCGA, the International Cancer Genome Consortium, ICGC, the Sequence Read Archive, SRA, among others). In order to make these

systems most useful, a further development of robust application programming interfaces (APIs) for discovering and querying large datasets on remote systems is on demand. In this line, for instance the Global Alliance for Genomics and Health (https://www.ga4gh.org/) adopted such standards for human genomic data, and it is expected that similar communities will follow.

Finally, very fundamental issues arise when sharing and distributing data such as authentication, encryption, and other security safeguards must be developed to ensure that genomic data remain private. A number of data science technologies, including R, Mahout, and other machine learning systems powered by Hadoop and other highly scalable systems are used in regular basis. Data science companies, Google, as well as open-source initiatives are already developing such components, with a large degree of success. However, genomics poses unique challenges in terms of data acquisition, distribution, storage, and especially analysis, waiting for innovations from outside the field is unlikely to be sufficient (see challenges).

III. Diagnosis and Prognosis methods applied to Precision Medicine.

Precision Medicine permits healthcare interventions to be tailored to groups of patients based on their disease susceptibility, diagnostic or prognostic, as well as their treatment response. In fact, an ideal system links analytics, clinical practice, research and data science with the goal of improving the efficiency and effectiveness of disease prevention, diagnosis, and treatment (Ginsburg, 2014; David et al., 2015; Ginsburg & Phillips, 2018).

The development of diagnosis and prognosis methodologies requires the understanding of processing and coding of genetic, epigenetic and post-genetic information, which cannot be deduced from the sequence of genome alone. In this sense, for example, proteins as molecular machines are more directly related with the phenotypes of the different pathological states, and therefore are closer to the underlying disease-causing pathways. Omics-based diagnosis and prognosis methodologies are expected to improve in specificity and possibility, with single tests for the decision of treatment pathway, therapy choice or even disease risk for multiple diseases simultaneously (Slade et al., 2015; Matthews et al., 2016).

IV. Ethical aspects associated with the handling of omics data

Our world is moving towards precision personalized medicine (PPM) where patients will eventually be treated not only according their symptoms and disease associated but also taking into account their genetic variants. Subtle genomic variations might have enormous

implications regarding the convenience, or not, to prescribe a drug for a person. Of course, this has downstream implications. Patients will have to be taught, be duly informed and be advised to accept to give regulated access to their genomes in order to explore potential beneficial or detrimental variations affecting their health and treatments, hence the precision personalized medicine that has been aimed and expected since long. Measures to enforce and ensure the privacy of all these patient genomes will have to be developed, including the use of blockchain methods (Ozercan et al. 2018; Kuo et al. 2020).

Impact in basic science panorama and potential applications

I. Data acquisition and integration

It is clear that the main impact with respect to the acquisition of omics data is associated with the challenge of achieving adequate coverage of all omics technologies (that is, conducting true multi-omics studies) and achieving adequate integration of the different complementary omics technologies (within the 8 levels described above). Multi-omics studies, where different omics datasets are retrieved for the same sample or patient, are increasingly popular (Noor et al., 2019). While these omics datasets were first analysed separately and the results combined a posteriori, it became clear that only integrating and analysing them in a synergistic fashion made it possible to obtain a clear biological picture.

Machine learning approaches are frequently used for analysing multi-omics datasets due to their intrinsic versatility to handle and combine diverse data. In this regard, multilevel learning (Serra et al., 2016) is especially suited for handling data of diverse nature and even coded indifferent ways. When mechanistical information is available besides the data themselves, networks are the mathematical object of choice to represent it. For example, a network representation of a known biological pathway with the multi-omic data put on top of the nodes (genes/proteins) (Barabási & Oltvai, 2004). Then, network based approaches are used to mine these networks in the search for new knowledge. In fact, interactomics (which measure the molecular interactions between proteins, between proteins and DNA or RNA, or between any type of biomolecule) can only be adequately studied using networks and applying graph theory (De Las Rivas & Fontanillo, 2010; De Las Rivas & Fontanillo, 2012). This has spawned the new field of network biology (Barabási & Oltvai, 2004).

Not only the integration of data coming from different omics (i.e., inter-omics integration) is important, but also the combination of different instances or different datasets of the same type of genomic data (i.e., intra-omics integration). For example, different human genomes produced with different sequencing technologies, or different expression

transcriptomic datasets produced in different labs with samples from different populations,. In this regard, technical issues concerning standards for representing the data, consistent database indentifiers for proteins and genes, ontologies for representing knowledge in a standardized way, etc. should be considered.

II. Big Data management and analysis

The analysis of genomic data involves a wide diverse range of approaches because of the variety of steps involved in reading a genome sequence and deriving useful information from it. Our ultimate goal is to be able to interpret certain aspects of genomic sequences and answer different questions related to them. A set of questions are, for instance, how DNA mutations, expression changes, or other molecular measurements relate to development, behaviour, evolution, or to disease from a biomedical perspective.

Accomplishing this goal within the Big Data framework, will clearly require a multidisciplinary approach with the integration of several experts of the biological and biomedical domain that formulate the key biomedical questions, together with computer scientists and engineers capable of managing and applying large-scale machine learning systems in robust computing infrastructures, that can support flexible and dynamic queries to search for patterns in very large collections with very high dimensionality. Additionally, mathematicians and statisticians will provide a third key axis of this multidisciplinary approach, which will be essential to translate data into appropriate numbers and quantitative expressions to achieve true Big Data science in the next generation.

On the other hand, recent re-implementations of machine learning and data analytics on genomics data has generated the widespread impression that such methods are capable of solving most problems without the need for conventional scientific methods of inquiry, however it is essential to understand how biology works to make sense out of data and to properly use these computing technologies. In fact, while machine learning has been there even before bioinformatics existed, it became integrated into the field via protein sequences analyses almost 30 years ago (Rost & Sander, 1993) to predict protein secondary structures. Thus, while these and akin methods have been used widely in the field, it is now within the genomics and transcriptomics subdomains when a full explosion of machine learning methods have emerged, which are agnostic to theory.

Big Data analytics is gaining weight in the clinical domain, for example in the prediction of individual risk factors for different types of diseases, for clinical decision support, and for practicing precision medicine using genomic information (Alonso-Betanzos

& Bolon-Canedo, 2018). In particular, oncology artificial intelligence (AI) has contributed to the resolution of certain specific biomedical problems in cancer studies, over the past decade. Deep learning (DL), a subfield of AI that is highly flexible and supports automatic feature extraction, and is increasingly being applied in various areas of both basic and clinical cancer research (Shimizu and Nakayama, 2020). Overall, big data studies and associated new technologies will continue to guide novel, exciting research that will ultimately improve healthcare and medicine, but we are also realistic that concerns remain about privacy, equity, security, and benefit to all (Car et al., 2019).

Another important field where big data could gain weight is pharmacogenomics, which has surpassed pharmacogenetics, leading to new perspectives in drug design and identification of drug response and drug resistance factors, once caveats dealing with data collection, processing, analysis and interpretation are overcome (Barrot et al., 2019). Finally, proteogenomics (Nesvizhskii, 2014), as an integrative approach at the interface of genomics and proteomics, is a field that could benefit enormously precision personalized medicine (PPM) (Nishimura & Nakamura, 2016), making this data integration a source of novel knowledge to improve the genomic understanding of disease.

III. Diagnosis and Prognosis methods applied to Precision Medicine

In the last years, the use of mass spectrometry in the diagnosis and prognosis methods has become an indispensable tool, especially in the search and detection of epigenetic biomarkers. Genomics has permitted to advance in the genomic mapping of nucleosome positions, the knowledge of DNA and histone modifications as well as chromatin-bound factors (Hawkins et al., 2010; Zhou et al., 2011). However, the study of the mechanisms of how epigenetics can influence or modify the biological functions is still needed, and therefore the use of mass spectrometry and proteomic or peptidomic approaches have an important impact on epigenetics research, allowing the development of methods for the identification of disease-specific proteins and endogenous peptides. In addition to participate in the development of diagnosis and prognosis methods for a better treatment response, the use of protemics and peptidomics provide a better understanding of disease and a more effective use of modern therapies (DiMeo et al, 2016). As an example, mass spectrometry approaches have contributed to the development of cancer epigenetics through the identification of biomarkers for diagnosis and prognosis by improving the understanding of chromatin regulation mechanisms. As an advanced product of the efficient combination of genome-centered omics plus proteome-centered omics, a new area called proteogenomics has emerged in recent years.

One of the first relevant proteogenomic studies in biomedical area was published by Mertins and collaborators working on breast cancer (Mertins et al., 2016). These authors demonstrated that proteogenomic analysis of breast cancer elucidates the functional consequences of somatic mutations, narrows candidate nominations for driver genes within large genomic deletions and amplified regions, and identifies therapeutic targets.

IV. Ethical aspects associated with the handling of omics data

Handling and using patient's genomic data will have to be done according to the law. In Europe, the recent General Data Protection Regulation (GDPR, 2016/679) imposes strict and severe limits to the free movement and sharing of these genomic data. At the same time, GDPR grants the patients and their associations full rights to handle their own data, including genomic data, which can then be offered to researchers and clinicians for further analysis, This new possibilities pose a number of challenges to researchers and clinicians, being confronted with the fact to protect the privacy of all genomic information and, at the same time, respond to patient's request to access to their genomes (or to a collection of genomes from patients of a given association representing them legally) (Schickhardt et al. 2020). These problems require urgent solutions, otherwise many sets of useful genomic data kept in different institutions and different countries will remain blocked and not be permitted for sharing, and thus everyone will be losing. Some innovative strategies have been presented, including the European Health Research and Innovation Cloud (HRIC; Aarestrup et al. 2020). As presented by these authors, the HRIC will enable data sharing and analysis for health research across the EU, in compliance with our rigorous data protection legislation while preserving the full trust of the patients and volunteers participating. Similarly, as we will begin accessing people's genomes we will need a clear strategy to deal with the incidental findings, about how and when to communicate them to individuals (and their physicians), particularly those affecting or compromising the current or future health conditions of subjects under study (Ayuso et al. 2015).

Key challenging points

I. Quantitative, efficient and secure acquisition of omics data

Omics technologies are becoming essential tools in most biological and biomedical fields. In order to satisfy these growing needs, there are some issues and limitations that will need to be overcome in the coming years, including:

- Reduce the sequencing costs at least two orders of magnitude.
- Collect other data and metadata associated to the sequences.

- Develop new genome-focused systems for representing sequencing data beyond string characters.
- Improve technologies for data storage (memory, disk, …) and transmission (networks).
- Improve compression and indexing systems for efficient data storage and access.
- Develop authentication, encryption, and other security safeguards to ensure data privacy.
- Improve quantitative omics (i.e., measuring quantitative changes) and longitudinal omics (i.e., measuring changes over time events).
- Improve the data acquisition and integration of clinical, phenotypic information with omics biomolecular data
- Develop technical and analytical approaches for combining intra- and inter-omics datasets.

II. Interpretation: a recurrent challenge in Bioinformatics.

Since biology started to become quantitative, many expectations were put on the computational side and in the new field of bioinformatics (Ouzounis, 2012), with the naive idea that computers will solve all our problems. Most initial promises relied on the final automatic interpretation of data (Thornthon, 1998), which rapidly moved from an expectation to a key challenge, at current time. The truth is that data interpretation has become a recurring challenge, especially when data integration was declared the first bottleneck. This initial challenge was proclaimed by expert teams in genome sequencing and genomic databases: "managing huge data volumes, integrating information from various discovery platforms and discerning phenotypic implications" (Scherer et al., 2007). By contrast, in some particular areas of genomics, it has been long proposed that the bottleneck in our knowledge is the lack of available data. However, this is true only in some situations, because for instance the availability of thousands of general population genomes, have helped to revisit the penetrance of certain diseases (Check Hayden, 2016) based on frequency of mutations.

Then comes the hype. As an illustrative case, the report of the "genetic heroes" or those "bullet-proof genomes", those 13 people who should be dead, as mendelian mutations were found in their genomes (Chen et al., 2016). This latter was widely refuted, as methodology and interpretation were faulty and over-hyped (MacArthur, 2016). This controversial example shows that accurate, reproducible data analysis remains as the great challenge of large-scale omics studies. In this regard, in many cases a frequently recommended increase of sample size will not aid to solve the problems associated with data analysis and data interpretation, where functional and biological assignment remains a critical step. Therefore, this is likely one of the most challenging aspects in current biomedical research to achieve true precision personalized medicine. For instance, in medical genomics,

the identification of genomic variants in an individual genome and correct interpretation of what the impact of a particular mutation is, remains highly problematic. In other words, discovering whether a particular mutation is actually causative of a given pathogenic state or not remains an unresolved question. Therefore, while the obvious cases linked to well-defined diseases are easily addressed, the large number of mutations of unknown significance are still badly resolved, and the choice of our predictive computational tools and input data greatly affects the results (McCarthy et al., 2014). In fact, it has been estimated that potential phenotypic consequences for variation in >75% of the ~20,000 annotated genes in the human genome are lacking (Posey et al., 2019). This poses a paradox situation since the more data we have, the more difficult is its interpretation. In this regard dozens of bioinformatics and computational biology methods have been developed to answer this precise question, which indicates the relevance of the matter. Indeed, benchmarking of procedures and annotators have been applied in clinical settings, where poor results and poor consensus required final discussions to reach a consensus guidelines (Amendola et al., 2016). This example describes a common problem in computational genomics, where the development or application of yet another method, does not improve the interpretability of the results. This is due in part to the fact of always using a human reference genome, so another one of the key challenges in medical genomics is moving towards reference-free genomic assemblies and analysis, or even to the simultaneous use of multiple references applying fuzzy and deep learning methodologies.

To summarise, biological big data can only be efficiently managed and analyzed by expert computational biologists in close contact with physicians and biological researchers, who only together can achieve a transition from association studies to causality studies.

III. Computation (CPU) capabilities

Improvements to CPU capabilities, could help, but trends in computing power are often geared towards floating point operations and do not necessarily provide improvements in genome analysis, in which string operations and memory management often pose the most significant challenges. Moreover, considering technological challenge, the main bottleneck of Big Data analysis in the future may not be in CPU capabilities, but in the input/output (I/O) hardware that shuttles data between storage and processors a problem requiring research into new parallel I/O hardware and algorithms that can effectively utilize them.

IV. Proper training in biology, computing, and statistics

Another challenge that became apparent trough the development of the field was the "automation utopia" versus the "people paradox": the realisation that the application of computer sciences and machine learning to biology results in an increase in the demand of well-trained people (Miller & Attwood, 2003). This realization has become even more evident with the skills required to analyse the data we have on hands. In the particular case of genomics, it poses unique challenges in terms of data acquisition, distribution, storage and analyses. We must face these challenges ourselves, starting with integrating data science into graduate, undergraduate, and high-school curricula to train the next generations of quantitative biologists, bioinformaticians, and computer scientists and bioengineers. This particular challenge will be very well illustrated in the following chapters of this theme, since in all the highlighted areas of (epi)genomic research there is a serious need of not only computational infrastructure but also of computational biologists.

Moreover, the lack of statistical thinking is a norm in the biomedical and biological field, and it becomes evident when it comes to experimental design and choices of methodological pipelines. For example, p-values are being widely used but often not well understood, therefore widely abused in many biomedical research publications (Leek & Peng, 2015). Furthermore, in many omics studies very little scrutiny and debate is usually presented for the experimental design needed to achieve a given statistical significance.

V. We need theory when applying AI to biological data

One of the main challenges in biological and biomedical research is the heterogeneity of the data, as any biological system or organism is composed of tens of thousands of components. And we need to know the relationships among these components. Moreover, there is a widespread conception that Artificial Intelligence (AI) will solve most problems without any scientific logic behind, just like trusting in hidden patterns will give us the outcomes we want. This misconception is accelerated by the ease of obtaining data and the ease to implement many methods as black boxes. No matters the depth and the sophistication of data-driven methods, such as deep learning neural networks (DLNN), in the end they merely are used to provide a simple output fitting curves to existing data. However, these methods require far larger quantities of data than anticipated by big data aficionados in order to produce statistically reliable results. Even more, in many cases theory regarding the methods, and

knowledge regarding the processes we want to get answers are needed and cannot be displaced yet (Coveney et al, 2016).

VI. Reproducible biomarkers for clinical use

In addition to the difficulties of integrating data from the different omics, complications also arise in the discovery and application of stable reproducible biomarkers due to difficulties in handling heterogeneous samples, adequate recognition of statistical errors in identifications (i.e., calculation of false positive and false negative rates), possible cross-reactivity or hidden factors in the techniques, incomplete molecular identification, or final incorrect data analysis and interpretation. Thus, the standardization of the omic methodologies through the use of last generation instruments to avoid variability of procedures is a challenging point to overcome in order to successfully implement the use of newly discovered biomarkers in clinical applications.

VII. Ethical guidelines in the omics and precision medicine era

The continuous improvement of omics technology in terms of the quality and throughput of the generated data is revolutionizing all areas of biomedical research. However, these rapid advances have also raised ethical questions that demand answers, guidelines are regulation: (i) What do we do with the non-actionable information of a potential disease identified in a screening? Should we notify the patients? (ii) How will be this information anonymised? Nowadays, is very easy to get to a name a zip code using genomic data. This will have implications regarding ancestry, disease, etc. How are we as a society going to deal with this? (iii) Precision medicine must respect ethical and moral concerns of all groups and cultures and ensure safety of information in an environment where hospital computer systems could be in the focus of cyber-attacks. (iv) Researchers and institutions, that can be either public or private, must protect the integrity of their data and their research and ensure that findings are reproducible. The use of independent verification and validation bodies should be encouraged. Policies to either release the data, or to regulate this should be implemented.

Together with omics methods, recent advances in genome editing technologies, will make personalized precision medicine a reality in the near future. For example, it will be soon possible to alter the genome of any living organism, including human beings, thanks to the popular, affordable and efficient genomic tools, such as CRISPR-Cas, which have been already presented in the previous chapter (see chapter 3.1). Whether the aimed strategy will target patient's cells ex vivo or in vivo we will have to assess the ethical aspects and potential

undesired collateral damage we might infringe on the people's genome. Off-target effects and mosaicism are usual unwanted problems associated to any current genome editing approach. The consequences are different when you are dealing with animal or plant species, or human beings. Clearing undesired mutations through breeding schemes, selecting the most appropriate individuals is something acceptable for animals and plant species, whereas similar strategies are ethically unacceptable for patients. Therefore, reflecting on the future responsible use of genome editing technologies is a must and will have to be addressed.

The ethical aspects of our current ability to manipulate genomes expand beyond the clinical examples. In plants, the possibility to implement simple mutations associated with traits of agroeconomic interest faces ethical and legal concerns in many countries, including the Europe Union. The sentence of the European Court of Justice, from July 2018, stating that genome edited crops will not be exempt to comply with the EU GMO Directive (2001/18) clearly represented a break to the development of new genome-edited plants, heavily penalized by that legal decision. In contrast, in other parts of the world, such concerns are not found and, hence, their possibilities to trade and expand their markets will negatively influence our immediate future in Europe (Hundleby and Harwood, 2019).

CSIC advantage position and multi- / inter-disciplinarity

The CSIC is well positioned at the international level for acquiring and managing genomics data. The Institution has a number of in-house sequencing services and other omics services, as well as agreements with other organizations to use shared facilities. More importantly, it has the required human expertise for carrying out the data analyses. In this respect, in the last years, many research centers of the CSIC in the life-sciences areas have set up some sort of in-house Bioinformatics Service and related facilities that provide certain analysis service for internal and external researches. However, there are not many research groups leaded by staff scientists working actively in the field of Bioinformatics, Computational Biology or Computational Genomics. In fact, we did a brief exploration and found that, considering the entire CSIC, there are not more than 12-15 research groups working on these fields.

On the other hand, and despite the limitations stated above, there are also a number of examples that illustrate how collaborations among CSIC research are making influential contributions to the omics and precision medicine fields:

- The ability to generate, store and compare human genomes from different origin requires to producing reference data sets to be employed for assembling and comparing them with patients genomic data. Within the CSIC there are experts in detecting and analysing human genome data from

different populations. Ethically, it will be important to ensure access to these alternative genome data sets, different from the standard genomes derived from white anglosaxon groups currently used world-wide. Recently, a CSIC team discovered the unexpected genomic complexity and differences with the accepted reference genome in several populations from Africa (Lorente-Galdos et al. 2019; Gelabert et al. 2019). In Spain, a recent study of 267 genomes concluded that there are significant differences in disease-related genetic variability which identities Spanish genomes, as opposed to other genome SNPs that appear to be characteristic in other populations (Dopazo et al. 2016).

- Bioinformaticians within the CSIC have been developing new tools anticipating the need for better algorithms and improved web-based platforms for devising CRISPR-Cas tools, such as the BREAKING-CAS tool (Oliveros et al. 2016) and the comparative analysis of the existing programs on the subject (Torres-Pérez et al. 2019). Similarly, there have been efforts to develop new tools to predict potential epigenetic landmarks in genomes, as this can impact on our ability to correctly interpret individual genomes and to provide a more comprehensive information of sites within the genome where alterations are likely to be pathogenic (Pazos et al. 2018).

Finally, regarding the ethical aspects of omics research and precision medicine, there are already initiatives, undertaken by CSIC researchers, promoting responsible research and innovation (RRI) in the genome editing universe, at the international level, such as ARRIGE (Association for Responsible Research and Innovation in Genome Editing) (Montoliu et al. 2018; Hirsch et al. 2019). Similarly, at the CSIC, there are plant molecular biologists that have reflected on the legal and ethical aspects of the modification of plant genomes, analyzing their pros and cons, within the complex European scenario (Casacuberta and Puigdomenech, 2018a; Casacuberta and Puigdomenech, 2018b)

Plan and Resources

A major concern regarding the position of CSIC in the omics field is whether it has the capacity to adapt quickly to this rapidly changing area. Due to its large size and bureaucratic structure, the CSIC typically evolves slowly and with delayed responses to these rapid changes. For example, it took years of hard-work to set up a Genomics Facility for performing cDNA expression experiments and analysis and, when it was up and running, the technology changed to RNA-seq. With the current structure of the CSIC, we believe that the best solution to alleviate this is to have in-house trained versatile staff able to respond and adapt to these changes coping with a limited and slowly-evolving infrastructure. In this sense, it is especially important the role of Bioinformaticians, and the CSIC should consider including more of these profiles in future calls for job positions.

As stated in previous sections, the use and integration of the different and multiple omics technologies is a key step towards the achievement of modern precision personalized medicine (PPM). Probably at present time the two main problems in the CSIC to achieve this challenge are: (i) the frequent fragmentation of the expertise and experience of the CSIC researchers who often work in a specific area of the omics technologies; (ii) the frequent detachment of the CSIC research centers form the hospitals, the health centers and the medical units. Early in this 21st century, the main leading countries of the world and Europe are investing heavily growing capital in basic and applied research in Medical Genomics, Genome Medicine and personalized Precision Medicine. Thus, in order to be internationally competitive in the Precision Medicine field, we encourage the CSIC to heavily invest in computation, information and data technologies and infrastructure, since the development of advanced modern societies is mainly conducted by data, information and knowledge, which it has become the principal wealth of nations worldwide (Iriart, 2019).

References

Aarestrup FM, Albeyatti A, Armitage WJ, *et al.* Towards a European health research and innovation cloud (HRIC). *Genome Med*. **2020**, 12(1): 18. doi: 10.1186/s13073-020-0713-z

Alonso-Betanzos A, Bolón-Canedo V. Big-Data Analysis, Cluster Analysis, and Machine-Learning Approaches. In: Kerkhof P., Miller V. (eds) Sex-Specific Analysis of Cardiovascular Function. *Adv Exp Med Biol*. **2018**, 1065: 607-626. Springer, Cham. doi: 10.1007/978-3-319-77932-4_37

Amendola LM, Jarvik GP, Leo MC, *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines Among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium *Am J Hum Genet.* **2016** 98(6): 1067-1076. doi: 10.1016/j.ajhg.2016.03.024.

Ayuso C, Millan JM, Dal-Re R. Management and return of incidental genomic findings in clinical trials. *Pharmacogenomics J*. **2015**, 15(1): 1-5. doi: 10.1038/tpj.2014.62

Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. **2004**, 5(2): 101-113. doi: 10.1038/nrg1272

Barrot CC, Woillard JB, Picard N. Big data in pharmacogenomics: current applications, perspectives and pitfalls.*Pharmacogenomics* **2019**, 20(8): 609-620. doi: 10.2217/pgs-2018-0184.

Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med*. **2014**, 370(25): 2418-2425. doi:10.1056/NEJMra1312543

Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. *BMC Med.* **2019**, 17(1): 143. doi:10.1186/s12916-019-1382-x

Casacuberta JM, Puigdomènech P. Proportionate and scientifically sound risk assessment of gene-edited plants. *EMBO Rep*. **2018a**, 19(10): e46907. doi: 10.15252/embr.201846907

Casacuberta JM, Puigdomènech P. European politicians must put greater trust in plant scientists. *Nature*. **2018b**, 561(7721):33. doi: 10.1038/d41586-018-06129-2

Chakravorty S, Hegde M. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. *Annu Rev Genomics Hum Genet.* **2017**, 18: 229-256. doi:10.1146/annurev-genom-083115-022545

Chen R, Mias GI, Li-Pook-Than J, *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**,148(6): 1293-1307. doi:10.1016/j.cell.2012.02.009

Chen R, Shi L, Hakenberg J, *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol*. **2016**, 34(5): 531-538. doi: 10.1038/nbt.3514

Coveney PV, Dougherty ER, Highfield RR. Big data need big theory too. *Philos Trans A Math Phys Eng Sci*. **2016**, 374(2080): 20160153. doi:10.1098/rsta.2016.0153

Check Hayden E. A radical revision of human genetics. *Nature* **2016**, 538(7624): 154-157. doi: 10.1038/538154a.
David SP, Johnson SG, Berger AC, *et al.* Making Personalized Health Care Even More Personalized: Insights From Activities of the IOM Genomics Roundtable. *Ann Fam Med*. **2015**, 13(4): 373-380. doi: 10.1370/afm.1772
De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*. **2010**, 6(6) :e1000807. doi: 10.1371/journal.pcbi.1000807
De Las Rivas J, Fontanillo C. Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief Funct Genomics*. **2012**, 11(6):489-496. doi: 10.1093/bfgp/els036
Di Meo A, Pasic MD, Yousef GM. Proteomics and peptidomics: moving toward precision medicine in urological malignancies. *Oncotarget*. **2016**, 7(32): 52460-52474. doi: 10.18632/oncotarget.8931
Dopazo J, Amadoz A, Bleda M, *et al.* 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Mol Biol Evol*. **2016**, 33(5): 1205-1218. doi: 10.1093/molbev/msw005
Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet*. **2014**, 15(6): 409-421. doi: 10.1038/nrg3723
Gelabert P, Ferrando-Bernal M, de-Dios T, *et al.* Genome-wide data from the Bubi of Bioko Island clarifies the Atlantic fringe of the Bantu dispersal. *BMC Genomics*. **2019**, 20(1): 179. doi: 10.1186/s12864-019-5529-0
Ginsburg G. Medical genomics: Gather and use genetic data in health care. *Nature*. **2014**, 508(7497): 451-453. doi: 10.1038/508451a
Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. *Health Aff (Millwood)*. **2018**, 37(5): 694-701. doi: 10.1377/hlthaff.2017.1624
Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet*. **2010**, 11(7): 476-486. doi: 10.1038/nrg2795
Hirsch F, Lemaitre C, Chneiweiss H, Montoliu L. Genome Editing: promoting responsible research. *Pharmaceut Med*. **2019**, 33(3): 187-191. doi: 10.1007/s40290-019-00276-1
Hundleby PAC, Harwood WA. Impacts of the EU GMO regulatory framework for plant genome editing. *Food Energy Secur*. **2019**; 8(2): e00161. doi: 10.1002/fes3.161
Iriart JAB. Precision medicine/personalized medicine: a critical analysis of movements in the transformation of biomedicine in the early 21st century. *Cad Saude Publica*. **2019**, 35(3): e00153118. doi: 10.1590/0102-311X00153118
Kuo TT, Kim J, Gabriel RA. Privacy-preserving model learning on a blockchain network-of-networks. *J Am Med Inform Assoc*. **2020**, 27(3): 343-354. doi: 10.1093/jamia/ocz214
Leek J, Peng R. Statistics: p-values are just the tip of the iceberg. *Nature* **2015**, 520(7549): 612. doi: 10.1038/520612a.
Loh PR, Baym M, Berger B. Compressive genomics. *Nat Biotechnol*. **2012**, 30(7): 627-630. doi: 10.1038/nbt.2241
Lorente-Galdos B, Lao O, Serra-Vidal G, *et al.* Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biol*. **2019**, 20(1): 77. doi: 10.1186/s13059-019-1684-5
MacArthur D. Superheroes of disease resistance. *Nat Biotechnol.* **2016**, 34: 512–513. doi: 10.1038/nbt.3555
Marx V. Drilling into big cancer-genome data. *Nat Methods*. **2013**, 10(4): 293-297. doi: 10.1038/nmeth.2410
Massoni-Badosa R, Iacono G, Moutinho C, *et al.* Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol*. **2020**, 21(1): 112. doi: 10.1186/s13059-020-02032-0
Matthews H, Hanison J, Nirmalan N. "Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes*. **2016**, 4(3): 28. doi: 10.3390/proteomes4030028
McCarthy DJ, Humburg P, Kanapin A, *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **2014**, 6: 26 doi: 10.1186/gm543
Mereu E, Lafzi A, Moutinho C, *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol*. **2020**, doi: 10.1038/s41587-020-0469-4
Mertins P, Mani DR, Ruggles KV, *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **2016**, 534(7605):55-62. doi:10.1038/nature18003
Miller CJ, Attwood TK. Bioinformatics goes back to the future. *Nat Rev Mol Cell Biol*. **2003**, 4(2): 157-162. doi:10.1038/nrm1013
Montoliu L, Merchant J, Hirsch F, *et al.* ARRIGE Arrives: toward the responsible use of genome editing. *CRISPR J*. **2018**, 1(2):128-129. doi:10.1089/crispr.2018.29012.mon
Nesvizhskii A. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014**, 11: 1114–1125. doi: 10.1038/nmeth.3144
Nishimura T, Nakamura H. Developments for Personalized Medicine of Lung Cancer Subtypes: Mass Spectrometry-Based Clinical Proteogenomic Analysis of Oncogenic Mutations. *Adv Exp Med Biol*. **2016**, 926: 115-137. doi: 10.1007/978-3-319-42316-6_8

Noor E, Cherkaoui S and Sauer U. Biological insights through omics data integration. *Curr Opin Syst Biol.* **2019**, 15: 39-47. doi: 10.1016/j.coisb.2019.03.007
Oliveros JC, Franch M, Tabas-Madrid D, *et al.* Breaking-Cas-interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes. *Nucleic Acids Res*. **2016**, 44(W1): W267-W271. doi: 10.1093/nar/gkw407
Ouzounis, C. Rise and demise of bioinformatics? Promise and Progress. *Plos Comp. Biol*. **2012**, 8(4): e1002487. doi: 10.1371/journal.pcbi.1002487
Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies in genomics. *Genome Res*. **2018**, 28(9): 1255-1263. doi: 10.1101/gr.207464.116
Rakocevic G, Semenyuk V, Lee WP, *et al.* Fast and accurate genomic analyses using genome graphs. *Nat Genet*. **2019**, 51(2):354-362. doi:10.1038/s41588-018-0316-4
Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A.* **1993**, 90(16): 7558-7562. doi:10.1073/pnas.90.16.7558
Pazos F, Garcia-Moreno A, Oliveros JC. Automatic detection of genomic regions with informative epigenetic patterns. *BMC Genomics*. **2018**, 19(1): 847. doi: 10.1186/s12864-018-5286-5
Posey JE, O'Donnell-Luria AH, Chong JX, *et al.* Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet Med*. **2019**, 21(4): 798-812. doi: 10.1038/s41436-018-0408-7
Scherer SW, Lee C, Birney E, *et al.* Challenges and standards in integrating surveys of structural variation. *Nat Genet.* **2007**, 39: S7-15. doi: 10.1038/ng2093
Schickhardt C, Fleischer H, Winkler EC. Do patients and research subjects have a right to receive their genomic raw data? An ethical and legal analysis. *BMC Med Ethics*. **2020**, 21(1): 7. doi: 10.1186/s12910-020-0446-y
Serra A, Fratello M, Greco D, Tagliaferri R. Data integration in genomics and systems biology. 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, **2016**, pp.1272-1279, doi: 10.1109/CEC.2016.7743934.
Shen T, Pajaro-Van de Stadt SH, Yeat NC, Lin JC. Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Front Genet*. **2015**, 6: 215. doi: 10.3389/fgene.2015.00215
Shimizu H, Nakayama KI. Artificial inteligence in oncology. *Cancer Sci.* **2020**, 111(5): 1452-1460. doi: 10.1111/cas.14377.
Slade I, Riddell D, Turnbull C, Hanson H, Rahman N; MCG programme. Development of cancer genetic services in the UK: A national consultation. *Genome Med*. **2015**, 7(1): 18. doi: 10.1186/s13073-015-0128-4
Stephens ZD, Lee SY, Faghri F, *et al.* Big Data: Astronomical or Genomical?. *PLoS Biol*. **2015**, 13(7): e1002195. doi: 10.1371/journal.pbio.1002195
Thornton JM. The future of bioinformatics. *Trends Guide to Bioinformatics* **1998**, 30-31.
Torres-Perez R, Garcia-Martin JA, Montoliu L, Oliveros JC, Pazos F. WeReview: CRISPR Tools-Live Repository of Computational Tools for Assisting CRISPR/Cas Experiments. *Bioengineering* **2019**, 6(3):63. doi: 10.3390/bioengineering6030063
Wong KC. Big data challenges in genome informatics. *Biophys Rev*. **2019**, 11(1): 51-54. doi: 10.1007/s12551-018-0493-5
Yanai I, Chmielnicki E. Computational biologists: moving to the driver's seat. *Genome Biol.* **2017**, 18(1): 223. doi: 10.1186/s13059-017-1357-1
Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet*. **2011**; 12(1): 7-18. doi: 10.1038/nrg2905

3.3 3D GENOME ARCHITECTURE

ABSTRACT

The tridimensional (3D) organization of the genome is just beginning to be understood and, despite the tremendous progress witnessed in the recent years, many essential questions remain unanswered. Major methodological and conceptual challenges need to be overcome to reach a comprehensive view of the physiological impact of 3D genome architecture in health and disease. CSIC is in a privileged position to undertake these challenges.

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS (in alphabetical order)

- Ferran Azorín (IBMB, Barcelona, *Coordinator*)
- Albert Jordán (IBMB, Barcelona)
- Ignacio Maeso (CABD, Sevilla)
- Miguel Manzanares (CBMSO, Madrid)
- Álvaro Rada (IBBTEC, Santander)
- Joaquim Roca (IBMB, Barcelona)
- Josep Rotllant (IIM, Vigo, *Deputy Coordinator*)
- Guillermo Vicent (IBMB, Barcelona)

Executive Summary

Our knowledge about the regulation of genomic functions has benefited enormously from the unravelling of the structural organization of the genome at different levels. The initial elucidation of the structure of DNA revolutionized Biology and Medicine in many different ways. Likewise, the identification of the nucleosome as the basic structural subunit of chromatin set the basis for the understanding of the epigenetic mechanisms that are central to the regulation of genome functions, from RNA transcription and DNA replication, recombination and repair, to chromosome segregation and genome stability. In more recent years, the hierarchical organization of chromatin structure inside the cell nucleus has been well established. Nucleosome clutches, chromatin loops, topologically associating domains (TADs), compartments and chromosome territories appear as levels of organization that, ultimately, conform the tridimensional (3D) organization of the genome in the nucleus. Deciphering 3D architecture and conformational dynamics of the genome at nanoscale resolution will have an important impact on our understanding of how developmental and differentiation programs are established and maintained in the cell, the adaptive response of biological systems to environmental changes, and the etiology of genome dysfunctions, in particular, those leading to disease. However, many essential questions remain unanswered. Current experimental approaches allow us to address only few aspects of 3D chromatin organization and, for the most part, are only applicable to abundant cell types, leading to a still incomplete, static and potentially distorted view of the 3D structure of the genome. A theoretical framework to understand the forces and principles governing chromatin folding is also missing. On the other hand, characterization of the macromolecular machinery and the mechanisms involved in establishment and maintenance of 3D chromatin organization remains elusive. Moreover, the few comparative works done so far are revealing important differences on how different eukaryotic species structure their chromatin in the nuclear space. And last, but not least, the functional and evolutionary consequences of the 3D organization of the genome and its causal effects on DNA-templated processes are still poorly understood. Addressing these challenges requires a fully interdisciplinary approach involving the use of molecular, cellular, genomics, genetics, evolutionary, biophysical, physical, imaging and computational approaches. Our ultimate goal would be the construction of accurate and predictive mechanistic models that describe 3D genome architecture and uncover its functional relevance in health and disease. Many CSIC teams demonstrate strong complementary expertise in these different fields, which makes CSIC especially well situated to successfully address this challenge.

Introduction and general description

There is an intimate relationship between genome structure and function that is not fully understood yet. At the highest resolution, the nucleosome represents the basic structural subunit of chromatin and its discovery was essential for the discovery and subsequent understanding of epigenetic mechanisms that we now know are central to all aspects of genome function (i.e. RNA transcription, DNA replication, recombination, repair, chromosome segregation and genome stability). More recently, technological advances have provided us with a more complete and global view of the 3D organization of chromatin inside the cell nucleus. According to this emerging picture, the need to pack large eukaryotic genomes within the limited dimension of the nucleus is fulfilled through the hierarchical organization of DNA into structures of increasing compaction and complexity. The monotonous repetition of nucleosomes along the DNA molecule constitutes the elemental structural organization of the chromatin fiber. Chromatin fibers are subsequently folded into higher-order levels of structural organization operating at increasing genomic scales. Super-resolution microscopy analysis has shown that the interaction with linker histones H1 results in the formation of nucleosome clutches of variable size (Ricci et al., 2015). On the other hand, chromosome conformation capture (3C) and imaging approaches have revealed two additional levels of folding of the chromatin fiber, namely topologically associating domains (TADs) and compartments. TADs are insulated self-interacting genomic regions on the scale of few kilobases (kb) to megabases (Mb) (Dixon et al., 2012; Nora et al., 2012), while compartments reflect spatial clustering of large genomic regions, from tens to hundreds of Mb, that share similar epigenetic states (Bonev et al., 2017; Lieberman-Aiden et al., 2009; Rao et al., 2014). On top of that, the use of fluorescent whole chromosome paint probes unveiled that chromosomes occupy discrete areas within the nuclear space (Kempfer and Pombo, 2020). These structures, known as chromosome territories, showed limited intermingling between them and appear to distribute at preferential positions within the nucleus. Large gene-poor chromosomes preferentially locate at the nuclear periphery, close to the nuclear lamina, while small gene-rich chromosomes are frequently located within the central part of the nucleus.

3D genome architecture is likely to play important regulatory roles. Although direct causative roles in biological function remain scarce, 3D structure strongly correlates with genome function. In particular, two major types of compartments have been described: the *"A compartment"* that corresponds to transcriptionally active euchromatic regions marked with

active histone modifications, and the *"B compartment"* formed by transcriptionally silent heterochromatin decorated with repressive histone modifications. In this regard, TADs have been proposed to represent fundamental regulatory units in which most of the interactions between enhancers and their target genes take place. Moreover, during *Drosophila* embryo development, no signs of 3D chromatin folding are detected until the zygotic genome becomes transcriptionally active at the maternal-to-zygotic transition (Hug et al., 2017). Nonetheless, the evidence linking 3D structure and transcription is far from conclusive. Several studies have analyzed the effects on 3D chromatin organization of inhibiting transcription. In the bacteria *C. crescentus* and *B. subtilis*, inhibition of transcription results in a loss of contact domains (Le et al., 2013; Wang et al., 2017). However, similar experiments performed in *Drosophila* argue that transcription itself is not required for the organization in TADs and compartments (Hug et al., 2017). Thus, whether 3D folding is a cause or a consequence of function remains a matter of debate. Evolutionarily conservation of 3D chromatin organization is also under intense investigation. TAD-like self-interaction domains have been observed in a wide range of organisms, from mammals and other animal species (i.e. *Drosophila* (Rowley et al., 2017; Wang et al., 2018b)), to plants and fungi (Dong et al., 2017; Hsieh et al., 2015; Mizuguchi et al., 2014; Ricci et al., 2019; Wang et al., 2018a; Winter et al., 2018) and even in bacteria (Dame et al., 2020; Le et al., 2013). In this regard, a primordial mechanism that folds the large DNA molecules of genomes into looped domains appears to be conserved through evolution. This ancestral mechanism relies on the activity of SMC complexes (i.e. cohesin, condensin) and type-II DNA topoisomerases, both of which remain omnipresent in bacteria, archea and eukarya. However, many features of 3D chromatin organization are far from being universally conserved. A better understanding of the factors and mechanisms involved in establishment and/or maintenance of 3D chromatin organization must help to clarify these questions.

Despite the tremendous progress in studying 3D genome architecture that we have witnessed in the recent years, many essential questions remain unanswered. 3C-related methods, namely Hi-C, have greatly improved our current understanding of 3D chromatin organization, especially at a descriptive level. However, Hi-C (and other 3C-related methods) has intrinsic limitations, as it provides static, population-average (bulk) and pairwise measurements of physical interactions between pairs of sequences across the genome, which could in principle mask or distort fundamental principles of genome organization. In this regard, the actual structural features of TADs are still a matter of debate. Therefore, there is of great need to develop novel methodologies that provide us with a more dynamic, single-cell

and multi-way view of 3D genome architecture without losing the genome-wide and high-resolution properties of the Hi-C data.

A theoretical framework to understand the forces and principles governing chromatin folding is also missing. On the other hand, characterization of the macromolecular machinery and the mechanisms involved in establishment and maintenance of 3D chromatin organization remains elusive. Moreover, the few comparative works done so far are revealing important differences on how different eukaryotic species structure their chromatin in the nuclear space, but the molecular basis of these differences are not well understood. And last, but not least, the functional consequences of the 3D organization of the genome and its causal effects on DNA-templated processes are still poorly understood. Next, we review our current understanding of 3D genome organization and outline the main methodological, conceptual and translational challenges that the field is facing.

Impact in basic science panorama and potential applications

TECHNICAL AND METHODOLOGICAL CHALLENGES

Current methods for studying 3D chromosome and genome structure are based on few experimental approaches, mainly Hi-C and super-resolution microscopy. These methodologies present limitations, bias and technical difficulties that can obscure and blur the results obtained. Indeed, Hi-C based methods are based on cross-linking and ligation of genomic regions in close 3D proximity. But the need for ligation has also limited resolution and precluded regions that lie outside the ligation range from being incorporated into the overall structure. Moreover, fixation is chemical-dependent and so potentially biased to reveal only a subset of interactions. Metastable or transient interactions may be undetected or underrepresented. On the other hand, Hi-C matrices, from which TADs are inferred, do not represent chromatin contacts present in any one cell, but reflect instead pools of colligation events generated during multiple dynamic processes (i.e. loop extrusion, transcription, remodeling), which are in different positions in different cells. Super-resolution microscopy methods overcome some of these limitations since, in principle, they allow dynamic visualization of individual cells. However, detection of specific actors (protein tags and epitopes, DNA sequences) is limited and so biased by the number of available probes (dyes, fluorophores, antibodies, oligonucleotides) and their capacity to produce sustained and quantitative signals. Most importantly, there are only a few live cell imaging methods that can be used to analyze physical distances over time within individual cells. Finally, the current Hi-C and super-resolution technologies cannot trace yet the path of individual DNA molecules

and the internal architecture of macromolecular ensembles *in vivo*, which demand true nano-scale resolution. Future efforts should be devoted to implement or combine methods that provide us with a whole-genome, high-resolution, multi-way, single-cell and dynamic view of 3D genome organization. Next, we discuss methodological developments to address these challenges.

I. Imaging

Super-resolution and electron microscopy have been used to dissect genomic features, including sequence-specific super-resolution imaging of contact domains (Boettiger et al., 2016; Ou et al., 2017). Such super-resolution studies have provided first glimpses of the physical nature of contact domains, such as their volumes and shapes. Moreover, super-resolution microscopy methods using array-derived oligonucleotides, such as OligoSTORM and OligoDNA-PAINT, can be used to obtain high-resolution, single-cell and multi-way views of DNA folding in any cell type or tissue of interest (Bintu et al., 2018; Mateo et al., 2019). In addition, these novel imaging methods can be combined with RNA-FISH to simultaneously analyze ongoing transcription. However, these methods still offer limited genomic coverage given the limited number of oligos that can be used, the restricted size and number of regions that can be analyzed at the same time as well as the finite number of cells that can be monitored. Improvements in multi-color live cell imaging and the implementation of novel approaches using CRISPR-Cas9 based imaging (Chen et al., 2013) and dye based barcoding strategies (Beliveau et al., 2015) may also greatly increase our understanding of 3D genome organization in single cells. Another major limitation of the previously described methods is that they provide a static view of 3D structure. However, considering that the DNA-protein and protein-protein interactions that control DNA folding are highly dynamic, then, it is most logical that genome architecture is also dynamic. Methods based on the tagging of loci of interest (e.g. MS2 tagging system) can be used to analyze dynamics of 3D genome organization in individual life cells (Alexander et al., 2019; Fukaya et al., 2016). Unfortunately, these methods suffer from limited genomic coverage and resolution.

II. Hi-C

To date, most Hi-C data generated are based on pairs of contacts. New technologies should allow the analysis of multi-contacts as well as inter-chromosomal interactions, which today we believe are less frequent but could well be undetectable with the current technologies. Sequencing-based methods orthogonal to Hi-C have been recently established (e.g. GAM, SPRITE), which can provide with genome-wide, high-resolution and multi-way views of 3D genome organization (Beagrie et al., 2017; Quinodoz et al., 2018). Moreover, these new

methods seem to be particularly suited to detect very long-range inter-chromosomal contacts. Unfortunately, like Hi-C, these methods are based on population-average measurements across hundreds/thousands of cells. On the other hand, although single-cell Hi-C methods have been already implemented, they still suffer from low resolution and pairwise measurements of physical interactions (Nagano et al., 2013). Improvements in either of these aspects of the methodology will have major impacts on our understanding of 3D genome architecture.

III. Modeling

Another challenge is to reconstruct high-resolution 3-D models of large genomes from Hi-C data, which is needed for studying detailed interactions between genes and regulatory elements. The enormous time complexity and data sparsity associated with high-resolution modeling are strong constraints. Despite the improvement in 3-D structure modeling approaches, the lack of a real structure with which to contrast these models remains a challenge. In particular, it is currently difficult to confirm the true modeling capability of 3-D genome methods. In addition, given the possible connection of 3-D genome structure alteration with disease, it is important to make 3-D genome modeling methods easy for biomedical scientists to use in their research.

IV. cryo-EM

There is currently a revolution occurring in cryo-EM based characterization of large macromolecular complexes (Bai et al., 2015). Whether these methods can be extended to study large chromosomal domains inside of cells is unclear. However, improved and adapted cryo-EM methods will likely be essential for unraveling the structure of the protein complexes that are critical for 3D genome organization.

V. Phase-separation

One of the challenges will be to reconcile the chromatin domains and TADs mapped from Hi-C data with the presence of condensates or membrane-less nuclear organelles formed through different types of phase-separation processes. Efforts must be put into the development of techniques that allow the molecular characterization of these condensates *in vivo*, allowing their visualization inside the cells, the isolation of their content and fine mapping of their 3D interactions. Undoubtedly, this constitute a big challenge due to the transient, heterogeneous and dynamic nature of these structures and therefore requires a multidisciplinary approach, a close collaboration between cell biologists, physicists and chemists will be essential in the development of new technologies that allow to address these questions.

VI. DNA topology

Current Hi-C and imaging technologies cannot trace yet the path of individual DNA molecules. Thus, novel methodologies relying on the topological analysis of intracellular DNA are needed to infer how the path of DNA molecules is deformed (i.e. DNA helical twisting and axial bending) by individual chromatin elements *in vivo*. Forces acting on DNA molecules affect the conformational equilibrium and drive transitions of chromatin architecture. This is the case of the DNA superhelicity (double-helical tension) generated during genome transactions (transcription, replication), which propagates and interplays deeply with chromatin structure and function. How to assess the superhelical tension of intracellular DNA has been a long-standing challenge.

VII. In vitro studies

Another important challenge is the *in vitro* reconstruction of macromolecular ensembles that direct/reflect chromatin folding states (SMC complexes, chromatin remodelers, condensate- and phase-transition agents), and subsequent analysis of their physico-chemical and mechanistic properties (optical and magnetic tweezers, AFM, cryo-EM, DNA topology, enzyme biochemistry).

FACTORS AND MECHANISMS GOVERNING 3D GENOME ARCHITECTURE

Most of what we know about 3D genome organization and its underlying molecular mechanisms comes from the study of vertebrates, especially mammals, and a handful of additional model species, in particular the fruit fly *Drosophila melanogaster*. Here we will briefly review the mechanisms that are known to control the different levels of 3D genome folding in vertebrates and their potential conservation through evolution. Most importantly, we will also highlight important mechanistic questions that remain unresolved. Although 3D genome architecture is likely to impact all DNA-templated processes, here we will preferentially focus on transcription during interphase.

I. TADs and Compartments

At a large genome scale (tens to hundreds of Mb), Hi-C studies revealed that chromosomes are organized into compartments resulting from the spatial clustering of genomic regions with a similar chromatin and transcriptional state (Bonev et al., 2017; Le et al., 2013; Rao et al., 2014). Two major types of compartments have been described: the *"A compartment"* includes genomic loci that are gene-rich, transcriptionally active and marked with active histone modifications (e.g. H3K27ac, H3K4me2/3); the *"B compartment"* is formed by gene-poor loci that are transcriptionally silent and marked with repressive histone modifications (e.g. H3K9me2/3, H3K27me3). The *"B compartment"* can be further subdivided into (i)

constitutive heterochromatin loci typically found in the nuclear periphery (as part of Lamina-Associated Domains (LADs)) and nucleoli, marked with H3K9me2/3 and DNA hypermethylated; (ii) facultative heterochromatin loci, which are bound by PcG protein complexes, DNA hypomethylated and located within the nuclear interior.

At a sub-megabase scale (tens to hundreds of Kb), compartments can be subdivided into TADs (Dixon et al., 2012; Nora et al., 2012). TADs are separated from each other by TAD boundaries or borders, which typically coincide with CTCF binding sites and, to a lesser extent, with housekeeping genes, tRNA genes and SINE repeats (Dixon et al., 2012). Importantly, TADs have been proposed to represent fundamental regulatory units in which most of the interactions between enhancers and their target genes take place. Therefore, TADs might (i) facilitate the interactions between enhancers and their target genes and (ii) insulate genes from establishing ectopic interactions with the wrong enhancers (Spielmann et al., 2018).

Following the discovery of TADs and compartments, major efforts have been devoted to elucidate the mechanisms implicated in their formation. Recent studies have conclusively demonstrated that TAD formation depends on both CTCF and the Cohesin complex (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Zhang et al., 2019), since the depletion of either of these two proteins leads to an almost complete loss of all TADs. Moreover, TAD formation seems to be explained by a "loop extrusion" mechanism, whereby cohesin is initially loaded at enhancers and promoters to then form progressively larger loops that eventually stall at TAD boundaries formed by convergent CTCF motifs (Fudenberg et al., 2016; Rao et al., 2014). The mechanistic details of how this "loop extrusion" model actually works are currently being elucidated. For example, it has been now demonstrated that human cohesin can extrude DNA loops symmetrically, rapidly and in an ATP-dependent manner (Davidson et al., 2019; Kim et al., 2019; Vian et al., 2018). Similarly, recent structural studies explain how the interaction interface between cohesin and CTCF leads to the preferential emergence of TAD boundaries at convergent CTCF sites (Li et al., 2020). Despite these major advances, the functional relevance and even the existence of TADs is a subject of intense scientific debate. It has been proposed that TADs could represent an emergent property from cell population averaging as measured by Hi-C and that they could represent a computational artifact devoid of biological significance (Rowley et al., 2017). However, recent work based on single-cell Hi-C and super-resolution microscopy confirmed that TAD-like structures actually exist within individual cells (Bintu et al., 2018; Nagano et al., 2013). Moreover, although TAD boundaries showed variation between individual cells, they still frequently

overlapped with CTCF sites (Bintu et al., 2018). Interestingly, in the absence of Cohesin, TAD-like structures were still observed within individual cells, although their boundaries became highly variable and did not coincide with CTCF sites (Bintu et al., 2018).

TADs and compartments were initially considered as two hierarchically related layers of 3D genome organization operating at different genomic scales. However, recent studies in which cohesin was inducibly depleted suggested that they might actually represent two independent modes of chromatin organization (Rao et al., 2017; Schwarzer et al., 2017). Upon loss of cohesin TADs vanished completely, while compartment segregation was even reinforced and became evident at a finer sub-megabase scale that reflected the underlying chromatin state. This lead to the suggestion that cohesin-dependent loop extrusion and homotypic chromatin interactions represent two independent and opposing mechanisms of 3D organization (Schwarzer et al., 2017). Moreover, recent work has also clarified some of the mechanisms whereby homotypic chromatin interactions can lead to the segregation of distal loci into specific compartments. Importantly, these mechanisms seem to be specific for each compartment type and are currently better understood for the *B-compartment*. Namely, the spatial segregation of loci marked with constitutive heterochromatin seems to depend on tethering to the nuclear lamina or nucleoli and phase separation driven by HP1 (Larson et al., 2017; Strom et al., 2017; van Steensel and Belmont, 2017). On the other, the spatial clustering of loci bound by PcG is mediated by subunits of the canonical PRC1 complex, including phase separation driven by CBX2 and polymerization due to the SAM domains present in the PHC1/2 subunits (Isono et al., 2013; Plys et al., 2019; Tatavosian et al., 2019).

The discovery of TADs and Compartments has dramatically changed the way we think about genome architecture. However, there are still important questions regarding the molecular forces and mechanisms controlling these two structural features that need to be solved and that could represent important areas of future research.

II. Short-range genomic interactions

The relatively low resolution of initial Hi-C studies precluded a precise view of intra-TAD architecture, but this has changed due to the increased resolution of Hi-C studies and the use of targeted 3C-based approaches (4C-seq, capture Hi-C, ChIA-PET, HiChIP) that provide lower genomic coverage by higher resolution (Hughes et al., 2014; Mumbach et al., 2016; Tang et al., 2015; van de Werken et al., 2012). As a result, a complex 3D organization within TADs is starting to emerge, which includes multiple and sometimes overlapping topological entities, such as sub-TADs, loop domains or insulated neighborhoods (Dowen et al., 2014; Phillips-Cremins et al., 2013; Rao et al., 2014). The mechanisms controlling these intra-TAD

layers of 3D organization seem to be similar to those describe for TADs and, thus, include cohesin-dependent loop extrusion and interactions between CTCF and cohesin. However, these intra-TAD structures are more variable and cell-type specific and their functional significance has not been extensively studied yet.

Another major group of intra-TAD interactions include those that bring enhancers and their target genes into physical proximity. With the discovery of TADs and the confirmation of cohesin dependent loop extrusion as the mechanism involved in TAD formation, it was initially thought that most enhancer-gene interactions would depend on cohesin and loop extrusion (Dixon et al., 2012; Kagey et al., 2010; Phillips-Cremins et al., 2013). Similarly, many enhancer-gene interactions were suggested to depend on dimerization of CTCF molecules bound to enhancers and gene promoters (Guo et al., 2015). However, accumulating evidences indicate that CTCF and Cohesin are only involved in a limited number of enhancer-gene contacts and alternative mechanisms might be more prevalent (Rao et al., 2017; Schwarzer et al., 2017). For example, in an analogous manner to CTCF, YY1 can bind to both enhancers and promoters and form dimers that facilitate the interactions between these regulatory elements (Beagan et al., 2017). Dimerization is also the mechanism through which the co-activator protein LDB1 can facilitate long-range enhancer-gene contacts (Deng et al., 2014). On the other hand, the biophysical process of liquid-liquid phase separation might also be implicated in enhancer-gene interactions (Hnisz et al., 2017). In this model, enhancer and promoters serve as recruitment platforms for multiple proteins (*e.g.* TFs, RNA Pol2, Mediator, histone modifications) and RNA molecules that then engage into weak but multivalent and cooperative interactions. As a result, membraneless organelles or condensates with gel-like properties can form, within which enhancer-gene contacts are facilitated and established (Hnisz et al., 2017). A central component of the phase separation model are the intrinsically disordered regions (IDR) found in multiple transcription factors and co-activators, such as certain subunits of the Mediator complex (Boija et al., 2018; Sabari et al., 2018). These IDR facilitate phase separation by serving as flexible and pleiotropic platforms for protein-protein interactions. However, although the existence of transcriptional condensates has been confirmed *in vivo*, most of the insights regarding the mechanisms implicated in their formation come from *in vitro* experiments. For example, MED1-IDR can form phase-separated droplets in vitro, yet complete loss of the Mediator complex has almost no impact on 3D genome organization or enhancer-gene interactions (El Khattabi et al., 2019; Sabari et al., 2018).

Enhancers can display poised or primed states, which are hypothesized to facilitate the future activation of their target genes (Creyghton et al., 2010; Rada-Iglesias et al., 2011). Both poised and primed enhancers have been reported to physically interact with their target genes before transitioning to an active state (Cruz-Molina et al., 2017; Ghavi-Helm et al., 2014). In the case of poised enhancers, these pre-formed contacts seem to be mediated by PcG complexes bound to both the enhancers and their target gene promoters. These PcG-dependent contacts are mediated by the polymerization capacity of the SAM domains present at PHC1/2 proteins, which are key components of the canonical PRC1 complex (Isono et al., 2013). In the case of primed enhancers, the pre-formed interactions with the target genes might require the presence of H3K4me1 and MLL3/4 proteins (Yan et al., 2018). It has been proposed that H3K4me1 might facilitate the recruitment of cohesin to enhancers, which can then bring genes and enhancers into physical proximity (Yan et al., 2018). Last but not least, silencers represent another major class of regulatory elements that can negatively influence the expression of their target genes. Silencers have been historically difficult to identify and characterize and, thus, the topological features of silencers remain largely unknown. However, recent studies indicate that some silencers might physically contact with their inactive target genes (Ngan et al., 2020; Pang and Snyder, 2020).

In summary, intra-TAD contacts are highly complex, involving multiple and diverse 3D structures. Consequently, there are still many mechanistic questions that should be addressed in the coming years.

III. Are TADs universal units of genome organization?

Self-interaction domains superficially similar to mammalian TADs have been observed in a wide range of organisms, from other animal species (i.e. *Drosophila* (Rowley et al., 2017; Wang et al., 2018b)), to different eukaryotic lineages such as plants and fungi (Dong et al., 2017; Hsieh et al., 2015; Mizuguchi et al., 2014; Ricci et al., 2019; Wang et al., 2018a; Winter et al., 2018) and bacteria (Dame et al., 2020; Le et al., 2013). However, many features of these TAD-like structures are far from being universally conserved, with differences in their size, the sharpness of their boundaries, intensity of the contacts and genomic content (i.e. gene rich and gene poor regions, abundance of transposable and repetitive elements, etc). Still, a common theme seems to emerge from the comparison of different plant, fungi, apicomplexan and animal species (Rowley et al., 2017). In all these cases, transcriptional activity and the chromatin features associated to it create boundaries between contact domains, partitioning the genome into alternating transcriptionally active and inactive genomic regions (Rowley et al., 2017).

Nevertheless, this shared organizational principle can lead to very different contact patterns, depending on the degree of genome compactness of each species and the distribution and density of transcribed genes. For instance, in *Drosophila*, transcriptionally active genes are frequently found in clusters of different sizes that separate large TADs corresponding to repressed regions (Polycomb repressed enriched in H3K27me3 as well as classical heterochromatin enriched in H3K9me2) (Rowley et al., 2017; Wang et al., 2018b). In contrast, in the filamentous fungi *Epichloë festucae*, there is an alternation between domains containing very repeat-rich blocks of DNA and gene-rich regions that are almost repeat-free (Winter et al., 2018).

IV. Evolution of long-range regulatory interactions

Besides these genomic differences, several lineages have evolved additional molecular mechanisms on top of the pre-existing compartmentalization found across studied eukaryotes, such as the presence of insulator/architectural proteins. In vertebrates, there is an additional layer of organization and homotypic chromatin interactions can be at least partially superseded by the presence of inverted CTCF binding sites. As we mentioned above, CTCF together with other proteins such as YY1, allow the formation of chromatin loops between long-range cis-regulatory elements and their target gene promoters. Given that YY1 and CTCF are exclusively found in animals and bilaterian animals respectively (Heger et al., 2012; Irimia and Maeso, 2019), the origin of these architectural proteins could explain the evolution of long-range regulation and large-scale TADs typically found around animal developmental genes (Harmston et al., 2017). This hypothesis would be supported by the fact that some nematode species such as *Caenorhabditis elegans*, have lost CTCF and YY1 and these loses seem to be concomitant with the dismantling of most of the ancestral long-range regulation in these lineages, which are characterized by their highly compact, gene-dense genomes (Crane et al., 2015; Heger et al., 2012; Heger et al., 2009; Jabbari et al., 2018). The situation is however more complex than this putative scenario, and understanding the relationship between the evolution of long-range regulatory interactions and architectural proteins will require a significant amount of work in a much wider range of animal and non-animal species. For instance in *Drosophila*, CTCF is not required for the establishment of chromatin loops (Rowley et al., 2017). The evolution of multiple novel architectural proteins in the insect and fly lineages (Heger et al., 2013; Pauli et al., 2016) could have contributed to the loss of the putatively ancestral role of CTCF, but this hypothesis has not been completely confirmed.

Furthermore, although long-range regulation is ancestral to all animals (Gaiti et al., 2017a; Gaiti et al., 2017b; Grau-Bove et al., 2017; Irimia et al., 2013; Irimia et al., 2012; Schwaiger et al., 2014; Sebe-Pedros et al., 2016), CTCF is not, since it originated after the divergence of bilaterian animals (Heger et al., 2012). Thus, distal regulation in metazoan ancestors was CTCF-independent, and though not yet tested experimentally, this might also be the case in extant non-bilaterian lineages such as sponges and cnidarians (Irimia and Maeso, 2019).

Finally, recent studies across different plant species have shown that distal cis-regulatory elements are widespread among angiosperms and, similar to what happens in certain animal lineages such as nematodes, only those species with extremely compact gene-dense genomes such as the model species *Arabidopsis thaliana* are largely devoid of long-range regulatory interactions (Lu et al., 2019; Ricci et al., 2019). How long-range cis-regulatory loops are maintained in plants is a completely open question and it is currently unknown if these loops are the consequence of compartmental segregation, if they are established by (so far undescribed in plants) sequence-specific architectural proteins or the combined action of both type of molecular mechanisms (Ricci et al., 2019).

In sum, there seem to be many different types of TAD-like structures across different organisms, which can be formed by different molecular mechanisms, some of which are shared across eukaryotes while others are lineage-specific. Thus, although these diverse mechanisms may lead to the formation of superficially similarly interaction domains, it is currently unclear to what extent these 3D domains represent ancestral structural features or have evolved convergently as a result of shared functional properties (Szabo et al., 2019).

PHYSIOLOGICAL IMPACT OF 3D GENOME STRUCTURE

Higher-order chromatin structures emerge as putative building blocks of the genome that are supposed to have a functional role. However, finding this role has turned more complex than anticipated. Direct causative roles in biological function have been hard to find, and, in other cases, it appears that structure might be a secondary result of other genome functions. In this section, we address what is known about the role of 3D chromatin structure in other basic functions of the genome, as well as in different physiological contexts.

Interplay of 3D structure with genome function

I. Replication

Several studies have addressed the relationship between DNA replication and 3D genome structure (reviewed in (Marchal et al., 2019)). Early and late replication timing domains correspond roughly to *A* and *B* compartments respectively, while, at smaller genomic scale,

there is a good correlation between TADs and replication domains. The recent description of early replicating control elements (ERCEs) suggest a causal link where replication has an instructive role on 3D genome structure (Sima et al., 2019). Further evidence for replication not depending on 3D structure comes from studies in early stages of development, where replication takes place while 3D structure has not yet been established (reviewed in (Hug and Vaquerizas, 2018)).

II.Transcription

It has long been assumed that the primary effect and direct consequence of 3D chromatin organization would be the differential transcription of the genes located in different structural domains. This is most evident for *A* and *B* compartments, that correspond to euchromatin and heterochromatin respectively, and are enriched in active and inactive genes in turn. Genes in regions that switch from *A* to *B* compartment during differentiation reduce their expression and *vice versa* in the case of changes from *B* to *A* (Dixon et al., 2015). Regarding the functional relevance of TADs as fundamental regulatory units of gene expression, there are a number of conflicting reports. On one hand, structural variants (deletions, inversions, duplications) that disrupt TAD organization can lead to either a loss of endogenous enhancer-gene interactions ("enhancer disconnection") or a gain of ectopic enhancer-gene interactions ("enhancers adoption") that can lead to pathological gene silencing or activation, respectively (Laugsch et al., 2019; Lupiáñez et al., 2015; Smol et al., 2020; Spielmann et al., 2018). On the other hand, the structural disruption of certain TADs does not have any major effects on gene expression (Ghavi-Helm et al., 2019; Laugsch et al., 2019), while the global disruption of TAD organization due to the loss of either CTCF or Cohesin leads to moderate gene expression changes (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017). As for loops or self-interacting domains, these structures would bring into close proximity enhancers and other cis-regulatory elements with their target genes, thus facilitating their proper regulation (reviewed in (Schoenfelder and Fraser, 2019)). However, careful analysis of paradigmatic cases of long-range enhancer elements, such as those of the *Shh* and *Sox2* loci, indicates that TAD structure or physical enhancer-promoter contact mediated by chromatin looping is not a prerequisite for activity (Alexander et al., 2019; Benabdallah et al., 2019; Williamson et al., 2019).

Therefore, there is contradicting evidence for the role of 3D genome organization in regulating gene expression, and maybe it has a facilitating rather than an instructive role. In addition, transcriptional activity has an impact on chromatin structure. Therefore, there is a

dynamic cross talk between transcription and genome organization, where each can modulate the activity of the other (reviewed in (van Steensel and Furlong, 2019)).

Physiological significance of 3D genome structure

I. Development and evolution

Normal development depends, not only on the linear sequence of the genome incorporating millions of CREs, but also on the 3D organization of chromatin. Chromatin organizes the interaction between CREs and their target genes and, thus, modulates biological processes crucial to cell differentiation and development. TADs can be simply defined as functional genomic units of gene regulation, in which CREs interact with their target genes. TADs are positionally very stable between cell types and tissues independently of transcriptional status (Dixon et al., 2015). In addition, the activity of promoters and enhancers seems to be very uncoordinated within the same TAD. This has suggested the existence of a very stable and pre-formed topology that establishes the physical proximity between enhancers and their target genes, though doubts about its actual role in the specific regulatory processes within cells and tissues have been raised (Lupiáñez et al., 2015). Linage-specific alterations of the 3D genome organization often occurs within TADs or sub-TADs. Therefore, to determine the influence of 3D chromatin organization on linage-specific genetic regulation, it is essential to generate maps of chromatin interactions to a sufficiently high resolution to distinguish individual regulatory elements and for a wide range of tissues and cells during normal development. Currently, Capture-C/Hi-C and PLAC-seq/HiChIP (Hui & Wei., 2019) are being used to generate such high-resolution maps. Thus, further development of these technologies appears crucial.

The study of the positions of CREs and the genes they regulate in relation to TADs offers numerous opportunities to the study of gene expression variation during evolution. To date, the limited evidence available has been obtained from a study of conserved non-coding evolutionary elements (CNEs) (Gómez-Marín et al., 2015). These elements are organized in syntenic locations, primarily around key developmental genes (Gómez-Marín et al., 2015). The study of these elements has allowed to conclude that, at least around developmental genes, TADs are evolutionarily conserved structures that may play a role in maintaining the correlation between CREs and their target genes. Another important point in the evolutionary conservation of TADs is that during the course of evolution, animal genomes have experienced a profound reorganization, changing the relative order of TADs and this restructuring has caused important changes in gene expression that have resulted in the

appearance of evolutionary novelties. What remains to be described however, are the molecular mechanisms underlying this reorganization of TADs during evolution. Moreover, we still do not know precisely how CREs interact with their target genes. Understanding the mechanisms that facilitate functional interactions both within and between TADs is essential to understanding the control of gene expression during development in an evolutionarily perspective.

II. Disease.

The description of 3D genome organization quickly led to address if it could be possibly involved in different aspects of human disease. This relationship can be twofold. In the first place, a direct link could be occurring in which disruption of 3D chromatin structure leads to a pathological state. On the other hand, analysis of genome structure can give us novel insight into disease mechanisms.

Structural variations of the genome, such as deletions or inversions, can lead to relocation or loss of TAD boundaries, resulting in rearrangement of enhancer-promoter interactions that can cause pathology because of incorrect gene expression. These alterations have been named "TADopathies", and can explain the molecular basis of autosomal dominant adult-onset leukodystrophy (ADLD) and some congenital limb malformations (Matharu and Ahituv, 2015). While it appears that TAD rearrangements are not necessarily causative of disease (Smol et al., 2020), TAD disruption should be taken into account when analyzing the pathological effect of structural variations linked to disease. Furthermore, it should be stressed that their effect could be occurring on genes that are not included in the pathological deletion or inversion. For example, deletion of a TAD boundary can result in mis-expression of genes located outside of the deleted segment (Spielmann et al., 2018).

Mutation of chromosomal architectural proteins, such as CTCF or the cohesin complex, are frequently found in different types of cancer. In addition, changes in the binding of CTCF to DNA have been shown to be causative of disease. In a striking example described in gliomas, hypermethylation of a CTCF site, caused by a gain-of-function mutation of *IDH*, that produces a metabolite that leads to decreased activity of TET demethylases, leads to dysregulation of a boundary and ectopic activation of *PDGFRA*, which encodes a potent oncogenic driver (Krijger and de Laat, 2016). Surely more examples related to aberrant methylation of CTCF sites await discovery, again showing that the knowledge of chromatin 3D structure can reveal unexpected genes as causative in a particular disease or condition. We must also be aware that the non-random and locally heterogeneous nature of genome structure has a direct impact in the distribution and nature of mutations, such as those we see in cancer.

Different chromatin states impose constrains to sequence as well as physical availability of nucleotides to mutagenic agents, what has a direct impact on their mutability (Akdemir et al., 2018; Schuster-Böckler and Lehner, 2012).

Finally, the study of 3D genome structure provides an invaluable tool to link disease variants identified by Genome Wide Association Studies (GWAS) to the underlying casual genes. A very high proportion of disease or trait-associated variants lie in no-coding intergenic regions of the genome, most probably containing regulatory elements. These elements control the transcription of genes, but often assigning the correct gene to a putative regulatory element is not obvious. Although the default is assigning the nearest gene to GWAS variants, this may not be correct. Knowledge on the 3D genomic structure of the region, such as the distribution of TADs or loops, can aid in correctly identifying target genes of disease variants (Krijger and de Laat, 2016).

Key challenging points

I. *Towards a dynamic, single-cell and multi-way view of 3D genome organization*

Our current understanding of 3D genome organization lacks a more dynamic, single-cell and multi-way view. Therefore, there is a great need to develop novel methodologies and techniques that provide us with a dynamic high-resolution structural view of the whole-genome at the single-cell level. Ideally, such methods should be combined with the analysis of other DNA-templated processes (transcription, replication, DNA repair) in order to answer some major questions regarding the biological relevance and mechanistic basis of 3D genome architecture.

II. Towards an evolutionary view of 3D genome organization

It is currently difficult to know how much of what we have learned in mammals and some common model organisms can be extrapolated to other species and to differentiate ancestral versus lineage-specific features, hampering our ability to draw general conclusions about the mechanisms responsible for 3D genome organization and to which biological functions it contributes in different organisms. As a matter of fact, the few non-vertebrate and non-animal species studied so far have revealed important differences across lineages. Hence, it appears imperative to characterize 3D chromatin organization in a much wider range of species than those currently tested, including representatives of diverse eukaryotic lineages.

III. Towards a unified mechanistic and theoretical framework of 3D genome organization

A unified view of the mechanisms governing 3D genome organization is lacking. The segregation of genomic loci into compartments strongly correlates with the underlying chromatin and epigenetic landscapes. However, the mechanisms that enable the preferential interaction between *loci* with similar chromatin state (i.e. homotypic chromatin interactions) remain so far largely undetermined. Phase-separation, either liquid or polymer-based, could provide a general biophysical basis for chromatin folding. Furthermore, the formation of phased-separated transcriptional condensates emerges as a potential mechanism to co-regulate gene expression. However, phase-separated condensates are difficult to study *in vivo*. Current models propose multivalent interactions between many protein and RNA components, suggesting a highly redundant system that might be robust to the disruption of individual components (e.g. Mediator). Therefore, which kind of experimental approaches should be designed to dissect the mechanisms and molecular forces controlling phase separation?. What are the dynamics of transcriptional condensate formation and which mechanisms control their dismantling? How can we integrate transcriptional condensates based on phase separation and TADs based on loop extrusion?. How many genes and enhancers can be found and co-regulated within a single phase-separated condensate?. Do specialized condensates exist within individual cells in which only certain genes and enhancers are transcribed and which is the basis of this specificity?. Vertebrate genomes contain thousands of enhancers, which are bound by a large number of highly diverse transcription factors and co-activators. Taking this into consideration, different types of enhancers might utilize distinct mechanisms (e.g. polymerization) to communicate with their target genes and that should be systematically dissected and the dynamics of these interactions compared to those occurring within condensates. These are only some of the many mechanistic questions that await an answer.

IV. Towards predictive models of 3D genome organization and its functional consequences

Our current knowledge is still insufficient to model 3D genome folding directly from genomic data and, even more, we do not fully understand the functional implications of chromatin folding. For instance, the disruption of only certain TADs has measurable consequences on gene expression, but we do not understand the genetic or epigenetic features that distinguish “functional” from “non-functional” TADs. Similarly, we do not know whether all TADs and TAD boundaries are mechanistically and functionally equivalent. A large effort is required to gather sufficient information in a wide variety of cell types and during development and differentiation to be able to build models of 3D chromatin folding, and predict functional outcomes and pathological consequences of structural variants. In parallel, computational instruments must be generated to make 3D genome folding analysis amenable to non-specialists, particularly for those in the biomedical field.

CSIC advantage position and multi/inter-disciplinarity, strategic plan and resources

Addressing the challenges that the 3D GENOME ARCHITECTURE field is facing requires a coordinated study from a multidisciplinary approach, a close collaboration between geneticists, biochemists, cell biologists, computational biologists, physicists and chemists will be essential in the development of new technologies that allow to address these questions. From this point of view, CSIC is in a privileged position to take this endeavor and become a node to bring together Spanish research groups interested in responding to the challenges posed. Most important, several CSIC groups are already involved in setting a European initiative in this field (LifeTime), which is at the final stages of preparation. At the institutional level, CSIC could impulse the participation of Spain in the LifeTime initiative, promote *ad hoc* meetings and coordinate the evaluation of relevant projects for funding. We anticipate that CSIC's efforts in this field will have strong scientific and economical returns, increasing international visibility through the participation in strong collaborative projects at the european and international levels, producing research of excellence and top-level publications, and, in turn, resulting in the development of patents to translate basic research into commercially viable applications.

The strategic plan will be implemented in the following steps:

i) *Screening.* Identify CSIC researchers, as well as external collaborators, who may be interested in participating in the 3D GENOME ARCHITECTURE initiative.

ii) *Meeting.* Coordinate, evaluate and identify the most relevant work topics through scientific meetings called for this specific purpose.

iii) *Projects*. Promote the elaboration of highly innovative (risky) collaborative projects on the identified topics.

iv) *External evaluation* of the projects

v) *Financial Support.* Provide intramural financial support (budget, personnel) to the selected projects to bring them to a competitive stage to be presented to national, European and international calls.

References

Akdemir, K.C., Le, V.T., Killcoyne, S., King, D.A., Li, Y.-P., Tian, Y., Inoue, A., Amin, S., Robinson, F.S., Herrera, R.E., *et al.* (2018). Process-specific somatic mutation distributions vary with three-dimensional genome structure. bioRxiv 426080.

Alexander, J.M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S., and Weiner, O.D. (2019). Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. eLife *8*.

Bai, X.C., McMullan, G., and Scheres, S.H. (2015). How cryo-EM is revolutionizing structural biology. Trends Biochem Sc *40*, 49-57.

Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res *27*, 1139–1152.
Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R., *et al.* (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519-524.
Beliveau, B.J., Boettiger, A.N., Avendano, M.S., Jungmann, R., McCole, R.B., Joyce, E.F., Kim-Kiselak, C., Bantignies, F., Fonseka, C.Y., Erceg, J., *et al.* (2015). Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. Nat Commun *6*, 7147.
Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D., Grimes, G.R., Therizols, P., and Bickmore, W.A. (2019). Decreased enhancer-promoter proximity accompanying enhancer activation. Mol Cell, 473-484.e477
Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science *362*, eaau1783.
Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.T., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418-422.
Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., *et al.* (2018). Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell *175*, 1842-1855.e1816.
Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., *et al.* (2017). Multiscale 3D genome rewiring during mouse neural development. Cell *171*, 557-572.e524.
Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., *et al.* (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479-1491.
Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240-244.
Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA *107*, 21931–21936.
Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W.F.J., Grosveld, F., Frommolt, P., Bazzi, H., *et al.* (2017). PRC2 facilitates the regulatory topology required for poised enhancer function during pluripotent stem cell differentiation. Cell Stem Cell *20*, 689-705.e689.
Dame, R.T., Rashid, F.M., and Grainger, D.C. (2020). Chromosome organization in bacteria: mechanistic insights into genome structure and function. Nat Rev Genet *21*, 227-242.
Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J.M. (2019). DNA loop extrusion by human cohesin. Science *366*, 1338–1345.
Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A., *et al.* (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. Cell *158*, 849–860.
Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., A.Y., L., Ye, Z., Kim, A., Rajagopal, N., Xie, W., *et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.
Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.
Dong, P., Tu, X., Chu, P.Y., Lu, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D chromatin architecture of large plant genomes determined by local A/B compartments. Mol Plant, 1497-1509.
Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schujiers, J., Lee, T.I., Zhao, K., *et al.* (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.
El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.R., Park, S., Wang, X., *et al.* (2019). A pliable Mediator acts as a functional rather than an architectural bridge between promoters and enhancers. Cell *178*, 1145-1158.e1120.
Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of chromosomal domains by loop extrusion. Cell Rep *15*, 2038–2049.
Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer control of transcriptional bursting. Cell *166*, 358–368.

Gaiti, F., Calcino, A.D., Tanurdzic, M., and Degnan, B.M. (2017a). Origin and evolution of the metazoan non-coding regulatory genome. Dev Biol *427*, 193-202.
Gaiti, F., Jindrich, K., Fernandez-Valverde, S.L., Roper, K.E., Degnan, B.M., and Tanurdzic, M. (2017b). Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. Elife *6*.
Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., *et al.* (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell *176*,, 377-390.e319.
Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O., and Furlong, E.E.M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat Genet *51*, 1272–1282
Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.
Gómez-Marín, C., Tena, J.J., Acemel, R.D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., *et al.* (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci USA *112*, 7542-7547.
Grau-Bove, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A., and Ruiz-Trillo, I. (2017). Dynamics of genomic innovation in the unicellular ancestry of animals. Elife *6*.
Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., *et al.* (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell *162*, 900–910.
Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun *8*, 441.
Heger, P., George, R., and Wiehe, T. (2013). Successive gain of insulator proteins in arthropod evolution. Evolution *67*, 2945-2956.
Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., and Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci USA *109*, 17507-17512.
Heger, P., Marin, B., and Schierenberg, E. (2009). Loss of the insulator protein CTCF during nematode evolution. BMC Mol Biol *10*, 84.
Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A phase separation model for transcriptional control. Cell *169*, 13–23
Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping nucleosome resolution chromosome folding in Yeast by Micro-C. Cell *162*, 108-119.
Hug, C.B., Grimaldi, A.G., Kruse, K., and Vaquerizas, J.M. (2017). Chromatin architecture emerges during zygotic genome activation independent of transcription. Cell *169*, 216-228.e219.
Hug, C.B., and Vaquerizas, J.M. (2018). The birth of the 3D genome during early embryonic development. Trends Genet *34*, 903-914.
Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet *46*, 205–212.
Irimia, M., and Maeso, I. (2019). Boosting Macroevolution: Genomic Changes Triggering Qualitative Expansions of Regulatory Potential. In Old Questions and Young Approaches to Animal Evolution 2019, Fascinating Life Sciences, M.-D. J., and V. B., eds. (Springer, Cham).
Irimia, M., Maeso, I., Roy, S.W., and Fraser, H.B. (2013). Ancient cis-regulatory constraints and the evolution of genome architecture. Trends Genet *29*, 521-528.
Irimia, M., Tena, J.J., Alexis, M.S., Fernandez-Miñan, A., Maeso, I., Bogdanovic, O., de la Calle-Mustienes, E., Roy, S.W., Gómez-Skarmeta, J.L., and Fraser, H.B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res *22*, 2356-2367.
Isono, K., Endo, T.A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B.E., and Koseki, H. (2013). SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. Dev Cell *26*, 565–577.
Jabbari, K., Heger, P., Sharma, R., and Wiehe, T. (2018). The Diverging Routes of BORIS and CTCF: An Interactomic and Phylogenomic Analysis. Life (Basel) *8*.
Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., *et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.
Kempfer, R., and Pombo, A. (2020). Methods for mapping 3D chromosome architecture. Nat Rev Genet *21*, 207-226.
Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J., and Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. Science *366*, 1345–1349.

Krijger, P.H., and de Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol *17*, 771-782.
Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature *547*, 236–240.
Laugsch, M., Bartusel, M., Rehimi, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., Zentis, P., Nikolic, M., Bleckwehl, T., Kolovos, P., *et al.* (2019). Modeling the pathological long-range regulatory effects of human structural variation with patient-specific hiPSCs. Cell Stem Cell *24*, 736-752.e712.
Le, T.B., Imakaev, M.V., Mirny, L.A., and Laub, M.T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. Science *342*, 731-734.
Li, Y., Haarhuis, J.H.I., Sedeño Cacciatore, Á., Oldenkamp, R., van Ruiten, M.S., Willems, L., Teunissen, H., Muir, K.W., de Wit, E., Rowland, B.D., *et al.* (2020). The structural basis for cohesin-CTCF-anchored loops. Nature *578*, 472–476.
Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.
Lu, Z., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X., and Schmitz, R.J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants *5*, 1250-1259.
Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., *et al.* (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.
Marchal, C., Sima, J., and Gilbert, D.M. (2019). Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol *20*, 721-737.
Mateo, L.J., Murphy, S.E., Hafner, A., Cinquini, I.S., Walker, C.A., and Boettiger, A.N. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature *568*, 49–54.
Matharu, N., and Ahituv, N. (2015). Minor loops in major Ffolds: enhancer-promoter Llooping, chromatin restructuring, and their association with transcriptional regulation and disease. PLoS Genet *11*, e1005640.
Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.M., Taneja, N., Folco, H.D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., *et al.* (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in *S. pombe*. Nature *516*, 432-435.
Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods *13*, 919–922.
Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59-64.
Ngan, C.Y., Wong, C.H., Tjong, H., Wang, W., Goldfeder, R.L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., *et al.* (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat Genet *52*, 264–272.
Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell *169*, 930-944.e922.
Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., *et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.
Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O'Shea, C.C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science *357*.
Pang, B., and Snyder, M.P. (2020). Systematic identification of silencers in human cells. Nat Genet *52*, 254–263.
Pauli , T., Vedder, L., Dowling, D., Petersen, M., Meusemann, K., Donath, A., Peters, R.S., Podsiadlowski, L., Mayer, C., Liu, S., *et al.* (2016). Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects : Insect insulator proteins. BMC Genomics *17*, 861.
Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., *et al.* (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.
Plys, A.J., Davis, C.P., Kim, J., Rizki, G., Keenen, M.M., Marr, S.K., and Kingston, R.E. (2019). Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. Genes Dev *33*, 799–813.
Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., *et al.* (2018). Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell *174*, 744-757.e724.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., *et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.
Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., *et al.* (2017). Cohesin loss eliminates all loop domains Cell *171*, 305-320.e324.
Ricci, M.A., Manzo, C., García-Parajo, M.F., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes *in vivo*. Cell *160*, 1145.
Ricci, W.A., Lu, Z., Ji, L., Marand, A.P., Ethridge, C.L., Murphy, N.G., Noshay, J.M., Galli, M., Mejía-Guerra, M.K., Colomé-Tatché, M., *et al.* (2019). Widespread long-range cis-regulatory elements in the maize genome. Nat Plants *5*, 1237-1249.
Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y., and Corces, V.G. (2017). Evolutionarily conserved principles predict 3D chromatin organization. Mol Cell *67*, 837-852 e837.
Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., *et al.* (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.
Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet *20*, 437-455.
Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504-507.
Schwaiger, M., Schonauer, A., Rendeiro, A.F., Pribitzer, C., Schauer, A., Gilles, A.F., Schinko, J.B., Renfer, E., Fredman, D., and Technau, U. (2014). Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome Res *24*, 639-650.
Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., *et al.* (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.
Sebe-Pedros, A., Ballare, C., Parra-Acero, H., Chiva, C., Tena, J.J., Sabido, E., Gomez-Skarmeta, J.L., Di Croce, L., and Ruiz-Trillo, I. (2016). The dynamic regulatory genome of capsaspora and the origin of animal multicellularity. Cell *165*, 1224-1237.
Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K.N., Holcomb, N.P., Turner, J.L., Paulsen, M.T., Rivera-Mulia, J.C., Trevilla-Garcia, C., *et al.* (2019). Identifying cis elements for spatiotemporal control of mammalian DNA replication. Cell *176*, 816-830.e818.
Smol, T., Sigé, J., Thuillier, C., Frénois, F., Brunelle, P., Rama, M., Roche-Lestienne, C., Manouvrier-Hanu, S., Petit, F., and Ghoumid, J. (2020). Lessons from the analysis of TAD boundary deletions in normal population. bioRxiv 20200401021188;.
Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat Rev Genet *19*, 453–467.
Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.
Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. Sci Adv.
Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., *et al.* (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell *163*, 1611–1627.
Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H., and Ren, X. (2019). Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. J Biol Chem *294*, 1451–1463.
van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A., *et al.* (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods *9*, 969–972.
van Steensel, B., and Belmont, A.S. (2017). Lamina-Associated Domains: Links with chromosome architecture, heterochromatin, and gene repression. Cell *169*, 780–791.
van Steensel, B., and Furlong, E.E.M. (2019). The role of transcription in shaping the spatial organization of the genome. Nat Rev Mol Cell Biol *20*, 327-337.
Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.R., Jung, S., Baranello, L., Huang, S.C., El Khattabi, L., Dose, M., Pruett, N., *et al.* (2018). The energetics and physiological impact of cohesin extrusion. Cell *175*, 292–294
Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., and Zhang, X. (2018a). Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat Plants *4*, 90-97.

Wang, Q., Sun, Q., Czajkowsky, D.M., and Shao, Z. (2018b). Sub-kb Hi-C in *D. melanogaster* reveals conserved characteristics of TADs between insect and mammalian cells. Nat Commun *9*, 188.
Wang, X., Brandao, H.B., Le, T.B., Laub, M.T., and Rudner, D.Z. (2017). *Bacillus subtilis* SMC complexes juxtapose chromosome arms as they travel from origin to terminus. Science *355*, 524-527.
Williamson, I., Kane, L., Devenney, P.S., Flyamer, I.M., Anderson, E., Kilanowski, F., Hill, R.E., Bickmore, W.A., and Lettice, L.A. (2019). Developmentally regulated Shh expression is robust to TAD perturbations. Development *146*, pii: dev179523.
Winter, D.J., Ganley, A.R.D., Young, C.A., Liachko, I., Schardl, C.L., Dupont, P.Y., Berry, D., Ram, A., Scott, B., and Cox, M.P. (2018). Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*. PLoS Genet *14*, e100746.
Yan, J., Chen, S.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z., *et al.* (2018). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. Cell Res *28*, 204–220.
Zhang, H., Emerson, D.J., Gilgenast, T.G., Titus, K.R., Lan, Y., Huang, P., Zhang, D., Wang, H., Keller, C.A., Giardine, B., *et al.* (2019). Chromatin structure dynamics during the mitosis-to-G1 phase transition. Nature *576*, 158–162.

3.4 THE NON-CODING GENOME

ABSTRACT

The non-coding genome contributes to establish a sophisticated regulatory network that it is essential for cellular function, with a direct impact on the growth, development, evolution and health of organisms. Most of its structural features, interacting factors and mechanisms of action are currently unknown. Deciphering these aspects together with the consequences that their function might have at multiple levels are important challenges for future investigations.

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS (in alphabetical order)

- Carlos Estella (CBMSO, Madrid)
- Sonia García (IBB, Barcelona)
- Crisanto Gutiérrez (CBMSO, Madrid, *Coordinator*)
- Cristina Hernández López de Munain (IPBLN, Granada, *Deputy Coordinator*)
- José Carlos Reyes (CABIMER, Sevilla)
- Esther Serrano (CBMSO, Madrid)

Executive Summary

Studying the function of every gene encoding proteins in the genome has concentrated, and will continue to concentrate, many efforts to understand how cells and organisms function. However, the coding genome constitutes a very minor fraction of the whole genome, in particular in the genome of multicellular organisms. The presence of very large amounts of non-coding genomic DNA (ncGENOME) is a feature of the genome of a wide diversity of species, including all evolutionary lineages analyzed, covering animals, plants, and yeasts, and being also present in prokaryotes and even in viruses with large genomes. The so far generally accepted idea that this huge fraction of the eukaryotic genomes consists on non-coding DNA has dramatically changed in recent years, transforming the concept of the "dark matter" genome into one of the stars of modern biology. This conceptual change is largely based on the enormous diversity of elements and products derived from the ncGENOME, and the identification of its crucial role in both normal and disease conditions, as reported in a handful of individual examples. Together, the different elements and products of the ncGENOME contribute to establish a sophisticated regulatory network that is essential for normal function of the coding genome. Most of the structural features, interacting factors, mechanisms of action, targets, and loop interactions of the ncGENOME are virtually unknown. Deciphering these various aspects of ncGENOME together with the identification of the consequences that their function might have at multiple levels are important challenges for future investigations. A non-exhaustive list of potential roles for the ncGENOME includes transcription, messenger RNA splicing and processing, transport, translation and decay, RNA interference, imprinting, epigenetic modifications, chromatin remodeling, assembly of subnuclear organelles by liquid-liquid phase separation processes, and three dimensional nuclear organization. One key aspect of the function of various components of the ncGENOME is that its derived non-coding RNAs associate with a plethora of proteins, other RNA molecules, and DNA in order to achieve their function. Therefore, studying the structural features of both the RNA moiety and its interacting proteins is of primary relevance to understand the function of the ncGENOME.

In summary, the ncGENOME is far larger than the coding genome and we know very little about it. Why is crucial to tackle the enormous challenge of understanding its role in the years to come? The answer to this question is relatively simple: the ncGENOME has a direct impact on the growth and development, and evolution of all multicellular organisms, including animals and plants, as well as in their health. Based on the current state of the art in this field, we outline a number of questions arising that we think should constitute the focus in the future.

Introduction and general description

The presence of very large amounts of non-coding genomic DNA (ncGENOME) is a feature of the genome that it is present in all analyzed organisms. Within this term we include the DNA and derived non-coding RNAs (ncRNAs). The importance of these transcripts was evidenced in 2012 by the ENCODE project that established that 75% of the human genome is transcribed, but only 2% of the transcripts are translated into proteins (ENCODE Consortium, 2012). The observation of these abundant transcripts has been confirmed in many different organisms.

The ncGENOME is composed of diverse elements that can be classified in three major groups: sequences that result in transcripts with a housekeeping function (hkGENOME), repeated sequences (repGENOME), and regions that either contain regulatory sequences or lead to the genesis of regulatory transcripts (regGENOME). The *hkGENOME* transcription results in the generation of transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs) with constitutive functions. The rep*GENOME* includes not only telomeric and centromeric regions, but also all different classes of repetitive DNA that together constitute a large fraction of the ncGENOME. The repGENOME sequences can be transcribed, as it is the case of telomeric repeat-containing RNAs (TERRAs), involved in protecting telomeres. It also includes repeats that can be tandemly arranged or dispersed, whose main constituents are the transposable elements (TEs), known not only to lead to chromosomal rearrangements but also to modulate the expression of nearby genes and of themselves through the generation of piwi RNAs (piRNAs). The reg*GENOME* is made up of a very diverse group of genomic sequences. These include the *cis*-acting elements that conform proximal and distal sequences that regulate gene expression at the transcription level, such as promoters, silencers, and enhancers that contain motifs to mediate binding of cellular factors. The regGENOME also includes the intronic and intergenic regions that lead to the generation of ncRNAs of various sizes and configurations, such as long ncRNAs (lncRNAs), micro RNAs (miRNAs), and enhancer RNAs (eRNAs), that work in *cis* or *trans* with diverse regulatory functions. Finally, there are sense and antisense linear transcripts and circular ncRNAs (circRNAs) derived from exonic and intronic regions that also play a regulatory role. Because circRNAs are, at least in part, a consequence of transcription of exons, we have classified them as transcripts derived from the coding genome, rather than from the ncGENOME. The elements that comprise the hkGENOME, repGENOME, and regGENOME are inter-regulated with each other and also crosstalk functionally and structurally with the coding genome and RNA/DNA-binding proteins in order to regulate development, growth, disease, and evolution.

During embryonic development, numerous processes need to be finely orchestrated in space and time through regulation of gene transcription. Achieving this requires integrating the regulatory input from many elements of the ncGENOME, an encrypted language of enormous complexity. Understanding how the ncGENOME can affect growth and welfare, and hence productivity of plants and animals used by humans, while maintaining a sustainable productivity, is crucial. Mutations or epigenetic alterations that affect the ncGENOME have great impact in the growth of organisms, including disease conditions. Genome-Wide Association Studies (GWAS) have evidenced that most (70–90%) of disease-associated genetic variation (single nucleotide polymorphisms, SNPs), and also in animals and plants, lie outside of gene bodies, in the ncGENOME. However, the vast majority of these SNPs are not functionally characterized. Answering to these and related questions represents a major challenge with important scientific, social, and economic implications, especially considering the constantly increasing human population that needs to be fed and maintained with good health.

Impact in basic science panorama and potential applications

I. hkGENOME

The hkGENOME generates transcripts with constitutive functions that play essential functions in a broad range of biological processes and cause multiple pathologies.

tRNAs

tRNAs are basic components of the translation machinery, the decoders of mRNAs in protein translation (Tuorto and Parlato 2019). They are relatively small compared to other RNA species (70-90 nt), with a cloverleaf tri-dimensional (3D) structure, an evolutionary innovation that allowed the standardization of the three-nucleotide genetic code. tRNAs play also fundamental roles in other cellular functions, e.g., modulation of cell proliferation and stress response (Thompson et al., 2008), as well as in gene expression regulation (Raina and Ibba, 2014). In particular, the activity of specific tRNA fragments (cleaved tRNA) is responsible of still mostly unknown functions. These fragments, previously considered non-functional degradation intermediates, are now recognized as major RNA species for which their regulatory roles are just starting to be understood, including stress response, apoptosis and cancer (Hanada et al., 2013). Other tRNAs are involved in additional biochemical processes, such as cell wall formation, protein labelling for degradation and antibiotic biosynthesis (Kanai, 2014).

rRNAs and rRNA genes (rDNAs)

rRNA are integral part of the ribosomes with structural and catalytic functions in protein synthesis. Most RNA in a cell (ca. 80%) is rRNA, and there are several different rRNA molecules, three in prokaryotes (5S, 16S, and 23S) and four in eukaryotes (5S, 5.8S, 18S, and 28/26S). The eukaryotic 5S rRNA is the most enigmatic and its functions are not yet resolved; interestingly, while the eukaryotic 5.8S corresponds to the prokaryotic 5S, no prokaryotic molecule corresponds to the 5S in eukaryotes, and it is also produced in a different location (the nucleolus) (Ha and Bhagavan, 2011). As tRNA, rRNA also mediates cellular stress conditions, and the synthesis of both RNA species is interdependent.

snRNAs: snoRNAs, scaRNAs and U-snRNAs

snRNAs are short transcripts (60-300 nt long) present in all eukaryotic organisms and its number increases with organism complexity. In addition, genomic organization of the ncRNAs-encoding genes follows an evolutionary tree. In *Sacharomyces cerevisiae* they are organized in independent genes, whereas in humans they are predominantly located within introns, being released by splicing. snRNAs have important functional and structural functions, being decisive in the formation of subnuclear organelles known as nuclear bodies. These membrane-less compartments are formed by local high concentrations of molecules that promote formation of weak non-covalent bonds, most probably driven by liquid-liquid phase separation (Khosraviani et al., 2019). These dynamic structures include nucleoli, speckles, paraspeckles, promyelocytic leukaemia bodies, Cajal bodies (CBs), histone locus bodies (HLBs), and polycomb bodies. Nuclear bodies occupy the interchromatin space and are highly enriched with specific nuclear factors and, in many cases, with structural ncRNA. Other ncRNAs have been structurally implicated in their formation, such as rRNAs in forming the nucleolus and the structural lncRNA in forming paraspeckles.

snoRNAs and scaRNAs are assembled with proteins in small-nuclear ribonucleoproteins (snRNPs), responsible for RNA post-transcriptional modifications, such as pseudouridylation and 2´-O-methylation of other RNAs. RNPs formed with snoRNAs (snoRNPs) are responsible for the modification of rRNAs in the nucleoli, whereas those formed with scaRNAs (scaRNPs) mediate the modification of U-snRNAs in speckles and CBs. U-snRNAs also assemble into RNPs (U-snRNPs) that function in pre-mRNAs processing. Recent data indicate that snoRNAs and scaRNAs have overlapping specificities and localization in the nucleus, suggesting that nucleoli and CBs may have interchangeable functions (Deryusheva and Gall, 2019).

In addition to their role in RNA modification and maturation, nuclear bodies are involved in guiding chromosome folding to provide a platform for the spatial organization of genomic

loci, affecting their expression. Thus, actively transcribed genes are associated with the periphery of speckles, the HLBs participate in the 3D organization of the histone gene clusters, and heterochromatin domains are mostly located near the nuclear periphery or the nucleolus (Khrosraviani et al., 2019). A provocative hypothesis argues that the clustered organization of genes within chromosomes and the capacity of ncRNAs to nucleate nuclear bodies have evolved together to facilitate RNA-driven assembly of nuclear hubs (Smith et al., 2020). Hence, snRNAs have the potential to modulate subnuclear structures with an unexplored role in shaping the genome to promote proper gene expression.

II. repGENOME

Eukaryotic genomes are largely composed of repetitive DNA sequences that constitute the repGENOME. These sequences can be classified as tandem or interspersed repeats. Tandem repeats include centromeres and telomeres (composed by satellite DNA), whereas interspersed repeats include TEs. Repeats are extremely variable both in abundance and sequence length, occurring in up to millions of copies per genome, ranging in size from a few base pairs to many thousands (Heslop-Harrison and Schwarzacher, 2011). These regions generate transcripts with important functions, such as those generated from telomers, known as TERRAs, that are essential for telomere protection, and the piRNAs, generated from TEs, that are involved in the regulation of their own transcription.

Telomeres and TERRAs

Telomeres are nucleoprotein complexes at the end of chromosomes that protect them from damaging during the replication process, and maintain chromosome homeostasis and aging. Their repetitive portion consists on a minisatellite of 6-8 nt, which sequence is highly conserved across large groups of organisms. Telomere length is species-specific, but it can be modulated through a balance between shortening and elongating signals during organism's lifetime. The telomerase, the enzyme responsible of elongating telomeres, and TERRAs, which regulate telomerase activity and maintain the heterochromatic state at chromosome ends, together with the telomeres conform the inter-dependent triad of the "telomere world" (Mensà et al., 2019).

Centromeres

Centromeres and pericentromeric regions, which are species-specific, are composed of satellite repeats. Centromeres organize chromosome movements from prophase to anaphase by interacting with microtubules of the spindle apparatus promoting faithful chromosome segregation. The composition of centromeric regions is known for very few taxa, mostly model

organisms (Hirsch and Jiang, 2013). Centromeres are usually accompanied by TEs in pericentromeric regions and the interaction between both genomic components is only starting to be understood. The combination of next generation sequencing (NGS) with chromatin immunoprecipitation (ChIP-seq) makes possible to analyse centromere composition in most species.

TEs and piRNAs

The most important fraction of repetitive DNA is composed by dispersed repeats, mostly TEs. Formerly known as jumping genes, TEs can change the DNA landscape and be source of genetic innovation by altering gene expression or promoting chromosomal rearrangements (Pantzartzi et al., 2018). TEs occupy most of eukaryotic genomes, e.g., nearly 50% of the human genome or up to 85% of maize. Most TEs in the genome have lost its mobility, being the remaining of previously active mobile elements, >200 millions of years ago, and random mutations have made them unrecognizable. The intrinsic characteristics of TEs, with its repetitive but also non-conserved sequence features, have made them a difficult subject of study by traditional approaches. Again, the advent of NGS has made possible their analysis. There are two large categories, the DNA-TEs, which use a cut-and-paste transposition mechanism, and retroelements, the most abundant in most eukaryotic genomes which move and amplify through an RNA-mediated copy-and-paste (retro)transposition. In the human genome, Alu sequences, a type of short interspersed nuclear elements (SINEs), classified as retroelements, are the most abundant. Another important fraction of the human genome (around 8%) is made up of endogenous retroviruses (ERVs), a type of retroelements that may have originated in ancient retrovirus insertions. Although historically considered as “selfish genetic elements”, TE contribute to a wide range of regulatory functions. Besides, recent findings support a relevant role of TEs in processes related to speciation, but also in health and disease (Reilly et al., 2013; Serrato-Capuchina and Matute, 2018).

piRNA a novel class of non-coding RNAs (belonging to the group of interference RNA) have recently found to have a role in the epigenetic and post-transcriptional silencing of transposable elements, usually preventing their expansion. However, the wide variation of piRNA sequences across species makes difficult to establish their functionality.

III. regGENOME

The regGENOME include not only the *cis*-acting regulatory elements and the boundaries of the topological associated domains (TADs)/insulators in chromatin, but also ncRNAs of diverse size

and configurations that are originated from intronic and exonic regions. The importance of TADs and their boundaries in 3D genome organization and gene expression controls have been extensively described in Chapter 3.3, so here we will focus on *cis*-regulatory elements and ncRNAs.

Cis-acting regulatory regions

Enhancers and promoters are sequences that have the ability to activate gene transcription in *cis*. While promoters act at a short distance from the transcriptional start site, enhancers and silencers do so at long distances, from a few hundreds of base pairs to several hundreds of kilobases, in an orientation-independent manner (Parker et al., 2013; Hnisz et al., 2013). The organization of chromatin in different TADs and sub-TADs with specific regulatory landscapes depends on the interactions between these *cis*-regulatory regions through the recruitment of transcription factors (TFs) (Franke and Gómez-Skarmeta, 2018; Schoenfelder and Fraser, 2019).

Enhancers have deserved high attention due to their essential role in the control of developmental programs. Enhancers contain numerous DNA binding sites for TFs, specific for different cell types or developmental stages (Heinz et al., 2015). TF binding serves to recruit co-activator complexes, such as Mediator complex or histone modifying enzymes, that together determine enhancer activity. In some instances, dense clusters of enhancers are brought together into close 3D proximity and collaborate to act as a single regulatory unit (super-enhancers) that can drive high levels of transcription (Sabari et al., 2018; Nair et al., 2019).

eRNAs correspond to a surprising discovery made in 2010, which correlated the presence of these enhancer-derived transcripts to that of mRNAs of target genes (De Santa et al., 2010; Kim et al., 2010). eRNAs are diverse with heterogeneous structure, length and post-transcriptional modifications, and exhibit tissue and lineage specificity. Although they were originally considered as a possible transcription by-product, growing evidence supports that they constitute functional biomolecules that promote gene transcription by enhancer-promoter looping and chromatin modifying (Arnold et al., 2020). However, these transcripts are typically highly unstable and lack strong sequence conservation, suggesting that some eRNAs might have rather subtle functions or even represent transcription by-products. Transcription of super-enhancers results in high quantities of eRNAs that interact with complexes, such as Mediator complex, cohesin, p300/CBP, and BRD4, to regulate transcription both locally and distantly.

ncRNAs

ncRNAs, intergenic and intronic/exonic transcripts that are not translated, have emerged as major players of the regGENOME during evolution, development and disease. Although the term

ncRNAs include also the transcripts generated from the hkGENOME and repGENOME, the description in this section is limited to the regGENOME-derived ncRNAs. These ncRNAs are involved in numerous cellular processes including transcription, splicing, and protein translation. These transcripts can regulate gene expression in *cis* (the genetic locus from where it has been generated) by binding to DNA, or in *trans* (elsewhere in the genome), as structural components of nuclear organelles or functioning as molecular decoys to titrate proteins and other RNAs. They can be classified based on their size in lncRNAs with >200 bp and short (sncRNAs) with <200 bp, and based on their configuration in linear or circRNAs.

sncRNAs include miRNAs with ~22 nucleotides. miRNA biogenesis and maturation pathways, as well as their mode of action have been well defined (Pu et al., 2019). These short transcripts act as post-transcriptional regulators of gene expression via direct binding to mRNAs, regulating diverse biological processes across all tissues and cell stages.

circRNAs are formed by covalently closed loops through back-splicing and can be exonic, intronic, and exon-intron. circRNAs derive from the coding genome, but are considered ncRNAs because they are not translated, although some can generate small peptides (Pamadurti et al., 2017). Similarly to mRNAs, circRNAs are abundant and well-conserved transcripts expressed in a tissue-, cell stage- and temporal-regulated fashion, and are altered in various diseases (Wang et al., 2017). circRNAs are involved in several biological activities by sponging miRNAs and RNA binding proteins. Interestingly, they are more stable than linear RNAs, having an average half-life 5 times longer than that of mRNAs (Jeyaraman et al., 2020). Based on these features, they have been proposed as interesting new biomarkers and therapeutic agents for cancer.

Little is known about the biology of lncRNAs, which display a high level of diversity and are originated from a significant part of the ncGENOME. In contrast with the 2,000 different miRNAs identified in humans, approximately 50,000 different lncRNAs have been detected, representing the largest part of the transcriptome in animals and plants, and are observed in practically all species, including yeast, prokaryotes, and even viruses (Alessio et al., 2020). These transcripts are transcribed by the RNA polymerase II, being normally capped and polyadenylated, and might be or not processed by the splicing machinery (Quinn and Chang, 2017). Based on their modes of action, they can be classified in *cis* or *trans*: An example of a *cis* mode of action includes when a lncRNA functions as a bridge between DNA and protein, serving as a scaffold to bring histone modifying complexes to specific loci, as it is the case for eRNAs. An example of a *trans* mode of action includes a lncRNA acting as molecular decoy to titrate proteins or miRNAs (Grüll and Massé, 2019). However, because their low abundance, there is a debate over whether they can efficiently deplete miRNAs (Ulitsky, 2018). In fact, the

existence of a network of interactions between lncRNAs and miRNAs is currently a major theme of research.

The role that lncRNAs plays in a many biological processes, such as transcription, imprinting, splicing, and translation with functional consequences in cell cycle, apoptosis, pluripotency and reprograming, heat-shock response, and disease, reflects the versatility of RNA itself, being able to fold into a variety of secondary structures and to bind to a large number of substrates. The lncRNA interactome includes proteins, other RNAs, and DNA. lncRNAs display very poor conservation compared with other RNAs (Necsulea and Kaessmann, 2014). Although they do not conserve their primary sequence, their function is maintained in different species indicating that their conservation is based on structural traits rather than on sequence traits. Therefore, their secondary/tertiary structure plays a pivotal role for lncRNA function. Interestingly, recent studies indicate that about 23% of the lncRNA interactome is composed by short peptides that are encoded within the lncRNAs (Matsumoto and Nakayama, 2018). Despite the current excitement about lncRNAs and their potential biological functions, it is worth mentioning that when their roles have been evaluated *in vivo* by individually deleting a selected group of 24 out of 727 of these transcripts in zebra fish corresponding to the 3,3% of the total lncRNAs present in this organism, it did not result in any obvious phenotypic defects (Goudarzi et al, 2019). Although this study suggests that lncRNAs might have none or rather subtle functional roles. However, other studies have demonstrated that mis-regulation or elimination of lncRNAs do have significant phenotypic consequences such as COOLAIR on flowering time in plants (Csorba et al., 2014), and sense/antisense transcripts originated at the antigen receptor genes on the control of V(D)J recombination (Abarrategui and Krangel, 2006; Giallorakis et al., 2010) or *Xist* on the regulation of dosage compensation of X chromosomes between males and females (Loda and Heard, 2019) in animals. Therefore, more systematically deletion analyses should be performed, including the simultaneous deletions of more than one transcript to rule out compensation effects among different transcripts, to clearly establish their *in vivo* function in this and other model organisms.

IV. The ncGENOME in development

All cell types of a multicellular organism share a nearly identical DNA sequence but perform very different functions during development and adulthood. This vast cell type diversity is accomplished by the precise regulation of gene expression in time and space by the regGENOME through chromatin compartmentalization, *cis*-regulatory elements, and post-

transcriptional mechanisms. Mutations affecting the regGENOME are a common source of phenotypic divergence and evolutionary change, and frequent cause of human disease.

Deciphering the spatial and temporal transcriptional code during development

DNA accessibility in chromatin is dynamic during development and different between cell types (Klemm et al., 2019). Active enhancers are often devoid of nucleosomes, with pioneer TFs being important for the opening of closed chromatin sites and the subsequent binding of additional TFs (Iwafuchi-Doi et al., 2016). Enhancer activity correlates with certain chromatin properties, e.g., nucleosome depletion and post-translational modifications, primarily H3K4me1 and H3K27ac (Yáñez-Cuna et al., 2013; Shlyueva et al., 2014). In addition, multiple enhancers and/or distinct TFs are employed by a single developmental gene to precisely activate its expression in different cells, at variable levels and at multiple times during development. In addition, it has been shown that eRNAs could regulate chromatin status, TF binding dynamics or chromatin loop stabilization providing another layer of transcriptional regulation (Lewis et al., 2020). Since the finding of enhancer function in early development (Martinez-Salas et al., 1989), understanding enhancer biology is currently an area of great interest, as there is an increasing appreciation of their importance in development, evolution and disease (Corradin and Scacheri, 2014).

Post-transcriptional mechanism in gene regulation through development

Post-transcriptional regulators such as lncRNAs, miRNAs, and circRNAs are key players in gene expression regulation during animal and plant development including almost every step of development from the downregulation of pluripotency of embryonic stem cells to the differentiation process, sex-determination, cell identity acquisition and maintenance, control of developmental time and transgenerational epigenetic inheritance and genomic imprinting. ncRNAs act in hierarchical networks; some regulate the initial patterning events, and its loss or misregulation generate lethal or strong phenotypes, whereas others perform subtler, but essential, functions (Davidson et al., 2003; Alberti and Cochella, 2017). ncRNAs are represented in all the positions of the hierarchy but their low levels of expression and redundant functions have hampered their study. The role of miRNAs as regulators of developmental time in *C. elegans* and the mechanisms involved in mammalian puberty show striking similarities, opening the question of how prevalent are miRNAs elements in the control of developmental time. After the discovery of *Xist* and *HOTAIR*, lncRNA are considered to preferentially function in specific cellular contexts, cell types, developmental stages and diseases. Their study has produced invaluable knowledge of the mechanisms of lncRNAs action in directing chromatin modifiers to *cis* and

trans. However, while the studies of lncRNA are still at its infantry, it has become a paradigm for understanding long-range mechanisms of gene regulation and dose compensation.

ncRNAs have become the modern candidates to reconcile the notion of adaptive epigenetic inherence (Heard and Martienssen, 2014). Transgenerational epigenetic inheritance, that is the transmission of epigenetic features from one generation to the next through the germline, persisting in subsequent generations, is well documented in numerous organisms, including plants, nematodes, fruit flies, and mammals (Tyebji et al., 2020). A major obstacle to transgenerational epigenetic inheritance is germline reprograming, whereby DNA methylation, histone variants and their modifications are reset (Tang et al., 2015). RNA molecules are excellent candidates to carry epigenetic information across generations due to their specificity and long life, less affected by reprogramming, although the mechanisms are largely unsolved. Small RNA signals are highly mobile and mediate heritable transcriptional silencing through generations as demonstrated in the germline of *C. elegans*. Likewise, small RNAs can also travel through vasculature and plasmodesmata in plants and through exosomes and even serum in mammals (Chen et al., 2016). Both core and species-specific components of this process continue to be discovered.

Genomic imprinting is a complex and highly regulated process, which consist of the monoallelic silencing of certain genes. lncRNAs regulate chromatin structure and gene expression of imprinted genes through interactions with histone modifying proteins, looping and by promoting intra-chromosomal chromatin compartmentalization (Kanduri, 2016). DNA methylation profiling revealed that, although the bulk of the genome (including imprinted loci) becomes demethylated in primordial germ cells, a number of loci, predominately associated with repetitive sequences, escape demethylation. The reason is still unclear, but they could represent prime candidates for possible transgenerational inheritance in mammals. More details about this interesting, yet poorly understood, type of inheritance and its implications for human health and disease can be found in Chapter 3.7.

V. Pathological relevance of the ncGENOME

The ncGENOME has a deep impact in disease. GWAS have evidenced that most (70–90%) of disease-associated SNPs lie in the ncGENOME (ENCODE Consortium, 2012; Maurano et al., 2012). While some of these SNPs are determinants for some disease development, other play a combinatorial contribution to them. In addition to inherited genetic variations, somatic mutations in the ncGENOME are also responsible for diseases, such as cancer. Association studies based on epigenetic traits (EWAS), mostly DNA methylation, have also started to be analyzed. Also in

this case, and even more than in the GWAS studies, most epigenetic variations are concentrated in regions that reside in the ncGENOME (Rakyan et al., 2011). Some examples are provided to illustrate how different elements of the ncGENOME are involved in disease.

rRNAs, tRNAs, and snRNAs

The dysregulation in rRNA synthesis can lead to disorders, including Alzheimer's or other neurodegenerative diseases (Tuorto and Parlato, 2019). Even more, cancer, premature aging, and neurological impairment in ataxia-telangiectasia and Bloom syndrome, among others, relate to increased cellular rDNA instability but its role is still unknown (Warmerdam and Wolthuis, 2019). Notably, the rDNA array has usually defied sequencing and assembly technologies due to its repeat nature, and, as a consequence, this genome portion remains a missing area in genome assemblies, even in well studied model organisms (Wang and Lemos, 2019).

The disruption in the maturation steps of any tRNA or rRNA and their related enzymes can impact cell homeostasis at multiple levels and be the cause of different disorders, such as cancer, infections, many neurodegenerative diseases and other pathological conditions (Tuorto and Parlato, 2019). Beyond understanding the role of tRNAs in human disease, in some cases these tRNA fragments may serve as useful biomarkers (Anderson and Ivanov, 2014).

Aberrant expression of snRNAs has been associated with cancer, cardiovascular diseases, and several neurological and neuromuscular disorders. Consequently, these ncRNAs have been proposed as attractive therapeutic agents able to activate or inhibit mRNA splicing and new biomarkers, due to their specific expression in particular tissues and stable circulation in biological fluids (Isakova and Quake, 2018).

Telomeres and Centromeres

Telomere shortening is strongly related to age-associated diseases. Mutations in telomere maintenance genes cause telomere erosion leading to different telomere dysfunction syndromes, such as Hoyeraal-Hreidarsson syndrome, dyskeratosis congenita, and aplastic anemia, all of them characterized by premature aging (Martínez and Blasco, 2017). Telomere shortening also provokes chromosomal rearrangements that contribute to tumor initiation or progression. Furthermore, ectopic expression of telomerase is commonly associated to malignant cell transformation.

Centromeres, probably due to the high density of repetitive sequences, are fragile regions prone to chromosome breakage and rearrangements (Barra and Fachinetti, 2018). Therefore, they are a potential source of genome instability that can be linked with human diseases. In fact, certain types of tumors are characterized by a high frequency of chromosome rearrangements

involving the pericentromeric region. In addition, centromeric alterations are found in human congenital diseases such as particular immunodeficiencies, centromeric region instability, and the facial anomalies syndrome. In this syndrome, mutations in *DNMT3B, ZBTB24, CDCA7, and HELLS* genes cause loss of DNA methylation and alteration in the heterochromatin structure of centromeric regions, leading to pericentromeric breaks and chromosomal rearrangements. It is still poorly understood how, at which point and to what extent centromere dysfunctions may participate in these pathologies, and new models and technologies are needed. In particular, centromere research was limited by the absence of sensitive sequencing technologies for repetitive DNA, thus new sequencing and bioinformatics developments will be essential to reveal the processes producing instability in these regions in human disease.

Enhancers

The human genome contains more than 1,000,000 of poorly characterized enhancers, which constitute an important part of the ncGENOME, and a large fraction of all SNPs associated with human diseases lie within them (Maurano et al., 2012). Diseases linked to enhancers are termed enhanceropathies, and include polydactyly, caused by mutations in one of the enhancers of the SHH gene or some □-thalassemias caused by translocations in the globin Locus Control Regions, a well-known super-enhancer (Smith and Shilatifard, 2014). Mutation in genes encoding proteins that regulate enhancer function, are also considered enhanceropathies. One example is Cornelia de Lange, a syndrome caused by mutations in genes encoding NIPBL, other cohesin complex subunits and the epigenetic reader BRD4. Translocations and chromosomal rearrangements also cause enhanceropathies by changing the 3D position of regulatory regions relative to their target genes causing new target genes to be ectopically expressed. Thus, the Burkitt´s lymphoma is caused by the translocation of the *IGH* enhancer in proximity to the *MYC* gene (Gillies et al., 1983). Epigenetic reprograming of enhancers are also involved in cancer predisposition (Aran and Hellman, 2013) and metastasis (Roe et al., 2017). These are only a few examples, however, most of the diseases and susceptibilities caused by enhancer alterations are uncharacterized and their study constitutes one of the major challenges of human genetics.

miRNAs

The role of miRNAs in cancer has attracted special attention. miRNA expression profiles are tumor-specific and differ between different types of tumors. In fact, miRNA profiling is used for tumor stratification and for prognosis prediction. Specific miRNAs act as oncogenes (onco-miRNAs), tumor suppressors, or both, in a context-dependent manner (Di Leva et al., 2014). Furthermore, early cancer detection in liquid biopsies by determination of circulating miRNAs is

a promising strategy (Toiyama et al., 2017), which may eventually be used also for other diseases.

lncRNAs

lncRNAs play important regulatory roles in almost every signalling pathway affecting cell proliferation and differentiation and, consequently, they are directly involved in malignant transformation (Huarte, 2015). For example, overexpression of the *HOTAIR* lncRNA causes metastasis in breast cancer cells and promotes silencing of the *HOXD* cluster among other genes, causing dedifferentiation and increasing cancer cell invasiveness and metastasis. Another example includes the antisense lncRNA *CDKN2B-AS1*, whose overexpression causes silencing of the INK tumor suppressor locus. lncRNAs can also play tumor suppressor roles regulating negatively the expression of oncogenes, as shown for *PVT1*, which control the expression of the contiguous *MYC* oncogene (Cho et al., 2018). In addition to cancer, lncRNAs are also associated to other diseases such as cardiovascular (Liu et al., 2014), neurodevelopmental (Ang et al., 2019), and celiac disease (Castellanos-Rubio et al., 2016). The number of lncRNA involved in disease is growing quickly and in the next few years probably thousands of then will be related to different pathologies. However, due to the concerns raised by a recent report (Goudarzi et al, 2019), it will also be important to validate their functional and pathological relevance *in vivo* and using appropriate and orthogonal methods.

Key challenging points

The sheer abundance, diversity, and our still shallow understanding of ncGENOME promote many challenging questions to be answered in future research that can be organized in four different groups. This list of questions to be answered is not exhaustive but should include the following:

Basic knowledge of the ncGENOME function and structure

- Why are there so many types of elements in the ncGENOME?
- What is the full map of interactions between the ncGENOME and their targets in animals and plants?
- How do ncRNA structures impact on liquid-liquid phase separation and what trigger organization of such structures?
- Do eRNAs drive enhancer-promoter looping and trigger gene transcription through a process mediated by liquid-liquid phase separation?
- How do ncRNAs modulate subnuclear structure and 3D genome organization?

- How do epigenetic modifications affect ncRNA function?
- How is organized the combinatorial regulatory interplay among different components of the ncGENOME?

Other challenges related with enhancer function include the identification of the molecular determinants defining not only the more complex enhancers but also the minimal set of elements conferring enhancer activity, as well as understanding of several enhancer features, including the function of the presence of specific post-translational modifications of histones, the binding of a variety of proteins including CTCF and cohesins, and eRNA function.

Other challenges related with ncRNAs include to clearly understand how lncRNAs exert their functions and how much of them depend on their structural features, their interaction with other molecules, or even their intranuclear localization in the 3D nucleus. Much effort should also be directed towards determining secondary and tertiary structures of lncRNAs to identify the crucial motifs for function. Another important line for future research will include the determination of the binding partners of these transcripts. Last but not the least, the study of epigenetic modifications in the context of ncRNA may uncover unknown mechanisms and new layers of complexity in regulatory networks that may impact health and disease.

Relevance of the ncGENOME in development and disease

- Why eukaryote genomes are so rich in non-coding DNA?
- What is the relevance of different ncGENOME elements in development, organogenesis, regeneration, senescence, and disease?
- Are all enhancers mechanistically similar or are there fundamentally different types of enhancers?
- How are enhancer-promoter specificities molecularly defined during development?
- What are the dynamics of enhancer-promoter interaction during development?
- Why are genes controlled by multiple enhancers in the same cellular context (enhancer redundancy)?
- What is the identity of pioneer TFs and how do they function in regulating enhancer function during development?
- How can we exploit the ncGENOME to generate synthetic regulatory elements eventually used as therapeutic agents?
- What is the impact of epigenetic modifications of ncRNAs on development and disease?

Systematic genome-wide high-throughput approach is needed to validate *in vivo* enhancer activity. Large-scale systematic studies of enhancer and promoter sequences and interactions combined with computational analyses are key to decipher the rules underlying the complex network of enhancer–promoter interactions during development. Furthermore, CRISPR-based strategies addressed to edit specific mutated enhancer or to modify its epigenetic state will uncover developmentally regulated patterns and provide new therapeutic tool to treat patients. In addition, of the thousands of annotated lncRNAs, only a small fraction has been functionally interrogated. Thus, is still a challenge to systematically identify and characterize all functional lncRNAs from an organism and their cell type-specific role during development, and also how they are epigenetically controlled. Because RNA molecules consist of specific sequences, it is realistic to predict that they will serve to design drug targets to modulate ncRNA (miRNAs, lncRNAs, and circRNAs) activity by modulating their function.

Relevance of the ncGENOME in evolution

- What are the structural and functional differences among the evolutionary lineages (animals, plants) in the various elements of the ncGENOME and their transcripts?
- What comparative genomics can tell us about universal and group-specific mechanisms?
- What is the contribution of the ncGENOME to speciation, evolution and ecophysiological interactions?
- What is the role of TEs in genome reorganization, transfer and dynamics across the tree of life?
- What is the impact of the ncGENOME on genome dynamics, including polyploidy or aneuploidy, among others?

It is relevant to highlight that when assembling genomes, the most difficult regions to resolve are constituted by the repetitive, non-coding regions. Thus, new NGS techniques able to deal with repetitive DNA will uncover the mysteries of repGENOME, evolutionary modes based on repetitive sequences, as well as the origin of alternative organisations in future. Future global analyses of changes in the regulatory landscapes between different organisms will also help to elucidate the genetic mechanisms for evolution.

Technical development of methods for the study of the ncGENOME

- What technological challenges will emerge to advance in strategies to analyze the ncGENOME beyond the current Hi-C, ChIP-seq, ATAC-seq, DNase-seq, high-resolution microscopy, among others?

- What are the technical challenges to advance in our capacity to identify ncGENOME elements and understand their structural and functional features?

As more extensively described in Chapter 3.1, the improvement of single cell technologies, powerful NGS tools, gene editing techniques, and super resolution microscopy would impact in our understanding of the ncGENOME. In addition, new computational and bioinformatics pipelines (see Chapter 3.2) will be essential to interpret this huge amount of data that at the end will provide us exciting insights on temporal and spatial ncGENOME regulation, helping us to identify novel mechanisms underlying development, disease and evolution traits. In addition, new DNA editing methods need to be developed to understand the ncGENOME impact on organism´s development and disease.

CSIC advantage position and multi/inter-disciplinarity

CSIC is a multi-disciplinary scientific institution with numerous research groups that cover different aspects of basic and translational science working with different animal and plant models, chemistry and structural analysis, bioinformatics, and social investigation. Several CSIC research groups work on different aspects of ncGENOME that will contribute to the development of this field in the near future. In the recent years, research in the study of the epiGENOME has experienced a boost, mostly due to the technical advance, in particular the extended use of NGS-related techniques. CSIC scientists are distributed over different research center allocated in distinct national geographic areas. These scientists also interact with international consortia, well-known scientific institutions and research groups that without any doubt will be key in the development of this field. Hence, CSIC is in a particularly privileged and unique position to undertake this challenge and contribute to the remodeling of our view in this area of research.

In order to understand the current position of CSIC in the ncGENOME research landscape, both at Spanish and European levels, a search has been performed across the scientific literature available in NCBI PubMed. We have used as queries certain keywords that are relevant to the field (Figure 1). General molecular biology or biochemistry keywords have also been included for comparison. According to this analysis, the contribution of Spanish researchers represents an average 2.52% of the global scientific production in this specific field - slightly below the 3.41% contribution of Spain to global scientific production in all fields (data from the "Indicadores del Sistema Español de Ciencia, Tecnología e Innovación 2019"). The weight of most of the keywords is similar except for "lncRNA" that is under-represented in CSIC and "genome size" that stands out as an important topic in our institution. When the activity of CSIC

is extracted from the overall Spanish scientific production on the topic, our average contribution represents a quarter of it (25.46%), in accordance with the fact that the CSIC is the leading public research body in Spain, and slightly above the average production of the CSIC in Spain in all fields (about 20%). Compared to other European countries and homolog research bodies, France and Germany contribute an average 4.05% and 5.45% of the world scientific production to the topic, respectively (out of which 46.19% can be attributed to the CNRS and 12.46% to the Max Planck Society).

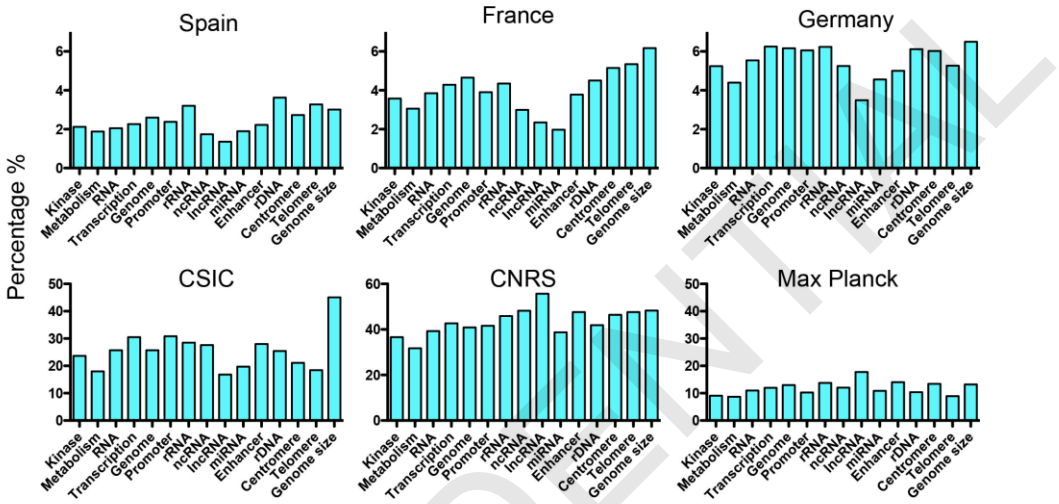

Figure 1.- Comparative analyses of ncGENOME-related key words in different European countries and institutions. The percentage of published articles containing the indicated key words is indicated. This information includes the historical data until May 2020 (NCBI, Pubmed)

A small number of CSIC groups work specifically in the ncGENOME. For example, only a total of 15 CSIC groups (out of 1619 active research groups) display in their group description ncGENOME related keywords such as those previously analyzed (data from CSIC Research Groups database). Of course, many other groups working in chromatin, epigenetics, transcription and genome dynamics may also investigate problems related to the ncGENOME; however, this data shows the small volume that this topic currently represents within CSIC. If CSIC considers that the ncGENOME is a priority area of research, more groups working in this challenge should be created (or more extant groups should incorporate this challenge).

Plan and resources

To reinforce this area of research as a priority, more research groups working on different aspects of ncGENOME should be created in the future. The incorporation of new CSIC scientists without any doubt will contribute to the development of this field in the future. The interactive

effort of a critical mass of groups working in different disciplines, including functional genomics, computational biology, developmental biology, gene regulation, and evolutionary biology will create a fruitful crosstalk to elucidate the molecular mechanism that control gene regulation, organism development, and genome evolution. In addition, an institute devoted to the study of ncGENOME would be very helpful to integrate the efforts in this area and to reach the goal of positioning CSIC at the cutting edge of the field. Understanding that this is a difficult goal to achieve in the short term, the creation of "horizontal" research programs dedicated to this topic can generate greater collaboration and synergies among the research groups that are physically separated. Similarly, it would be very important to have large common facilities similar to those existing in other countries that can integrate the needs for model organisms, including common services for transgenesis and gene editing, technical resources in Genomics, Bioinformatics, Microscopy, Chemistry and Structural Analyses, and Proteomics, among others. The common facilities will help to assist to potential users with well-trained staff and front-line equipment. In this regard, it is essential to stabilize highly qualified personnel at the facilities.

In addition, it will be important to create a competitive scientific career within the CSIC at the different levels, including the possibility of stabilizing researchers at a permanent post-doctoral level that would act as laboratory managers to create a solid scientific atmosphere for student training. In terms of the implications of the ncGENOME in pathology a closer connection between CSIC groups and clinicians should be pursued. In this sense, a clear area of improvement is related to the difficulties that CSIC researchers encounter in interacting with hospitals, for example in obtaining clinical samples. A close partnership between CSIC and hospitals would be highly desirable to provide an appropriate framework to facilitate such interactions. Related to this, it would be important for the CSIC to be considered as part of the National Health Research Institutes to allow its participation in competitive research callings to obtain more funds. Likewise, for the non-human research, a fluid association with other Institutions (OPIs) engaged in research in plants and animals will be very welcome.

An important need is that CSIC could function as an intramural funding agency to support its own research with project grants and contracts. In addition, it would be important to potentiate the interactions and visibility of our Institution with other important networks within the European community and other world-class international competitive research institutions. At present, it would be helpful to create a CSIC platform to promote a cooperative interaction among the distinct groups working in this area, which are dispersed in various research institutes located in different geographical areas. These meetings will help to detect the deficiencies facing the field within our Institution and how to address them for future improvement.

This challenge/chapter 3.4 is highly interconnected with several other challenges treated in this White Book. Especially, the functional and pathological relevance of the ncGENOME is tightly related with topics covered in challenges 3.2, 3.3 and 3.5, including genomics, 3D genome organization, cancer, neurodegeneration, aging, rare diseases, and personalized and precision medicine.

References

Abarrategui, I., Krangel, M.S. (2006) Regulation of T-cell receptor a gene recombination by transcription. Nat. Immunol. 7, 1109-1115.

Alberti, C., Cochella, L. (2017) A framework for understanding the roles of miRNAs in animal development. Development. 144, 2548-2559.

Alessio, E., Bonadio, R.S., Buson, L., Chemello, F., Cagnin, S. (2020) A single cell but many different transcripts: A journey into the world of long non-coding RNAs. Int. J. Mol. Sci. 21, 302.

Anderson, P., Ivanov, P. (2014). tRNA fragments in human health and disease. FEBS Lett. 588, 4297-4304.

Ang, C.E., Ma, Q., Wapinski, O.L., Fan, S., Flynn, R.A., Lee, Q.Y., Coe, B., Onoguchi, M., et al. (2019) The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders. eLife 8, e41770.

Arnold, P.R., Wells, A.D., Li, X.C. (2020) Diversity and emerging roles of enhancer RNA in regulation of gene expression and cell fate. Front. Cell. Dev. Biol. 7, 377

Barra, V., Fachinetti, D. (2018) The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat. Commun. 9, 1-17.

Castellanos-Rubio, A., Fernández-Jiménez, N., Kratchmarov, R., Luo, X., Bhagat, G., Green, P.H., Schneider, R., Kiledjian, et al. (2016) A long noncoding RNA associated with susceptibility to celiac disease. Science 352, 91-95.

Chen, Q., Yan, W., Duan, E. (2016) Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. Nat. Rev. Genet. 17, 733-743.

Cho, S.W., Xu, J., Sun, R., Mumbach, M.R., Carter, A.C., Chen, Y.G., Yost, K.E., Kim, J., et al. (2018) Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. Cell 173, 1398-1412.

Corradin, O., Scacheri, P.C. (2014) Enhancer variants: evaluating functions in common disease. Genome Med. 6, 85.

Csorba, T., Questa, J.I., Sun, Q., Dean, C. (2014) Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. Proc. Natl. Acad. Sci. USA 111, 16160-16165.

Davidson, E.H., McClay, D.R., Hood, L. (2003) Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci U S A. 100, 1475-1480.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., et al. (2010) A large fraction of extragenic RNA pol II enhancers. PLoS Biol. 8, e1000384.

Deryusheva, S., Gall, J.G. (2019) scaRNAs and snoRNAs: Are they limited to specific classes of substrate RNAs? RNA 25, 17-22.

Di Leva, G., Garofalo, M., Croce, C.M. (2014) MicroRNAs in cancer. Annu. Rev. Pathol. 9, 287-314.

ENCODE Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Franke, M., Gómez-Skarmeta, J.L. (2018) An evolutionary perspective of regulatory landscape dynamics in development and disease. Curr. Opin. Cell. Biol. 55, 24-29.

Gillies, S.D., Morrison, S.L., Oi, V.T., Tonegawa, S. (1983) A tissue-specific enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell, 717-728.

Giallourakis, C.C., Franklin, A., Guo, C., Cheng, H.L., Yoon, H.S., Gallagher, M., Perlot, T., Andzelm, M., et al. (2010) Elements between the IgH variable (V) and diversity (D) clusters influence antisense transcription and lineage-specific V(D)J recombination. Proc. Natl. Acad. Sci.USA 107, 22207-22212.

Grüll, M.P., Massé, E. (2019) Mimicry, deception and competition: the life of competing endogenous RNAs. Wiley Interdiscip. Rev. RNA 10, e1525.

Goudarzi M, Berg K, Pieper LM, Schier AF. (2019) Individual long non-coding RNAs have no overt functions in zebrafish embryogenesis, viability and fertility. eLife, 8:e40815. Published 2019 Jan 8.

Ha, C.E., Bhagavan, N.V. (2011). Essentials of medical biochemistry: with clinical cases. Academic Press.

Hanada, T., Weitzer, S., Mair, B., Bernreuther, C., Wainger, B.J., Ichida, J., Hanada, R., Orthofer, M., et al. (2013) CLP1 links tRNA metabolism to progressive motor-neuron loss. Nature 495, 474–480.

Heard, E., Martienssen, R.A. (2014) Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 157, 95-109.

Heinz, S., Romanoski, C.E., Benner, C., Glass ,C.K. (2015) The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell. Biol. 16, 144-154.
Heslop-Harrison, J.S.P., Schwarzacher, T. (2011) Organization of the plant genome in chromosomes. Plant J. 66, 18-33.
Hirsch, C.D., Jiang, J. (2013) Centromeres: Sequences, structure, and biology. Pp. 59-70 in: Wendel, J., Greilhuber, J., Dolezel, J. and Leitch, I.J. (eds.), Plant genome diversity, vol. 1, Plant genomes, their residents, and their evolutionary dynamics. Vienna: Springer.
Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., Young, R.A. (2013) Super-enhancers in the control of cell identity and disease. Cell 155, 934-947.
Huarte, M. (2015) The emerging role of lncRNAs in cancer. Nat. Med. 21, 1253-1261.
Isakova, A., Quake, S.R. (2018) A mouse tissue atlas of small nuclear non-coding RNA. BioRxiv 430561.
Iwafuchi-Doi, M., Zaret, K.S. (2016) Cell fate control by pioneer transcription factors. Development 143, 1833-1837.
Jeyaraman, S., Hanif, E.A.M., Mutalib, N.S.A., Jamal, R., Abu, N. (2020) Circular RNAs: Potential regulators of treatment resistance in human cancers. Front. Genet. 10, 1369.
Kanai, A. (2014) Welcome to the new tRNA world! Front. Genet. 5, 336.
Kanduri, C. (2016) Long noncoding RNAs: Lessons from genomic imprinting. Biochim. Biophys. Acta 1859, 102-111.
Khrosraviani, N., Ostrowski, L.A., Mekhail, K. (2019) Roles for non-coding RNAs in spatial genome organization. Front. Cell. Dev. Biol. 7, 336.
Klemm, S.L., Shipony, Z., Greenleaf. W.J. (2019) Chromatin accessibility and the regulatory epigenome. Nat Rev. Genet. 20, 207-220.
Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., et al. (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182-187.
Lewis, M.W., Li, S., Franco, H.L. (2019) Transcriptional control by enhancers and enhancer RNAs. Transcription. 10, 171-186.
Liu, J.Y., Yao, J., Li, X.M., Song, Y.C., Wang, X.Q., Li, Y.J., Yan, B., Jiang, Q. (2014) Pathogenic role of lncRNA-MALAT1 in endothelial cell dysfunction in diabetes mellitus. Cell Death Dis. 5, e1506.
Martínez, P., Blasco, M.A. (2017) Telomere-driven diseases and telomere-targeting therapies. J. Cell. Biol. 216, 875-887.
Loda, A., Heard, E. (2019) Xist RNA in action: past, present and future. PLoS Genet 15: e1008333.
Martinez-Salas E, Linney E, Hassel J, DePamphilis, ML. 1989. The need for enhancers in gene expression first appears during mouse development with formation of zygotic nucleus. Genes Dev. 3, 1493-1506.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-1195.
Matsumoto, A., Nakayama, A.I. (2018) Hidden peptides encoded by putative non-coding RNAs. Cell Struct. Func. 43, 75-83.
Mensà, E., Latini, S., Ramini, D., Storci, G., Bonafè, M., Olivieri, F. (2019) The telomere world and aging: analytical challenges and future perspectives. Ageing Res. Rev. 50, 27-42.
Nair, S.J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M.J., Wang, S., Suter, T. et al. (2019) Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nat. Struct. Mol. Biol. 26, 193-2013.
Necsulea, A., Kaessmann, H. (2014) Evolutionary dynamics of coding and non-coding transcriptomes. Nat. Rev. Genet. 15, 734-748.
Pamadurti, N.R., Bartok, O., Jens, M., Ashwall-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., et al. (2017) Translation of circRNAs. Mol Cell 66, 9-21.
Pantzartzi, C.N., Pergner, J., Kozmik, Z. (2018) The role of transposable elements in functional evolution of amphioxus genome: the case of opsin gene family. Sci. Rep 8, 1-11.
Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S. et al. (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. USA 110, 17921-26.
Pu, M., Chen, J., Tao, Z., Miao, L., Qui, X., Wang, Y., Ren, J. (2019) Regulatory network of miRNA on its target: Coordination between transcriptional and post-transcriptional regulation of gene expression. Cell. Mol. Life Sci. 76, 441-451.
Raina, M., Ibba, M. (2014) tRNAs as regulators of biological processes. Front. Genet. 5, 171.
Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S. (2011) Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. 12, 529-541.
Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., Gage, F.H. (2013).The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. 33, 17577-17586.

Roe, J.S., Hwang, C.I., Somerville, T.D.D., Milazzo, J.P., Lee, E.J., Da Silva, B., Maiorino, L., Tiriac, H. et al. (2017) Enhancer reprogramming promotes pancreatic cancer metastasis. Cell 170, 875-888.
Quinn, J.J., Chang, H.Y. (2017) Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. 17, 47-62.
Sabari, B.R., Dall´Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannet, M.N. et al. (2018) Coactivator condensation at super-enhancers links phase-separation and gene control. Science 361, eaar3958.
Schoenfelder, S., Fraser, S. (2019) Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. 20, 437-455.
Serrato-Capuchina, A., Matute, D.R. (2018) The role of transposable elements in speciation. Genes 9, 254.
Shlyueva, D., Stampfel, G., Stark, A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. 15, 272-286.
Smith, K.P., Hall, L.L., Lawrence, J.B. (2020) Nuclear hubs built on RNAs and clustered organization of the genome. Curr. Opin. Cell. Biol. 64, 67-76.
Smith, E, Shilatifard, A (2014) Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. 21, 210-219.
Tang, W.W., Dietmann, S, Irie, N., Leitch, H.G., Floros, V.I., Bradshaw, C.R., Hackett, J.A., Chinnery, P.F., et al. (2015) A unique gene regulatory network resets the human germline epigenome for development. Cell 161, 1453-1467.
Thompson, D.M., Lu, C., Green, P.J., Parker, R. (2008) tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA 14, 2095-2103.
Toiyama, Y., Okugawa, Y., Tanaka, K., Araki, T., Uchida, K., Hishida, A., Uchino, M., Ikeuchi, H., et al. (2017) A panel of methylated microRNA biomarkers for identifying high-risk patients with ulcerative colitis-assocaited colorectal cancer. Gastroenterology 153, 1634-1646.
Tuorto, F., Parlato, R. (2019) rRNA and tRNA bridges to neuronal homeostasis in health and disease. J. Mol. Biol. 431:1763-1779.
Tyebji, S., Hannan, A.J., Tonkin, C.J. (2020) Pathogenic Infection in Male Mice Changes Sperm Small RNA Profiles and Transgenerationally Alters Offspring Behavior. Cell Rep. 31:107573.
Ulitsky, I. (2018) Interactions between short and long noncoding RNAs. FEBS Lett. 592, 2874-2883.
Wang, M., Lemos, B. (2019) Ribosomal DNA harbors an evolutionarily conserved clock of biological aging. Genome Res. 29, 325-333.
Wang, Y., Mo, Y., Gong, Z., Yang, X., Yang, M., Zhang, S., Xiong, F., Xiang, B. et al. (2017) Circular RNAs in human cancer. Mol. Cancer 16, 1-8.
Warmerdam, D.O., Wolthuis, R.M. (2019) Keeping ribosomal DNA intact: a repeating challenge. Chromosome Res, 27:57-72.

3.5 Functional Epigenetics and Epitranscriptomics and Their Role in Health and Disease

ABSTRACT

In recent years, there have been great efforts to characterize the epigenome and epitranscriptome of different cell types and organisms. However, most of those studies are descriptive and we still ignore the role of many epi-modifications. We discuss the need of gaining functional and mechanistic insight and the large impact that this knowledge will have in the understanding and treatment of numerous diseases.

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS (in alphabetical order)

- Angel Barco (IN, Alicante, *Coordinator*)
- Sandra Blanco Benavente (IBMCC, Salamanca)
- Elena Gómez Díaz (IPBLN, Granada)
- María Gómez Vicentefranqueira (CBMSO, Madrid, *Deputy Coordinator*).
- Javier Martin (IPBLN, Granada)
- Marián Martínez Balbás (IBMB, Barcelona)
- Jordi Pérez-Tur (IBV, Valencia)
- Isidro Sánchez Garcia (IBMCC, Salamanca)
- Carles Suñé (IPBLN, Granada)
- Mario Vallejo (IIBM, Madrid)

Executive Summary

During recent years, there have been great international efforts to characterize the epigenome of different cell types and organisms. More than 100 distinct covalent modifications of the chromatin affecting both the DNA and the histone proteins have been identified to date. The emergence of epitranscriptomics is more recent and new studies are unveiling new and unexpected layers of regulation of gene expression. Despite this progress, we still ignore the specific function, if any, of most of these epigenetic and epitranscriptomic modifications and their impact on transcription, translation and cell biology.

In this chapter, we will first discuss the importance of functionally and mechanistically characterizing the epigenome and the epitranscriptome in different cellular contexts. Next, we will emphasize the relevance of such studies for the understanding and treatment of numerous diseases, including rare syndromes, cancer, degenerative, autoimmune and infectious diseases, and metabolic, neurological and psychiatric disorders. The investigation of epigenetic mechanisms in all these conditions has already contributed to a better understanding of their etiopathology. Next, we discuss the challenges that the field still faces. The constant technological development is unveiling new mechanisms and events that sculpt the epigenome and epitranscriptome, leading to the production of an enormous amount of genomic and transcriptomic information that should be processed and integrated to fully understand the function of these epi-changes and their implication in disease. Although changes in the chromatin have been reported in many disorders and these changes often correlate with the progression of the disease, it remains unknown whether these epigenetic alterations are a cause or a consequence of the pathology. Innovative technologies for precise manipulation of the epigenome and epitranscriptome should enable us to tackle the causality conundrum in the near future. In addition, compounds with the potential to reestablish the normal epigenetic and epitranscriptomic landscape should be identified and evaluated in clinical trials. These epigenetic drugs may open new avenues for therapy of a great number of disorders.

The CSIC counts with numerous and excellent research groups working on epigenetics from different perspectives. In the final part of this chapter, we will discuss the resources dedicated at CSIC to confront these challenges and the actions that could be put in place in our institution to improve its position in this rapidly developing and essential field of research.

Introduction and general description

The epigenome

Although all somatic cells of a multicellular organism have the same genome (i.e., identical DNA sequence), different cell types have different transcriptomes (set of all expressed RNA molecules), different proteomes (set of all proteins) and, hence, different functions. This is largely achieved throughout modifications in the chromatin. The basic structural unit of eukaryotic chromatin is the nucleosome, in which approximately 150 bp of DNA is wrapped around a basic protein core comprising two copies of histones H2A, H2B, H3 and H4. Chromatin modifications that affect both the DNA and the histones enable propagation of active and silent activity states from mother to daughter cells within a given cell lineage. Architectural and conformational changes in the nucleosomes, and gene regulatory feedback loops also contribute to modify the expression of genetic information without altering the genetic information itself. These mechanisms are collectively referred to as the "epigenome".

One of the most investigated epigenetic processes is DNA methylation, which occurs with different patterns in the genome of microorganisms, plants and animal phyla, while consistently playing protecting and regulatory roles. Other important epigenetic systems in eukaryotic organisms include the heterochromatin (HP1 and H3K9me3), Polycomb (PRC1 and PRC2) and Trithorax complexes. The transmission of epigenetic marks through cell division requires that they survive DNA replication and mitosis, what is particularly relevant for histone modifications, because nucleosomes do not have a DNA template-based duplication system. The afore referred complexes are thought to perpetuate functional responses by binding specific histone modifications and modifying other histone proteins in the nucleosome core in order to convey stable inheritance. Epigenetic inheritance usually involves the cooperation of partially overlapping signals, each of them adding a degree of stability, but also being each of them reversible, allowing plasticity in the presence of regulatory cues (Cavalli and Heard, 2019). The genomes of unicellular organisms also carry epigenetic information. Recent methylome analysis has shown that DNA methylation is widespread in the genomes of bacteria and archaea, including the small genomes of certain obligate parasites. These studies are shaking some old conceptions and show that epigenetic lineages enable the adaptation of bacterial populations to harsh or changing environments and modulate the interaction of pathogens with their eukaryotic hosts. In addition to their role in epigenetic inheritance, the modifications of the chromatin also play a critical role in cellular plasticity. Thanks to changes in their epigenome, non-dividing cells in organs, such as the

muscle or the brain, can adapt or respond differently to changes in their environment depending on their previous activation history.

Molecular biology, genomics, and mass spectrometry–based proteomics have identified over one hundred posttranslational modifications (PTMs) of specific histone residues that may work in a combinatorial manner generating thousands of patterns, many of whose functions are under intense investigation. DNA can be also modified in multiple ways, from the abundant methylation of cytosines in the chromatin of mammals and plants to less prevalent changes in both cytosines and the other bases. However, it is still under debate whether they are mere metabolic intermediates or whether they play a functional role in DNA biology. Careful chromatin investigation using chemical, cell and molecular biology approaches have provided valuable insights into the molecular function of epigenetic regulators. These insights have underscored the highly dynamic nature of the epigenome and provided the molecular rationale for therapeutically targeting these proteins. The outcome of 25 years of intensive research in the field of epigenetics and epigenomics have revealed an extraordinarily complex scenario with hundreds of proteins that introduce (writers), eliminate (erasers) or bind (readers) specific chromatin modifications. This complexity underscores the essential contribution of epigenetic mechanisms in development, aging, tissue regeneration and cell plasticity, as well as the involvement of epigenetic dysregulation in a great number of diseases.

The epitranscriptome.

Similar to DNA, RNA can be modified in diverse and complex ways. Chemical deposition can occur at all nucleotides and positions (carbon, nitrogen, oxygen at bases or ribose) and encompasses more than 170 modifications, including methylation, hydroxymethylation, hydroxylation, acetylation, pseudouridylation y glycosylation (Boccaletto et al., 2018). RNA posttranscriptional processing does not only encompass deposition of chemical groups, but also sequence editing, trimming and splicing of unnecessary sequences, addition of ribonucleotides not included in the genetic code, and binding to proteins to lock their tridimensional structure or gain catalytic properties. Together, these RNA modifications constitute the "epitranscriptome", a term only coined a few years ago. Two important advances triggered the emergence of this novel research area. First, the discovery that some disease linked genes encode for RNA modifying enzymes (e.g., *FTO* encoding a RNA 6-methyladenosine (m^6A) demethylase) (Jia et al., 2011). Second, the development of robust detection methods for mapping m^6A and other RNA changes has allowed the development of

mass sequencing methods to study the epitranscriptome (Dominissini et al., 2012; Meyer et al., 2012). Pioneering research on the physiological function of these modifications has revealed regulatory roles in a variety of cellular processes, including stem cell self-renewal and differentiation, proliferation, development, responses to environmental cues, migration, survival to stress, immune response, mitochondrial function, and the circadian clock (Frye et al., 2018). Understanding the role of these dynamic RNA posttranscriptional modifications represents a new frontier in research often referred to as "epitranscriptomics".

The intense research in the last years, together with the development of novel technologies and tools, has unveiled a complex array of new regulatory mechanisms of gene expression that affects the location, stability and translation efficiency of RNA molecules (Davalos et al., 2018). In addition to messenger RNAs (mRNAs), RNA modifications also occur in the much less explored non-coding genome, including small and long non-coding RNAs and transposable elements, whose roles in normal development and pathological processes might be of fundamental importance, as described in more detail in Chapter 3.4. Moreover, similar to epigenetic enzymes, we are discovering a large number of RNA-modifying enzymes that deposit (writers) or remove (erasers) specific modifications at different nucleotide positions, RNA domains or RNA types. In addition, groups of proteins have been identified that specifically bind to modified nucleotides (readers), thereby affecting the fate of RNA. Also, similar to histones and DNA, RNA binding proteins (RBPs) can be modified as part of the epitranscriptome repertoire. Mutations in all these proteins (writers, erasers, and readers) have been associated with various pathologies, from cancer to neuronal dysfunction, fertility or metabolism (Harries, 2019), which makes epitranscriptomic enzymes and their substrates promising therapeutic targets.

Impact in basic science panorama and potential applications

Functional epigenetics and epitranscriptomics

The study of epigenetics and epitranscriptomics has experienced a strong impulse in the past years. This is mainly due to the advent of technological advancements such as the ability to analyze the whole epigenome or the whole epitranscriptome in a single experiment and the possibility of studying a greater diversity of epi-marks. Recent developments have shown that 5-methylcytosine and its hydroxylated counterpart are just two of a number of modifications that DNA undergoes to modulate gene regulation. The finding that RNA molecules are subjected to an enormous plethora of chemical modifications, add new layers of complexity to the sophisticated gene expression control panel in a cell. These techniques include whole-

genome bisulfite sequencing to investigate DNA and RNA methylation, ChIP-seq to map histone modifications and transcription factor binding, ATAC-seq to explore chromatin accessibility and occupancy, Hi-C to elucidate chromatin architecture, iCLIP-seq to identify RBP biding sites in RNA, miCLIP, m^6A-seq or m^1A-seq to explore RNA methylation at adenines in RNA, RiboMeth-seq and Nm-seq to detect ribose methylation in RNA, aza-IP to detect cytosine methylation in RNA or ψ-seq and CeU-seq to detect pseurouridine among others. In addition, we have discovered that chromatin and RNA modifications can be cell type or cell state-specific, and organism-specific, which further increases the complexity of epigenetic and epitranscriptomic regulation.

Epigenetic and epitranscriptomic marks are dynamic in nature, and have the ability to appear or disappear in response to external stimuli and environmental influences, such as nutrients or stress. They also change from development to aged organisms. Mounting evidence demonstrates regulatory roles that enable quick cell adaptations for environment changes. The changes of the epigenome and epitranscriptome can be showing both the environmental past as well as the future susceptibility to disease. Thousands of studies have shown the association of aberrant deposition of epi-modifications with diseases ranging from rare to common, and from metabolic to autoimmune, psychiatric or cancer, as we will discuss in detail in the next section. Beyond human health, epigenetic and epitranscriptomic studies provide clues for crops and animal production improvement, for manipulating host-microbe interactions and to enhance relevant biotechnological products and the food industry (see Chapter 3.6 for more details). Contaminants are often chemical precursors of DNA, histone and RNA modifiers. As more extensively described in Chapter 3.6, understanding how the exposure to different environmental situations alter the epigenetic landscape will serve to predict what may happen in the future as well as to identify exposures that occurred in the past, therefore contributing to delineate the etiology of environmental diseases as well as enabling preventive interventions. Furthermore, growing evidence demonstrate the transmission of epigenetic information through the germ line (i.e., the so called "transgenerational epigenetic inheritance") (Horsthemke, 2018). Residual DNA and histone modifications in germ cells and long-lived RNA molecules have been postulated as possible carriers of epigenetic information across generations. These transgenerational mechanisms and their importance for health and disease will be more extensively covered in Chapter 3.7.

Epigenetic mechanisms in disease

The multidisciplinary collaborative research of geneticists, biochemists, medical chemists,

cell biologists, clinicians and bioinformaticians has yielded an enormous amount of knowledge about the fundamental role of epigenetic regulators in etiopathology. These advancements, and those still under development, have allowed the identification of several promising therapeutic targets that are currently being explored mainly in brain, autoimmune or cancer-related disorders. However, deciphering the mechanistic role of epigenetic changes in disease's origin and progression is challenging and will require the development of functional studies in model organisms and/or cell culture systems. Research on those models has shown that epi-marks can be modified by pharmacological interventions or changes in the environment, thereby opening new and unsuspected venues for therapy. Reproducing these collaborative efforts in the emerging field of epitranscriptomics represents an opportunity niche for future research. We will underscore in the next paragraphs the importance of such studies in different areas of biomedicine.

Epigenetic etiology of rare disorders: Recent genome-wide approaches have enabled the identification of an ever-increasing number of hereditary rare disorders caused by mutations in chromatin-acting factors, including DNA methyltransferases, histone modifying enzymes, chromatin remodeling factors and reader proteins (Bjornsson, 2015; Velasco and Francastel, 2019). Although rare diseases individually affect less than 1 person in 2,000, globally more than 30 million people are estimated to be affected by rare diseases only in the EU (The Lancet Diabetes, 2019). Most of these conditions do not have an approved treatment and represent a tremendous burden for patients, families and society. The etiology of these conditions involves defects in the establishment of epigenetic marks early during development or in the perpetuation of these marks at later stages. In many cases, these epigenetic alterations are associated with neuropathies, neurodevelopmental disorders, intellectual disability and immunodeficiency. In all the cases the epigenomic landscape is altered and changes of transcription profiles are observed. Interestingly, the deficiency of different epigenetic factors often generates common phenotypes and symptoms, suggesting their actions converge on common pathways and genes. We still poorly understand the nature of these nodes and the mechanisms that link the epigenetic changes to the clinical manifestations. There is a great need to identify the epigenetic mechanisms that govern the outcome of these conditions in order to identify reliable biomarkers to improve diagnosis and treatments that diminish or eliminate the most severe symptoms.

Cancer epigenetics: Despite the enormous amount of data gathered in the last four decades

about the biology of tumor cells, yet our ability to understand and control the development of cancer is unfortunately still limited. We do not yet know how to prevent the conversion of a pre-cancer cell into a tumor, mainly due to the fact that the early events triggering the commitment to a new cancer lineage remain largely unknown. A crucial point in the history of a tumor is the transition of a normal cell to a malignant state. Recent evidence from hematopoietic and epithelial tumors revealed that the contribution of oncogenes to cancer development is mediated mainly through epigenetic priming of cancer-initiating cells, suggesting that genetic lesions that initiate the cancer process might be dispensable for the posterior tumor progression and maintenance (Vicente-Dueñas et al., 2018). In the initial stages of cancer development, a normal cell is going to become a pre-cancer cell by the action of a given oncogenic hit or by the exposure to an environmental factor (like tobacco smoke in lung adenocarcinoma); both can "reset" the epigenetic and/or transcriptome status and reprogram the epigenome to give rise to a pre-cancer cell. Once its role in oncogenic reprogramming is performed, the initiating hit is no longer necessary for tumor progression. The malignant epigenetic priming can take place early in life and remain silent until specific second events will trigger cancer appearance. These second hits can happen randomly or can be triggered by environmental factors or aging (Sen et al., 2016; Tomasetti and Vogelstein, 2015). Therefore, a detailed understanding of the epigenetic rewiring is a prerequisite for the development of any potential cancer therapy directed at the epigenome of precancer cells. In addition, while we still do not know if epigenetic priming can be the sole driver of cancer, the advent in novel tools to detect the priming event before the development of a full-blown tumor are beginning to shed light on this mechanism. Progress in this area and development of novel therapeutic tools holds great promise for reverting epigenetic changes related to cancer. Similar to epigenetics, the epitranscriptome is a new layer of complexity in cancer biology. Researchers have been compiling data that implicate post-transcriptional modifications of RNA with roles either as tumor-suppressive or tumor-promoting function (Barbieri and Kouzarides, 2020).

Epigenetics of metabolic disorders: From an epidemiological point of view, the two most important metabolic diseases are obesity and type 2 diabetes. Both disorders have reached epidemic proportions worldwide and are thought to have an important epigenetic component. Their increased prevalence has been related to the improvement in living standards together with increased sedentarism and easy access to abundant fast and high-energy-containing food. Several studies have shown a strong correlation between epigenetic signatures and clinical

traits associated with obesity or adipose tissue distribution, often affecting genes related to insulin/glucose metabolism, lipid metabolism and adipogenesis, or regulation of food intake. However, it is not entirely clear whether all these changes are the cause or are secondary to metabolic dysfunction, especially in studies carried out using blood samples rather than adipose tissue (Wahl et al., 2017). The regions exhibiting abnormal DNA methylation are often associated with genes known to regulate metabolism and show differential gene expression, thus linking epigenetic mechanisms with islet dysfunction (Volkov et al., 2017). However, the individual contribution of these genes or sites to diabetes is rather small, in line with the complex polygenic and multifactorial nature of the disease. Several studies have also shown that high-fat diets and excessive intake of saturated fatty acids induce epigenetic and gene expression changes in muscle, adipose tissue and pancreatic islets. Moreover, epigenetic modifications during intrauterine development in mothers with poor nutritional or health conditions can lead to the appearance of metabolic disorders in their progeny years after birth, including obesity and diabetes, and even into the following generations (Sales et al., 2017). Therefore, an important aspect for future studies will be to understand the mechanisms by which feeding habits affect the epigenome, especially in predisposed individuals (see also Chapter 3.7). The post-transcriptional m^6A RNA modification also plays an important role in glucose and lipid metabolism, and some m^6A regulators may be involved in critical liver pathways related to obesity and the metabolic syndrome. Further investigations will pave the road to interventions that target specific epigenetic and epitranscriptomic pathways for treatment.

Epigenetics of degenerative diseases: Degenerative disorders are a heterogeneous group of conditions that may affect tissues, organs or the whole body. Although some of these conditions have a genetic origin, others are triggered by environmental factors or arise with aging. The role of epigenetics in aging and age-related diseases is well documented in the literature. Alterations in DNA, histone modifications and composition, and in the regulation of non-coding RNAs, are all part of the aging process and are thought to contribute to neurodegenerative conditions related to aging. The incidence of these diseases, including Alzheimer's disease and other dementias, Parkinson's disease and other movement disorders or neuromuscular disorders, nearly doubles with every decade of age and that is why they represent one of the main health problems of the aging population. As with aging, there is an overall reduction in 5-methylcytosine in several anatomical areas, together with increase/decrease at specific genes such as *SNCA* in Parkinson's or *IL-1* in Alzheimer's. Also

seen are alterations in histone modification levels or, even, in histone composition, as well as deregulation or differential expression of some microRNAs. The analysis of epigenetic modifications both at the genome-wide and locus-specific levels will uncover the genomic impact of the aging process and pinpoint approaches for intervention that could alleviate or prevent age-related degeneration and other undesirable effects of aging.

Epigenetics of neurological and psychiatric disorders: Our knowledge of the etiology of many of neurological and psychiatric disorders is still limited, in part due to their complex origin. Although some neurological conditions are originated by mutations in single genes, most mental disorders have a polygenic origin. Typically, these conditions have a heritable component, but the contribution of environmental factors ranges from less than 25% in conditions such as schizophrenia, bipolar disorder, autism spectrum disorders, or attention deficit hyperactivity disorder (ADHD), to more than 60% in anxiety disorders, obsessive–compulsive disorder (OCD), posttraumatic stress disorder (PTSD), and major depressive disorders. The investigation of epigenetic mechanisms in brain function has contributed to a better understanding of this non-heritable component of mental illness. This research has demonstrated that the epigenetic regulation of gene expression is not restricted to developmental processes, but also plays a critical role in mature neurons influencing a wide range of basic mechanisms in the adult brain, providing new cues about how the environment and our experiences can interact with our genome. Similarly, RNA modifications, such as m^6A, can drive region-specific post-transcriptional regulatory networks in the brain and contribute to brain diseases (Chang et al., 2017). Epigenetic and epitranscriptomic dysregulation seems a feature of numerous neurological and psychiatric disorders that can importantly contribute to their etiology. This is the case for numerous neurodevelopmental disorders associated with intellectual disability and autism, neurodegenerative diseases such as Alzheimer's or Huntington's, and psychiatric disorders including drug addiction and schizophrenia. The difficult access to brain tissue and its extreme complexity both in terms of cell diversity and number, causes specific challenges to the investigation in this area since most of the current methods to detect epi-changes rely on laborious biochemical or immunological procedures that require a significant amount of homogeneous cells.

Epigenetics of autoimmune diseases: The last few years have witnessed an increasing interest and appreciation for the role of epigenetic regulation in the healthy immune system and in autoimmunity. The development of genome-wide DNA methylation array-based technology

and high-throughput sequencing has allowed the evaluation of specific epigenetic marks across the genome in patients with a number of autoimmune diseases, as well as the identification and characterization of specific regions within the genome that are epigenetically altered compared with healthy controls. Nevertheless, the existing knowledge cannot fully explain whether epigenetic alterations cause or follow the increased immune activation, making their precise characterization a requirement for a comprehensive understanding of the pathogenetic mechanisms that complements genetic and clinical studies. Integrating data from disease-specific and cell-specific DNA methylation states, histone modifications, and non-coding RNA activity, in addition to genomics and transcriptomics data, and the application of novel methodologies such as single-cell RNAseq or gene editing will provide a better picture of the role of epigenetics in the etiology, prognosis and treatment of autoimmune diseases.

Epigenetics of infectious diseases: Human infectious diseases caused by bacteria, viruses, parasites, and fungi are the second most prevalent and represent 20% of all human diseases. The interactions between pathogen and host require rapid adaptation and evolution. Parasites survival depends on evading the host immune system to ensure persistence. In turn, the host must evolve defense mechanisms to avoid invasion and eliminate the invading microorganism (resistance), or to limit the damage caused by the infection (tolerance). Epigenetic processes in both, the host and pathogen, play a key role regulating host-pathogen interactions during infection, and in the evolvability and rapid adaptation of infectious agents (see also Chapter 3.6). An intriguing possibility is that epigenetic machinery of intracellular pathogens may directly alter the host genome (Sanchez-Romero and Casadesus, 2020). A better understanding of the regulatory mechanisms that control variant infection phenotypes is essential to prevent the emergence and re-emergence of infectious diseases and the failure of existing control/eradication interventions. Furthermore, although the epitranscriptome has mostly been investigated in eukaryotic organisms, RNA modifications are also present in the genome of numerous microorganisms. For example, the RNA genomes of many viruses hold numerous RNA modifications that influence their growth and infectivity (Netzband and Pager, 2020). The recent discovery of the enzymes in charge of adding and eliminating these post-transcriptional modifications, as well as the great technological advances in the determination and analysis of the epitranscriptome by means of immunoprecipitation and massive sequencing, have allowed us to begin to know the effect of these RNA post-transcriptional modifications in viral pathogenesis. The recent prevalence of RNA-virus

infections strongly argues in favor of the integration of epitranscriptomic studies in this area of research.

Key challenging points

I. Development of methods and tools for rapid and quantitative detection of epi-changes and nucleic acid-protein interactions with single nucleotide precision: Scientists already possess a powerful arsenal of techniques for mapping modifications across the genome and the transcriptome. Reduction in costs and refinements that enable the scale-down of these techniques to low cell numbers, in the range of single to a few thousand cells, are allowing a very rapid progress in our understanding of regulatory landscapes. However, we are still far from completing the map of epigenetic and especially epitranscriptomic marks in use by the different cell types in a given organism. Future progress will require sophisticated high-throughput techniques to expand and complement the existing ones. Emerging single direct molecule sequencing technologies and development of antibodies specific to various RNA modifications could enable charting transcript-specific epitranscriptomic marks across cell types. Furthermore, uncovering the combinatorial regulatory interplay between different RNA modifications is required to unveil the "epitranscriptomic code" and its relevance for human health and disease. In addition, it will be of the outmost importance the development of methodologies that could allow the unequivocal determination of the epi-modifications in a specific cell type regardless of the tissue of origin. The main objective should be to reach the ability to detect a particular modification, in a particular cell within gene/locus or even nucleotide resolution. Equally important is the development of innovative methods and sequencing technologies capable of simultaneously detect different epigenetic marks in the same locus or epitranscriptomic marks within the same transcript. Among these advances, the detailed description of alterations on the chromatin landscape and the global 3D chromatin structure seems essential to understand the contribution of chromatin to gene regulation and the establishment of disease. Translating tissue-specific epigenetic signatures into 3D chromatin architecture represents a new and promising line of research (see also Chapter 3.3 for more details). Complete high-resolution maps are required to sort out interactions between epigenetic profiles and quantitative trait loci or genetic variants associated with disease, as some of these are related to single nucleotide polymorphisms affecting CpG methylation sites.

II. System-wide understanding of DNA and RNA modifications and dynamics: The new techniques outlined above have the potential of generating a massive and difficult to handle

amount of data. Thus, we face the enormous challenge of integrating novel epigenomic and epitranscriptomic data with genetic, transcriptomic and proteomic outcomes, genomic associations and chromatin 3D interactions in cell-type, stimulus- and location-specific context to provide a better picture of their functional implications. The availability of genomic and epitranscriptomic profiles for the complete repertoire of writer, reader and eraser proteins and their substrates should lead to a comprehensive understanding of gene expression regulation and dynamics. Unveiling the crosstalk between histone, DNA and RNA modifications will require computational standardized solutions to identify, analyze and integrate high-throughput RNA modifications with epigenomic data. The development of new bioinformatic integrative data analyses using increased computing power will be required.

III. Mapping the epigenome and epitranscriptome in four dimensions: Both the epigenome and epitranscriptome are dynamic by nature and thus we will need to understand the mechanisms that drive their change during development, life experience, interaction with the environment and aging. It is essential to determine the transcriptional and epigenetic alterations that occur during normal development because this information will help us to determine the spatiotemporal window in which the deficiency of epigenetic or epitranscriptomic factors contributes most to the development of the disease, facilitating premature therapeutic interventions. This should dictate diagnostics based not only on the symptoms but also on molecular features, including altered patterns of DNA methylation, histone, or RNA modifications. This knowledge and the expansion of genomic screens, currently largely limited to the exome, to the non-coding genome should help to elucidate the molecular mechanisms underlying a large number of undiagnosed rare diseases. Moreover, a precise understanding of the dynamic of epi-modifications should also clarify the type of traits and information that can be transmitted to the offspring and effect future generations.

IV. Solving epigenetic and epitranscriptomic mechanisms of etiopathology. There is a major need to elaborate a complete catalog of epigenomic and epitranscriptomic variations associated to human diseases in general. They might serve as biomarkers for disease activity or disease course. To date, the majority of epigenome-wide association studies have been based on the use of arrays to identify CpG methylation sites, and therefore a very large proportion of the epigenome remains to be discovered. In addition, many possible epigenetic mechanisms remain unexplored and their research can take us in unexpected directions. Likewise, recent advances in epitranscriptomic research clearly associate alterations in RNA

modifying enzymes with disease occurrences such as cancer, demonstrating that inhibition of those enzymes may have enormous therapeutic potential. Yet epitranscriptome-wide association studies remain to be discovered and also hundreds of possible epitranscriptomic mechanisms remain unexplored. Once we complete the catalog of epi-modifications linked to diseases, we will still have to determine their specific contribution to the disease state, as well as the crosstalk between genetics and epigenetics, and between transcriptomics and epitranscriptomics. This enormous challenge should be accomplished, preferentially, on the initial stages of the disease because any epigenetic alterations observed in an autopsied tissue may reflect (especially in long-lasting diseases such as the neurodegenerative ones) changes related to disease advancement, which makes particularly challenging to differentiate between cause and consequence. In the field of cancer epigenetics, we still do not know whether the decision to initiate cancer takes place during tumor differentiation or if it is composed by a series of consecutive decisions. Future progress on cancer prevention may rely on the identification of specific exposures as triggers of epigenetic priming and on protecting susceptible individuals from being exposed to environmental factors that can trigger cancer (for example, infections in *gene susceptible*-carrying children). Efforts should be also devoted to study the changes in the epigenome and epitranscriptome of the host cell induced by the pathogen infection to influence host responses and to contribute to other forms of human disease.

V. Understanding disorders at the single cell level: The advent of single-cell approaches offers a unique opportunity to gain insights into mechanisms underlying cell identity, phenotype and response to stimuli, stressors and pathogens (Avraham et al., 2015). The individual nature of epi-changes, the fact that they differentially affect cells within a tissue, is one of the main obstacles when studying the influence or the role of epigenetic changes in etiopathology, particularly for brain diseases and other complex tissues. Whereas in genetic analysis, the use of tissues or cells targeted by certain diseases is not essential because most of the genetic variation causing a disease is found in all the cells, this cannot be applied to epigenetic diseases. Epigenetic marks change depending on the cell-type and, therefore, their study requires the specific analysis of those cells. The ongoing refinement and new development of genome-wide techniques to explore the epigenome, transcriptome and epitranscriptome at the single cell level should address the challenges caused by cellular diversity and reduce the amount of necessary tissue (see also Chapter 3.1 for more details about single-cell methods). These techniques will allow the analysis of epigenetic,

epitranscriptomic and gene expression changes in restricted cell populations. This may lead to an era of new discoveries that have the potential to radically change our understanding of diseases, particularly those affecting complex heterogeneous tissues such as the brain and tissues undergoing pathogen infection. This knowledge is pivotal, for example, for understanding phase variation and bet-hedging strategies involved in antibiotic resistance and immune evasion, and to anticipate rapid pathogen adaptation to new drugs and vaccines (Chattopadhyay et al., 2018). This should also clarify the striking neuron type specificity of most neurodegenerative conditions and the differences in success treating different types of cancer. Single cell analysis technology will be also a welcome addition to the arsenal of research tools to investigate metabolic disorders.

VI. Inferring causality of epi-modifications: Over the past decade, we have been able to draw epigenomic- and epitranscriptomic-wide maps and learned how a few players (mainly writers, readers, and erasers) disrupt these systems. While providing useful positional information and clues on the molecular mechanisms at play, much effort needs to be invested in developing functional assays that incorporate the data gathered using the aforementioned global methods to understand the effects of the complex interplay of these modifications. Changes in DNA methylation, histone posttranslational modifications (HPTM) and other changes in the chromatin have been reported in many disorders and these changes often correlate with the progression of the disease. However, it remains unknown whether these epigenetic alterations are cause or consequence, maybe even indirect, of the pathology. The ability to specifically edit the epigenome and the epitranscriptome holds promise of enhancing our understanding of how epi-modifications function and of enabling manipulation of cell phenotype for therapeutic purposes. The recent revolution in genome engineering technologies has allowed the use of highly specific DNA- and RNA-targeting tools to precisely deposit epigenetic changes or edit RNA sequences in a locus-specific manner, creating diverse epigenome and epitranscriptome editing platforms. For instance, CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 technology has been adapted to epigenetic editing through to creation of chimeric proteins between a nuclease-dead Cas9 (dCas9) with catalytic domains responsible for chromatin modification. Because guide RNAs are easy to design and can target the catalytic activity to virtually any region in the genome, one may, in principle, locally alter the epigenetic profile at any locus in different ways depending on the enzymatic activity coupled to dCas9. Importantly, these novel genetic tools (contrary to genome editing by the conventional CRISPR/Cas9 system) are equally effective in dividing and non dividing cells, and provide unprecedented means to increase or decrease the expression of any gene of

interest in cell types potentially resistant to gene editing approaches. This is an area still under development and novel and sophisticated tools for epi-editing are becoming available. Further development of the current technologies based on the CRISPR/dCas9 system to precisely manipulate the epigenome and the innovation of groundbreaking technologies for manipulation of the epitranscriptome should enable us to tackle the causality conundrum for most epi-modifications in the near future (Voigt and Reinberg, 2013).

VII. New epi-therapies: Over the last decade, medical chemists have produced an unprecedented array of small molecules that target proteins responsible for writing, reading or erasing epigenetic marks in the chromatin. Several of these epigenetic therapies have already reached the clinic to combat cancer, and many others have progressed to early-phase clinical trials in a plethora of conditions. The use of epigenetic drugs in combination with other therapies has opened very promising new avenues to fight disease (Michalak et al., 2019). For instance, the existence of an epigenetically-driven mechanism of tumor initiation opens new possibilities for preventing or abolishing cancer since epigenetic modifications, unlike genetic changes, can be erased, manipulated, and reinitiated even before a pre-cancerous cell might evolve into cancer.

Further investigation of compounds targeted to RNA modifiers linked to disease and the interplay between epitranscriptome and epigenome will undoubtedly identify novel therapeutic targets. Understanding the machineries and factors that introduce, remove and read chromatin and RNA modifications will allow their modulation through the development of novel drugs with pharmaceutical value. It is also urgent to expand the methods to deliver these epi-drugs; nanotechnology-based strategies will facilitate this task. In addition to pharmacological approaches, there is also a great deal of interest in exploring the possibility to directly correct epigenetic alterations using the epi-editing methods referred above (Hilton et al., 2015; Kwon et al., 2017). Although the use of this incipient technology in complex organs, like the brain, confronts the prominent challenges associated with gene therapy in these organs (i.e., biosafety, cell specificity, accessibility to diseased tissue, etc.), it still represents an important area of development that may allow personalized therapeutic approaches to correct chromatin alterations.

VIII. Social epigenomics and epidemiology: It becomes urgent to undertake large population studies to characterize the epigenetic and epitranscriptome landscape in baseline situations, and how these landscapes evolve with the aging process and environmental influence (see

also Chapters 3.6 and 3.7). There is also a strong need for understanding how different pathologies affect these landscapes and to incorporate disease heterogeneity into the study design. Epigenetic marks may foresee the appearance of diseases. This should reveal important cues regarding susceptibility to diseases across populations and the improvement of the environment leading to a better public health and social equity. The identification of epi-biomarkers that would predict disease course and treatment response is more difficult to achieve than classical genetic mutations. First, because they may be tissue specific; second because even with highly sensitive PCR approaches, it is still technically challenging to identify changes in the epigenetic profile of a given locus or in the epitranscriptome starting from a few cells. Transgenerational inheritance in mammals is still poorly understood, but if future research demonstrates a broader impact than anticipated, such insight would be important to protect the subsequent generations.

CSIC advantage position and multi/inter-disciplinarity

CSIC is in a good position to address the challenges outlined above due to its multidisciplinary character, including professionals covering different aspects of basic and translational research. CSIC counts with numerous and excellent research groups working on epigenetics from different perspectives from animal models and genomics to structural analysis, drug design and social aspects. The emergence of epitranscriptomics is much more recent, but there are already several groups working on the leading edge for this new area of research. The institution also maintains strong links with some international consortiums and cross-border institutions, which is essential to advance in the knowledge in this field.

CSIC participates in one sixth of the studies related to epigenetic and epitranscriptomic authored in Spain. Since the total contribution of CSIC to Spanish science is close to 20%, we could conclude that this area is underrepresented. The groups working in the epigenetics field are spread in different research institutes all over Spain. Probably, the larger clusters of researchers are located at the Centro Andaluz de Biología Molecular y Medicina Regenerativa (CABIMER, Seville), Instituto de Parasitología y Biomedicina “López-Neyra” (IPBLN, Granada), Centro de Biología Molecular “Severo Ochoa” (CBMSO, Madrid), Instituto de Neurociencias (IN, Alicante) and Instituto de Biología Molecular y Biomedicina (IBMB, Barcelona). According to the information available for the individual institutes and centers that are part of the *Biology and Biomedicine* scientific area, only one center has a department that is solely dedicated to the study of epigenetic mechanisms (IBFG) and only 6 groups, out of a total of 576 groups existing in the area, include the term “epigenetics” in their

description. In fact, consistent with the strategic plans of the different hosting institutes, the research on epigenetics is often oriented to other disciplines, such as parasitology, cancer research, neurosciences or basic biology. A large institute focused on this essential area of research and a better integration of the efforts conducted in this network of CSIC institutes is currently missing. Reinforcement in the emerging field of epitranscriptomics, as well as in specific areas of epigenetic research, such as the epigenomics of autoimmune or metabolic diseases would be desirable to increase the critical mass of CSIC researchers investigating the role of epi-processes on disease mechanisms, as it is underrepresented in the Biomedicine Area. Despite these weaknesses, the CSIC is in an excellent position to lead the research in some specific areas, both leading the answer to outstanding basic biological questions and progressing in the understanding of the epigenetic etiology of specific conditions.

Plan and resources

Most leading international research institutions include strong epigenetics programs in the form of an Institute or of transversal programs. The latter could be the easiest, and cheapest, way for the CSIC to strengthen its position in this field: organizing the groups interested in epigenetics and epitranscriptomics in some sort of "horizontal" platform where to meet, share knowledge and interests, promote cooperative advancement of projects and orientate the future direction of this discipline in the CSIC. This could address some of the main problems detected in this area, such as the lack of knowledge of what resources are available and how to access them (an improvement of the CSIC services database could help mitigate this problem); the difficulties in keeping on the front line with regards of expensive and rapidly evolving equipment; the lack of personnel to run these facilities and assist potential users; and the problems in conserving the know-how due to the highly unstable situation of most CSIC research scientists that rely on short-term contracts. An increase in the number of researchers working in this area throughout new recruitments and the stabilization of highly qualified technical personnel would be required to successfully address the proposed challenges. The collaboration between basic science investigators with bioinformatic groups at CSIC is particularly necessary.

Spain hosts some very prominent research centers working on this area outside the CSIC, such as the Centre for Genomic Regulation (CRG) and the Josep Carreras Leukaemia Research Institute in Barcelona, or the Spanish National Cancer Research Center (CNIO) in Madrid. Particularly important in this area is the CRG and the associated sequencing facilities at the National Center for Genomic Analysis (CNAG). In comparison, our institution is

lagging behind in key technical resources such as next generation sequencing or bioinformatics services. Although some CSIC institutes maintain small sequencing facilities, the creation of a large facility for genome analysis in the frame of CSIC, similar to those at CNAG-CRG, could greatly enhance the investigation in this area. Alternatively, the establishment of some sort of agreement for the participation of the CSIC in CNAG could increase the capacity of both institutions avoiding the duplication of facilities.

In the research of epigenetic mechanisms in etiopathology, one potential difficulty for studies with an important translational component is the need to access specific types of human samples (such as tumor samples, pancreatic islets from diabetic/obese patients, brain tissue from psychiatric patients, iPSCs from patients suffering rare diseases), for which connections with hospitals, tissue banks and specialized services are required. Although group-to-group collaborations with clinical investigators would obviously help, from and strategic point of view finding an appropriate framework by which CSIC could be associated with clinical and translational research based in hospitals on an institutional basis would be important. One possible avenue towards this end would be to potentiate the presence of CSIC in Health Research Institutes (*Institutos de Investigación Sanitaria*) as well as the possibility to participate in projects funded through ISCIII. The incorporation of CSIC groups to the ISCIII Institutes of Biomedicine will allow us to have better collaboration with clinical groups and to have access to additional personnel and financial resources both at the national and international levels. Collaboration of research groups at the CSIC with the pharmacological companies developing epigenetic drugs and with clinicians conducting clinical trials examining these drugs should be also potentiated. In addition, the institution also needs to strengthen its integrations in networks such as Orphanet or the European Reference Networks (ERNs) for rare disease and to potentiate the collaboration of research groups with patient associations, which should enhance the diffusion of results to social stakeholders. Concentrating excellent research in epigenetic and epitranscriptomics at CSIC will not only create a world-leading scientific research community on this topic, but also lead to the establishment of new startups and attract industrial partners.

References

Avraham, R., Haseley, N., Brown, D., Penaranda, C., et al. (2015). Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. Cell *162*, 1309–1321.

Barbieri, I., and Kouzarides, T. (2020). Role of RNA modifications in cancer. Nat. Rev. Cancer.

Bjornsson, H.T. (2015). The Mendelian disorders of the epigenetic machinery. Genome Res *25*, 1473–1481.

Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., et al. (2018). MODOMICS: a database of RNA

modification pathways. 2017 update. Nucleic Acids Res *46*, D303–D307.

Cavalli, G., and Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. Nature *571*, 489–499.

Chang, M., Lv, H., Zhang, W., Ma, C., et al. (2017). Region-specific RNA m(6)A methylation represents a new layer of control in the gene regulatory network in the mouse brain. Open Biol *7*.

Chattopadhyay, P.K., Roederer, M., and Bolton, D.L. (2018). A deadly dance: the choreography of host-pathogen interactions, as revealed by single-cell technologies. Nat Commun *9*, 4638.

Davalos, V., Blanco, S., and Esteller, M. (2018). SnapShot: Messenger RNA Modifications. Cell *174*, 498-498 e1.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

Frye, M., Harada, B.T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. Science (80-.). *361*, 1346–1349.

Harries, L.W. (2019). RNA Biology Provides New Therapeutic Targets for Human Disease. Front Genet *10*, 205.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., et al. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol *33*, 510–517.

Horsthemke, B. (2018). A critical view on transgenerational epigenetic inheritance in humans. Nat Commun *9*, 2973.

Jia, G., Fu, Y., Zhao, X., Dai, Q., et al. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol *7*, 885–887.

Kwon, D.Y., Zhao, Y.T., Lamonica, J.M., and Zhou, Z. (2017). Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. Nat Commun *8*, 15315.

Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., et al. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell *149*, 1635–1646.

Michalak, E.M., Burr, M.L., Bannister, A.J., and Dawson, M.A. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. Nat Rev Mol Cell Biol *20*, 573–589.

Netzband, R., and Pager, C.T. (2020). Epitranscriptomic marks: Emerging modulators of RNA virus gene expression. Wiley Interdiscip Rev RNA *11*, e1576.

Sales, V.M., Ferguson-Smith, A.C., and Patti, M.E. (2017). Epigenetic Mechanisms of Transmission of Metabolic Disease across Generations. Cell Metab *25*, 559–571.

Sanchez-Romero, M.A., and Casadesus, J. (2020). The bacterial epigenome. Nat Rev Microbiol *18*, 7–20.

Sen, P., Shah, P.P., Nativio, R., and Berger, S.L. (2016). Epigenetic Mechanisms of Longevity and Aging. Cell *166*, 822–839.

The Lancet Diabetes, E. (2019). Spotlight on rare diseases. Lancet Diabetes Endocrinol *7*, 75.

Tomasetti, C., and Vogelstein, B. (2015). Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (80-.). *347*, 78–81.

Velasco, G., and Francastel, C. (2019). Genetics meets DNA methylation in rare diseases. Clin Genet *95*, 210–220.

Vicente-Duenas, C., Hauer, J., Cobaleda, C., Borkhardt, A., et al. (2018). Epigenetic Priming in Cancer Initiation. Trends Cancer *4*, 408–417.

Voigt, P., and Reinberg, D. (2013). Epigenome editing. Nat Biotechnol *31*, 1097–1099.

Volkov, P., Bacos, K., Ofori, J.K., Esguerra, J.L., et al. (2017). Whole-Genome Bisulfite Sequencing of Human Pancreatic Islets Reveals Novel Differentially Methylated Regions in Type 2 Diabetes Pathogenesis. Diabetes *66*, 1074–1085.

Wahl, D., Coogan, S.C., Solon-Biet, S.M., de Cabo, R., et al. (2017). Cognitive and behavioral evaluation of nutritional interventions in rodent models of brain aging and dementia. Clin Interv Aging *12*, 1419–1428.

3.6 ENVIRONMENTAL GENOMICS AND EPIGENOMICS

ABSTRACT

Environmental pollution and climate change are greatly influencing all life forms in our planet. It is compelling to understand how pollutant exposure alter the genome, the epigenome and the microbiota in humans, animals and plants, and how microbial populations in nature respond to environmental fluctuations. Integrative genomics, epigenetics and metagenomics can inform the development of environmentally friendly agriculture and livestock solutions, and of new microbe-based biotechnological uses.

PARTICIPATING RESEARCHERS AND CENTERS (IN ALPHABETICAL ORDER)

- Miguel A. Bañares (ICP, Madrid)
- Myriam Calonje (IBVF, Sevilla)
- Mª Carmen Collado (IATA, Valencia)
- Gustavo Gómez (I2SysBio, Valencia)
- Joan Grimalt (IDAEA, Barcelona)
- Antonia Herrero (IBVF, Sevilla, *Coordinator*)
- Paloma Mas (CRAG, Barcelona)
- José M. Pardo (IBVF, Sevilla)
- Lourdes Ramos (IQOG, Madrid, *Deputy Coordinator*)
- Laia Ribas (ICM, Barcelona)
- Federico Valverde (IBVF, Sevilla)

Executive Summary

Environmental pollution and climate change have become one of the most serious threats to humans and other life forms on the planet. It is important to understand how these factors will influence life in our planet. This *Challenge* aims to unravel how the environment interacts with the genome and epigenome to shape the physiology, development and pathology of humans, animals and plants, and how environmental changes impact the evolution of microbial communities in nature.

A crucial need is to understand how pollutant exposure (chemicals and nanomaterials) and environmental changes influence the genome and epigenome throughout the life of an organism. Likewise, it is important to unravel how the genetic signature, including the epigenome, determines the susceptibility of each individual, as well as the transgenerational epigenetic inheritance in response to these threats. Understanding how the fluctuations in environmental conditions are translated into genomic and epigenomic changes will improve our knowledge on the regulatory mechanisms controlling human, animal and plant physiology and development. There is also an emerging need for new solutions for environmentally friendly agriculture and livestock that should be tackled with a better understanding of epigenetics as a source of phenotypic variability and adaptability, and integrative genomics for the discovery of new traits of agronomic interest. Microbial populations have crucial roles in counteracting environmental pollution and in global climate regulation. Metagenomics can provide an in-depth knowledge of the adaptations of microbial communities to changing environments, including imposed perturbations by human activity, which will enable prevention of disastrous impacts and the design of strategies for sustainable growth. Special attention should be paid to oceanic microorganisms, which globally contribute to about half of the total primary production on Earth, and massively impact global geochemical cycles and Earth's climate. Metagenomics in natural environments should include the viral complement, which has been revealed as a decisive factor in microbial population dynamics. In addition, understanding the crosstalk between microbiota and host genome and epigenome will contribute to promote cross-species beneficial interactions and to protect humans, animals and plants against diseases. Finally, functional metagenomics and epigenomics should provide a global repertoire of metabolic capacities and biomarkers, and guide the design of medical and biotechnological applications. The emerging relationships between microbes, humans, animals and plants resulting from the intensive geographical movements in a globally-connected world constitute a major compelling challenge, as observed on occasions of recent pandemics.

Introduction and general description

Environmental pollution and climate change are threatening all forms of life in our planet, including ours. Therefore, it is essential to study and understand the influence that these and other environmental factors have on living organisms in order to minimize and prevent deleterious consequences. More specifically, here we will focus on the interactions that environmental factors frequently have with the genome and epigenome of plants and animals, including humans, and with microbial populations in nature.

Environmental changes in general and exposure to pollutants in particular can damage the genome and/or the epigenome of living organisms. Furthermore, different individuals belonging to the same species might manifest different susceptibilities to damage upon exposure to the same environmental factors depending on their genetic or epigenetic information. Therefore, it is essential to understand the mutual interactions that are established between the environment and the (epi)genome, as this can have a major impact on our current understanding of development, physiology and disease.

Climate change demands new and environmental friendly forms of agriculture and livestock in order to satisfy the needs of constantly increasing human populations. To achieve these goals it will be important to use integrative genomic approaches to uncover and understand new traits of agronomic interest. Likewise, epigenetic mechanisms can be harnessed as source phenotypic variability and adaptability that can be tuned and targeted.

In order to minimize environmental pollution and climate change it is essential to fully understand microbial populations. One major strategy to achieve this is through metagenomics, which can be used to identify new organisms as well as the variability in microbial communities in response to environmental perturbations, particularly those caused by humans. It would be particularly important to apply these metagenomic approaches to the study of oceanic microorganisms, since they remain relatively unexplored despite their major contribution to the total primary production and climate regulation in our planet.

Finally, genomics and epigenomics approaches will also be essential to fully comprehend the interactions that are constantly established between microorganisms and their hosts (plants and animals) and that can have either beneficial or pathological consequences. This is particularly important considering the globally connected world in which we live and that can lead to emerging and unwanted interactions between microbes and other organisms, as sadly illustrated by recent pandemics.

Impact in basic science panorama and potential applications

Environmental pollutants and epigenetic transgenerational inheritance

Environmental pollution could only be counteracted on the basis of a detailed knowledge of the consequences of habitat perturbations on the life of humans and other living beings. The number of chemicals and nanomaterials synthesised by humans during the last century is huge. Many of them have contributed to improve our quality of life by reducing illness or facilitating new technologies and industrial processes. However, a significant number of these products have also been demonstrated to be deleterious for the environment and humans due to their recalcitrant and toxic nature. Accidental exposure to high concentrations of particularly toxic compounds and nanomaterials (e.g. polychlorinated dibenzo-p-dioxins, furans, ZnO, CeO_2) resulted in severe population damage and evidenced transgenerational effects due to their teratogenic and mutagenic character. In addition, chronic exposure to compounds with endocrine disruption activity may alter the endocrine system function(s) and consequently cause adverse health effects in an intact organism, its progeny or (sub)populations. These evidences led to the toxicological evaluation of these and several other chemically-related compounds, including their uses and maximum allowed levels in a variety of matrices (e.g., air, soil, water, food). National and international legislation and agreements have regulated the use of some of these compounds to protect human health and the environment, but these measures are clearly insufficient to prevent the metabolic adverse outcome pathways generated by chemical exposure. A major effort is currently done by the European Commission to assess adverse effect of nanomaterials and establish the appropriate governance. Despite the undoubted value of these regulations and their associated monitoring programs, the consequences of chronic exposure to residual levels of these, and many other still not characterised chemicals and nanomaterials in use, remains essentially unknown. However, recent studies with laboratory animals have pointed out a possible relationship between exposure to environmental toxicants, such as pesticides and plastic components, and abnormal reproductive or metabolic phenotypes that are transmitted transgenerationally (Ost *et al.*, 2014; Nilsson *et al.*, 2012). This new knowledge poses an urgent call for research on how exposure to environmental pollutants and other stressors can induce epigenomic changes in both humans and farm animals that are transmitted to next generations and that associate with disease phenotypes (see also Chapter 3.7 for other examples of epigenetic transgenerational inheritance).

Genomic-based characterization of microbial communities

Often unseen in our anthropocentric view of the world, microbes in nature, and particularly in the oceans, play crucial roles in making our planet a livable one (Falkowski *et al*., 1998). Oceanic picocyanobacteria such as *Prochlorococcus* and *Synechococcus* are the most abundant photosynthetic organisms and the principal primary producers in our planet (Farrant *et al*., 2016), whereas the heterotrophic *Pelagibacter* (SAR11) is the more abundant microorganism in the oceans, and likely in the planet. Some phototrophic microbes of the phytoplankton can fix atmospheric nitrogen, either free living or in symbiosis, such as those formed by diatoms as hosts and cyanobacteria as symbionts (diatom diazotrophic associations, DDAs) (Foster *et al*., 2011; Karl *et al*., 2016). These are globally distributed and perform the crucial task of replenishing nitrogen into the biosphere in a form usable by other living beings, thus facilitating CO_2 fixation. Hence, these microorganisms are main actors in the biogeochemical cycles of carbon, nitrogen, oxygen, phosphorus and iron in our planet, and extensive alterations in the dynamics of their populations would have a serious impact in the trophic chains of marine ecosystems, with consequences for all living beings in the planet.

Human activity, and associated environmental changes, can deeply impact microbial communities in nature, as well as their essential activities. Metagenomics, which consists in massive sequencing of DNA from natural samples and subsequent chromosome reconstruction, can provide a comprehensive view of the structure and dynamics of microbial population in spite of our capacity of cultivation of its members. This is a crucial point taking into account the estimations that only ca. 1 % of microbes inhabiting our planet have been cultivated. Since the pioneering contribution by Venter and collaborators (Venter *et al*., 2004) of whole-genome shotgun sequencing in seawater samples from the Sargasso Sea, diverse collaborative initiatives, such as the TARA Oceans (https://oceans.taraexpeditions.org/en/) (Bork *et al*., 2015) and the Malaspina expedition (http://www.expedicionmalaspina.es/Malaspina/Main.do#content:Home), have been devoted to the study of marine biodiversity and the effects that the global climate change is having on this diversity. These efforts have rendered a wealth of metagenomic data that have sustained unprecedented analysis of the structure of the microbial communities of specific oceanic zones. Also, an increasing number of projects have been directed at the analysis of the structure of the microbial communities in other specific environmental niches and, in some cases, at detecting specific metabolic pathways or gene-product activities (e.g., Dietrich *et al*., 2019). Having an in-depth knowledge of the microbial communities and the physicochemical characteristics of their environment would help to predict the effects of human activity-

imposed perturbations, prevent disastrous impacts, and design controlled communities for profitable biotechnological uses (Duarte *et al*., 2020).

Interactions between the microbiome and the host (epi)genome

Microbial communities inhabit nearly every terrestrial niche, have crucial roles in counteracting environmental pollution and can broadly influence human health and disease. Host and microbial communities have a key intimate relationship that benefits both. Hence, the microbial communities are provided with continuous source of nutrients while the host obtains a wide range of metabolites from bacterial digestion, pathogen and viral protection, and immune system education, among other beneficial functions. Thus, host-microbiota interactors are considered as a single evolutionary and biological unit: the *holobiont*, which represents a highly relevant field in biology and medical sciences (Simon *et al*., 2019). Moreover, community composition is more similar within than between different environments, and interpersonal dissimilarity within habitats is larger than intra-individual variability over time (Näpflin *et al*., 2019). The complexity of the microbial community depends on the particular habitat, and only selected microorganisms will be able to survive and colonize under the conditions characteristic of each habitat. Interestingly, although host-associated microbiota is likely acquired from the surrounding environment, the composition of microbial communities varies greatly from common free-living microorganisms. Furthermore, different factors such as host genetics and environment, including temperature, air pollution, xenobiotics and nutrient resources, shape the microbial composition. In fact, the effect of the exposome is still largely unexplored and the bidirectional interplay between chemical and nanomaterial exposures, the microbial communities and their hosts is only starting to be considered a relevant factor that, among other things, can shape human health and disease.

The main objectives in microbiota studies are to discern how the composition, diversity and functions of the constituent microorganisms influence and regulate host physiology and its association to health and disease. By understanding this, we will be able to learn how to manipulate the microbiome composition and metabolic activities of the microbiome, thus maximizing the health benefits to the host. Advances in this research area have been possible due to the development of different culture-independent genomic technologies based on massive sequencing in combination with innovative bioinformatics and system biology approaches. In addition, understanding the crosstalk between the microbiota and the genome and epigenome of their hosts will contribute to promote cross-species beneficial interactions

and to protect humans, animals and plants against disease. Cumulative evidence demonstrates that the microbiota regulates the host epigenome through specific microbial signals including metabolites, bile acids and other compounds (Sironi *et al*., 2015). Understanding the complex interactions between microbiota, environmental factors and host epigenome, including DNA methylation, histone modification and non-coding RNAs, is a compelling research challenge.

Genomics and Epigenomics in the development of novel agriculture and farming approaches

Humans depend on agriculture and farming for their daily energy intake, whereas plants and animals are increasingly challenged by their environment. As the world population increases, food security is threatened by limited and continuously deteriorated areas of arable land, which affect productivity and product quality. This is further worsened by the increased use of arable land to feed animals due to the growing demand of livestock products, effects that will be potentially multiplied by the devastating consequences of global warming. Therefore, obtaining resilient crops and animals capable to adapt to extreme environments and developing sustainable approaches to increase yield has become an urgent challenge in which genomics and epigenomics will play crucial roles. The knowledge derived from these disciplines will foster breeding programs and unveil key targets for manipulation to develop a-la-carte modified crops and animal cultures.

Key challenging points

I. Exposome: effect on human health of exposure to harmful environmental factors

Since the end of the last century, the incidence of certain non-infectious diseases has increased in human populations worldwide. Some of these diseases, such as obesity, polycystic ovary syndrome (PCOS) or male infertility, have been associated in animal models with exposure to specific environmental stressors (mainly toxic contaminants) and recognized as transgenerationally inherited (Guerrero-Bosagna and Jensen, 2015). This transgenerational inheritance most likely involves epigenetic mechanisms, which are still poorly understood and should be deeply investigated in coming years (see also Chaper 3.7). These findings have raised concerns about the possible effects for humans, and in particular for future generations, of chronic exposure to residual levels of complex mixtures of environmental pollutants. Factors such as intensive farming and industrial development, but also world globalization and climate change, contribute to a constant increase in the number of nanomaterials and chemicals in use. Even nowadays, and despite the high capabilities of state-of-the-art

analytical techniques, unravelling the composition of complex mixtures of organic pollutants present in most environmental and human samples is a challenging task that can be successfully addressed only by experienced laboratories equipped with advanced instrumentation. Similar considerations apply for the evaluation of the toxicity and fate of increasingly used nanomaterials. One major research objective for the future is to understand the molecular mechanisms whereby chemicals and nanomaterials can introduce modifications (genetic or epigenetic) in the DNA, which is also instrumental to better assess their safety and to improve their design. This molecular knowledge can help modelling the toxicity of chemicals and nanomaterials, which will be instrumental for grouping them and enable a capacity to read-across, thus predicting the adverse effects and mechanism. Lastly, proper identification of environmentally-induced (epi)genetic alterations and understanding their etiological relationship with disease phenotypes should allow the design of effective strategies to protect human health and reduce the incidence of non-communicable diseases in future generations.

II. The microbiome and host-pathogen interactions

Despite the great advances in *omic* technologies, the microbiome characterization is a growing field and still to be fully explored. Whole-genome sequencing approaches allow the reconstruction of the clones, characterize known clones and variants, and screen for virulence or resistance genes. Cultivation and 16S amplicon sequencing expanded the knowledge, but still a deeper resolution is needed. The studies of host-pathogen interactions have moved from the study of single genes to whole genome approaches, including both host and microbial genomes. Metagenomics can provide an in-depth knowledge of the adaptations of microbial communities to changing environments, and enable the identification of specific microbial strains. Metagenome-assembled genomes (MAGs) is a recently introduced method that allows microbial genome assembly from metagenomic data, and is providing new insights into the microbial diversity as well as the host-pathogen interactions (Quince *et al*., 2017). Studies with longitudinal sampling and multiple molecular perspectives are necessary to decipher the underlying dynamics and provide novel insights into host-pathogen interaction under specific environmental contexts. To investigate *holobiont* ecosystem dynamics, multi-*omic* approaches including genomics, transcriptomics, proteomics and metabolomics are needed. These approaches would provide a detailed molecular description and new mechanistic insight into the microbiota composition and metabolism, as well as the regulation of the host phenotype through microbiota interactions with the host transcriptome, epigenetic marks and metabolic pathways (Miro-Blanch and Yanes, 2019). Furthermore, due to the difficulty to sort the *omic* information from host-microbiome relationship and to explore microbial diversity at strain resolution, bioinformatics and computational

biology are actual compelling challenges in this field. The biocomputing aspects have been developed in chapter 3.2.

III. Integrative genomics to accelerate trait discovery

Genomics has been contributing to advances in crop and animal culture development for decades. The advent of Next Generation Sequencing (NGS) platforms changed the impact of sequencing on our knowledge of genomes and gene regulation. Once a genome sequence is available, all genes and genetic variants that contribute to agronomic traits can be identified, and changes made during the breeding process can be evaluated at the genotype level. Genome sequencing has become an initial step for ascertainment of the genome and evolution, while ensuing resequencing steps allow elucidating genetic variability among individuals. Nevertheless, there are still important limitations in NGS, which leads to highly fragmented genome assemblies that complicate the analysis. These problems can now be solved with the emergence of Third Generation Sequencing (TGS) technologies that enable the generation of long reads and allows the production of more accurate and contiguous genome assemblies (Jiao and Schneeberger, 2017; Korlach *et al*., 2017), therefore facilitating genomic studies.

Several genomic approaches have been applied to accelerate the detection of gene-trait associations. As plants and animals evolve in complex environments, gradually acquiring the ability to cope with different environmental conditions, a high number of desirable traits or phenotypes are defined as complex quantitative traits. Quantitative Trait Loci (QTL) studies have been used to identify regions of the genome that co-segregate with a given trait (*Hu et al., 2018*). However, QTL mapping suffers from two fundamental limitations: its limited resolution and the fact that only allelic diversity that segregates between the parents of a segregating population can be assayed (Korte and Farlow, 2013). Over the last years, Genome-Wide Association Studies (GWAS) connected with whole-genome sequencing strategies have evolved into a powerful tool to reconnect traits back to their underlying genetics, overcoming QTL limitations (Korte and Farlow, 2013). These new GWAS strategies provide higher resolution to identify multiple recombination events and can be employed to pinpoint genomic mutations linked to traits in diverse and unrelated populations, thus, allowing to explore the natural variations associated with phenotypic differences. Integrating GWAS across multiple studies, and combining their data with other high-throughput techniques such as transcriptomics, proteomics and metabolomics will accelerate the detection

of robust gene-trait associations and help to understand complex phenotypic traits. On the other hand, paleogenomics aims to reconstruct ancient genomes by direct sequencing of fossil material or ancestors of actual crops, thus helping to understand crop domestication and predicting how future populations will evolve in response to global warming (Pont *et al*., 2019). It also allows to introduce in modern breeding programs ancient traits lost during the natural breeding history, such as stress resistance, color or flavor.

IV. Epigenetics as a new source of increased variability and adaptability

Intensive breeding programs have reduced the genetic diversity in crops and livestock. Recent completion of genome sequencing has increased breeding efficiency by providing new tools that have helped to identify a substantial proportion of the inheritable sequence-based phenotypic variation. However, the sequence variability of genes that control agronomic traits cannot explain the full spectrum of phenotypic diversity observed in plants and animals, and there is still a significant proportion of unexplained heritability. Recent evidence indicates that epigenetic variation can explain the heritability of complex traits (Cortijo *et al*., 2014). Furthermore, studies in model organisms have provided a large amount of information on the implication of epigenetic mechanisms in the regulation of development and in the response to different abiotic and biotic stresses, thus playing an important role in shaping phenotypic plasticity.

The ability of a single genotype to express multiple phenotypes in response to different environments, either external or internal, is widespread amongst both animals and plants. Plasticity is generally considered to be adaptive and/or advantageous for sessile organisms that have to adapt in place to environmental conditions, and can be exploited to breed for more resilient crops. On the other hand, inclusion of plasticity in animal breeding models (e.g., sexual plasticity in cultured fish species) will be important to breed for increased robustness of animals, or in breeding programs (e.g, brood stock animals) that produce genetic material for a range of production environments. Nevertheless, our ability to harness animal and plant plasticity requires a better understanding of the underlying epigenetic mechanisms and to define epigenomic states.

Epigenetic information is mostly mediated by DNA methylation and histone posttranslational modifications, the so-called chromatin marks that alter the reading and writing of the genome, resulting in the regulation of chromatin architecture, gene activity, and expression without changes to the DNA sequence. While analyses of epigenetic regulation of

gene expression date back to the 80s, methods to analyze epigenetic modifications at a genome-wide scale were not developed until the early 2000s. Advances in DNA sequencing technology, the development of methods such as bisulfite sequencing and chromatin immunoprecipitation sequencing, and generation of highly specific antibodies against post-translationally modified histones have created the opportunity to generate epigenomic maps (Gallusci *et al*., 2017). However, it is important to take into consideration that each genome can give rise to a large number of "epigenomes", which tend to be tunable and highly dynamic. Therefore, there is a need to describe the epigenomic "ground state" at different developmental stages in different animal and plants species, and how this state changes in response to the environment. It is also important to determine whether trait-associated variants are enriched in tissue-specific epigenetic signatures. The generation of genome-wide maps of histone marks and the methylome within specific cells, tissues and organs under varying environmental conditions will pave the way to uncover the implication of particular cell types and tissues in specific traits. It would be also crucial to search for novel epigenomic marks to fully understand the diversity of epigenomic modifications that might need to be considered. Therefore, the overarching goal now is to conduct basic and applied research on how epigenetic/epigenomic processes contribute to development and response to environmental conditions. The data obtained from these studied should be included in integrative epigenomic databases. This will allow to conduct Epigenome-Wide Association Studies (EWAS), as it is done with humans (Verma, 2012), using different genotypes/tissues/cell-types/environmental conditions, which can provide valuable inputs for the development of epimarks that can be used in crop and animal improvement.

V. Understanding the epigenetic memory

Stress responses in animals and plants can be entrained and primed by prior stress episodes (Chang et al. 2020). The molecular basis of this stress memory is largely epigenetic and often transgenerational; i.e., acquired tolerance can be passed on to the progeny. Future challenges to harness epigenetic regulation of gene expression to give rise to more resilient crops and animal cultures will include deciphering the code of the stress-induced epigenetic landscape, how the environmental cues are translated into specific epimarks, and learning to preserve desirable epigenetic modifications throughout successive generations or to erase undesirable ones. Answering these questions will be essential for understanding the stability, reversibility, and heritability of epialleles, and the use of epigenetic engineering to improve resilience and

productivity of plants of agronomical interest. Identifying the molecular components connecting the environmental changes with the epigenome will provide novel targets susceptible of epigenetic engineering for improved resilience. Entrained stress avoidance by sensitized animals and plants often entails reduced growth rate and smaller sizes, traits that can be transferred epigenetically to the offspring. Efforts should be made to understand how transgenerational epigenetic changes curtail the full yield potential of agriculture and farming as to avoid the stress-avoidance syndrome (Maggio *et al*., 2018).

VI. Microbial genomics and environmental implications

A major impact of human activities in natural environments is the global warming in oceans, which likely will condition our planet suitability for life in the medium- and long-term (Cavicchioli *et al*., 2019). A predicted increase of the temperate zones of the oceans will affect the distribution of microorganisms, favoring those adapted to oligotrophic intertropical zones in detriment of others less tolerant to high temperatures. Concomitant with temperature increase, acidification in the oceans, as a consequence of increased CO_2 concentration in the atmosphere, can greatly alter the composition of marine phytoplankton (e.g., effects on many algae that include calcium carbonate structures), and impact essential parameters for oceanic productivity, such as incident light or nutrient recycling. Key developments would be the implementation of the capacity to analyze microbial communities by metagenomics on numerous niches, which should include open oceans, but also coastal and terrestrial environments, the improvement of the capacity of detection of all varieties of microorganisms in the samples, and the capacity to ascribe the participation of each community member to the biological activities detected in each environmental niche. Together, the implementation of these practices would provide invaluable capacity for assessing the extent and direction of microbial responses to human activities. Finally, in marine ecosystems, viruses capable of infecting bacteria and other microorganisms greatly exceed in number those of their hosts, and likely have a large impact in the dynamics of the populations of primary producers. This is a rather poorly characterized effect that deserves further attention and that can also be investigated using metagenomic approaches.

VII. Biotechnological developments by microbial genomics

Microbes are already used for a wide repertoire of biotechnological uses that include the production of high added-value chemicals, the detoxification of contaminated environments

and water treatments. As an example, the Spanish enterprise PharmaMar is devoted to the investigation of marine resources for the pharmaceutical industry. Metagenomics could allow the reconstruction of community-structured metabolic pathways, as well as the identification of new enzymes and reactions. Comprehensive dissection of microbial activities represents an unprecedented opportunity for the identification of new metabolic pathways, new enzymatic activities that can be used for the design of biotechnologically-relevant reactions and networks, and natural products with interest for humankind. For this, a challenge is represented by the development of procedures to generate shotgun expression libraries derived from metagenomic information, as well as to reconstitute pathways for activity screening. Together, an in-depth knowledge of microbial communities in nature and their responses to human perturbations will become essential for the mitigation of pernicious impacts and for the sustainable utilization of the ocean resources.

CSIC advantage position and multi/interdisciplinarity

Despite the complexity, novelty and interdisciplinary nature of most of the previously mentioned challenges, the CSIC holds a key position to become a reference institution in many of these research fields at international level in the short and medium-term. Several CSIC research groups are internationally recognized by their research activities in the previously mentioned fields. The impact of their research studies, published in high impact peer-reviewed scientific journals, and their invitation to be part of decision-making committees at national and international levels are considered objective and illustrative evidences of worldwide scientific recognition. Studies in these emerging but already active research areas will certainly continue, alone or, in most cases, in collaboration with highly renowned national and international groups. Also, new challenges will be successfully addressed and effectively solved in the next decades, thereby contributing to maintain and enhance the international visibility and heft of CSIC.

In CSIC, there are several research groups that work in disciplines related to genomics and epigenetics with the future aim to improve agriculture and farming production. Some of them are: CABD, IBVF, CRAG, CNB, IBMCP, ICM, CBM, CIB, I2Sysbio, IATA, EEZ, ICMAN, UCL-JCCLM, IGM-ULE, IHSMIATS, EEZ, and ULE. The molecular understanding of key events for DNA modification triggered by chemicals and nanomaterials is an emerging field already addressed in IDAEA and ICP. IDAEA and IQOG have a leading European position in the study of chemical compounds with known capacity for generation of metabolic adverse outcome pathways and epigenetic modifications in human DNA. Furthermore, due to the increasing biological importance of epigenetics and genomics, it is

foreseen that other CSIC groups currently working in environmental pollution, human, animals and plants may include these disciplines in their coming research projects. Further, all the challenges identified require inter and multidisciplinary approaches, which cover different areas of knowledge. Thus, to address the correct interpretation of input data from *omic* studies, there is a need of cross-interactions between several CSIC groups; e.g., those with human, animal and plant physiological expertise, or those having molecular and computational expertise. Also, the IBVF include research groups with high expertise in molecular biology and genetics of photosynthetic microorganisms, algae and cyanobacteria, as well as their use for biotechnological applications. These groups can represent the germ of future projects on microbial environmental genomics.

Plan and resources

I. Facilities and services to develop genomic and epigenomic studies

The maintenance and enhancement of the current international position of CSIC in the previously mentioned research lines during the next decades will only be possible through the acquisition and maintenance of instruments that allow to retain and increase the current quality levels of the CSIC research groups in this highly competitive scenario. In particular, resources needed include powerful centralized services for massive sequencing and proteomics and state-of-the-art separation-plus-detection instruments.

II. Accessibility, database creation and computational tools

Due to the massive amount of data created by *omic* studies, in the next few years there will be an exponential increase in computational data analysis (e.g., petabyte, PB= 10^{15} bytes). Thus, there is a need to create curated findable, accessible, interoperable, reusable (FAIR) massive database facilities able to storage bioinformatic data; repositories that are required for peer reviewed journals and scientific organizations, and which will enable further research based on those curated data. In addition, it is required to generate a "supercomputing center" able to handle the petabytes of information generated by this kind of data. In this public center, data among researchers could be easily shared and accessed, thus favoring a close collaboration between CSIC groups.

III. Personnel resources

Together with the availability of instruments and tools, the incorporation of experienced researchers, which can provide new or complementary knowledge to many of the above considered areas of expertise, as well as contribute to strength the currently weakened scientist network, appears mandatory. Personnel needs cannot disregard the urgency of counting on specialized technicians that could be permanently incorporated to the research groups as well as the centralized services.

IV. Communication strategy

Communication of the main research conclusions to decision-making authorities and to the general public should also become mandatory in the next decades, not only due to our compromise with knowledge communication, but also as an efficient strategy to control some of the possible environmental factors affecting microbial populations and human, animal, and plant health.

References

Bork P. Bowler C. de Vargas C. Gorsky G. Karsenti E. Wincker P. Tara Oceans studies plankton at planetary scale. *Science* 2015; 348: 873-873.

Cavicchioli R. Ripple WJ, Timmis, K.N. Azam F. Bakken L.R. Baylis M. *et al.* Scientists' warning to humanity: microorganims and climate change. *Nature Rev. Microbiol.* 2019; 17: 569-586.

Chang Y.N. Zhu C. Jiang J. Zhang H. Zhu J.K. Duan C.G. Epigenetic regulation in plant abiotic stress responses. *J. Integ.r Plant. Biol.* 2020; 62: 563-580.

Cortijo, S. Wardenaar R. Colomé-Tatché M. Gilly A. Etcheverry M. Labadie K. *et al.* Mapping the epigenetic basis of complex traits. *Science* 2014; 343: 1145–1148.

Dietrich J. Schmidt E., Kotze D.J. Hornung E. Setälä H. Yesilonis I. *et al.* Metagenomics reveals Bacterial and Archaeal adaptation to urban land-use: N catabolism, methanogenesis, and nutrient acquisition. *Front. Microbiol.* October 2019 | Volume 10 | Article 2330

Duarte C.M. Agusti S. Barbier E. Britten G.L. Castilla J.C. Gattuso J.P. Rebuilding marine life. *Nature* 2020; 580: 39-51.

doi: 10.1186/s13148-014-0043-3. eCollection 2015

Falkowski P.G. Barber R.T. Smetacek,V. Biogeochemical controls and feedbacks on ocean primary production. *Science* 1998; 281: 200-206.

Farrant G.K. Dore H. Cornejo-Castillo F.M. Partensky F. Ratin M. Ostrowski M. *et al.* Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc. Natl. Acad. Sci. USA* 2016; 113: E3365-3374.

Foster R.A. Kuypers M.M.M. Vagner, T. Paerl R.W. Musat N. and Zehr J.P. Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *ISME J*. 2011; 5: 1484–1493.

Gallusci P. Dai Z. Génard M. Gauffretau A. Leblanc-Fournier N. Richard-Molard C. Vile D. Brunel-Muguet S. Epigenetics for plant improvement: Current knowledge and modeling avenues. *Trends Plant Sci.* 2017 22: 610-623.

Guerrero-Bosagna C. Jensen P. Globalization, climate change, and transgenerational epigenetic inheritance: will our descendants be at risk? *Clinical Epigenetics* 2015; Jan 22;7(1):8.

Hu H. Scheben A. Edwards D. Advances in integrating genomics and bioinformatics in the plant breeding pipeline. Agriculture 2018; *8*: 75.

Jiao W-B. Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* 2017; 36: 64–70.

Karl D.M. Church M.J. Diore J.E. Letelier R.M. Mahaffey C. Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Pro. Natl. Acad. Sci. USA* 2016; 109: 1842–1849.

Korlach J. Gedman G. Kingan S.B. Chin C. Howard J.T. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* 2017; 6: 1–16.

Korte A. Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 2013; 9: 29.

Maggio A. Bressan R.A. Zhao Y. Park J. Yun D.J. It's Hard to avoid avoidance: Uncoupling the evolutionary connection between plant growth, productivity and stress "Tolerance". *Int. J. Mol. Sci.* 2018; 19.

Miro-Blanch J. Yanes O. Epigenetic regulation at the interplay between gut microbiota and host metabolism. *Front. Genet.* 2019; 10: 638. doi: 10.3389/fgene.2019.00638.

Näpflin K. O'Connor E.A. Becks L. Bensch S. Ellis V.A. Hafer-Hahmann N. Harding K.C. *et* al. Genomics of host-pathogen interactions: challenges and opportunities across ecological and spatiotemporal scales. *PeerJ.* 201; Nov 5: 7:e8013. doi: 10.7717/peerj.8013. eCollection 2019.

Nilsson E., Larsen G. Manikkan M. Guerrero-Bosagna C. Savenkova M.I. Skinner M.K. Environmentally induced epigenetic transgenerational inheritance of ovarian disease. *PLoS ONE* 2012; 7: e36129.

Ost A. Lempradt A. Casas E. Weigert M. Tiko T. Deniz M. *et al.* Peternal diet delfine offspring chromatin state and intergenerational obesity. *Cell* 2014; 159: 1352-64.

Pont C. Wagner S. Kremer A. Orlando L. Plomion C. Salse J. Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biology* 2019; 20: 29.

Quince C. Walker A.W. Simpson J.T. Loman N.J. Segata N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 2017; 35: 833-844.

Simon J. Marchesi,J.R. Mougel, C. *et al.* Host-microbiota interactions: from holobiont theory to analysis. *Microbiome* 2019; 7: 5.

Sironi M. Cagliani R. Forni D. Clerici M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* 2015; 16: 224-36. doi: 10.1038/nrg3905.

Verma, M. Epigenome-wide association studies (EWAS) in cancer. *Curr. Genomics* 2012; 13: 308-313.

Venter J.C. Remington K., Heidelberg J.F. Halpern A.L. Rusch D. Eisen J.A. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004; 304: 66-74.

3.7 EPIGENOMICS AND LIFE STYLE

ABSTRACT

That epigenetic changes during lifetime (i.e. those with a biological purpose, epigenetic drift and epigenetic clocks) depend on a complex mixture of factors is well-known, as is the fact that, depending on the loci, they can be modulated by genetic and/or external factors, lifestyle being one of the most important. However, the underlying molecular mechanism remains largely unknown and addressing this will be an important challenge for CSIC in this field.

PARTICIPATING RESEARCHERS AND RESEARCH CENTERS (in alphabetical order)

- Laura Bravo Clemente ((ICTAN, Madrid)
- Carmen Collado (IATA, Valencia).
- Mario Fernández Fraga (CINN, Oviedo, *Coordinator*).
- Raúl Fernández Pérez, (CINN, Oviedo).
- Belén Gomara (IQOG, Madrid).
- Rocío González Urdingio (CINN, Oviedo).
- María Ángeles Martín Arribas (ICTAN, Madrid).
- Sonia Ramos Rivero (ICTAN, Madrid, *Deputy Coordinator*).

Executive Summary

The past two decades have served to consolidate the concept that epigenetic modifications change during lifetime. Certainly, numerous scientific works have described several epigenetic signatures associated with increased age and, remarkably, epigenetic "clocks" have been recently developed which can predict biological age by measuring levels of epigenetic marks such as DNA methylation. These clocks also serve as biomarkers of physiology because they are altered in pathological states. However, although these predictors of chronological age have been well characterized, the underlying molecular mechanisms are still largely unknown. There is thus an urgent need to describe these molecular mechanisms and to determine whether epigenetic clocks are a cause or consequence of the increase of age.

In recent years, numerous (mainly descriptive) works have proposed that epigenetic changes during lifetime can be modulated, at least in part, by external stimuli. It has also been proposed that lifestyle during pregnancy can program the epigenome during embryonic development and determine certain disease phenotypes in adult life. If true, this would imply that lifestyle might shape health and disease phenotypes through epigenetic mechanisms. Moreover, at present it is unclear whether transgenerational epigenetic inheritance or intrauterine exposures influence offspring's health and disease susceptibility. Therefore, the role of epigenetics in transgenerational inheritance requires to be explored.

Diet is one of the most important factors in lifestyle and has the potential to modulate gene expression programs through epigenetic mechanism, thus affecting individual health and life expectancy. A challenge for the future will be to determine both how nutrients and bioactive components in food influence epigenetic function and the underlying molecular mechanisms involved. This will be important in order to describe new effective nutrition-based preventative and therapeutic approaches. In this way, nutritional epigenetics can be instrumental in developing personalized programs that contribute to reducing the risk of disease and improving health. Moreover, in addition to beneficial nutrients, foods might also contain variable amounts of pollutants which exhibit different mechanisms of toxicity and bioactivity. Thus, nutrigenomic studies in the future should consider not only the beneficial effects of nutrients but also the possible harmful effects of contaminants present in foodstuffs.

Another central component of lifestyle is physical activity. Indeed, it is well known that regular exercise influence health and prevent disorders, such as cardiovascular and metabolic diseases or cancer. As with diet, epigenetic mechanisms have also been proposed to be the molecular link between these health benefits and physical activity. However, as most of the studies are focused on specific candidate genes in target tissues, it is necessary to carry out

extensive genome-wide analysis of epigenetic modifications in the future. It will be also necessary to identify specific epigenomic signatures associated with different types of physical activity (e.g., gentle aerobics, strength or endurance activity, etc.).

In addition to diet and physical activity, other important aspects of lifestyle, such as alcohol and recreational drugs use, pharmacological treatments, etc., must also be taken into consideration. For all these factors, future research should not only be limited to describing epigenetic changes in response to a given environmental cue, but also to trying to identify the functional and physiological consequences of these changes.

Finally, because epigenetic changes are in principle reversible, they provide an avenue for the development of therapies to counter complex processes, such as aging. Among the challenges for the future are the identification of the alterations involved, the separation of the biologically relevant changes from the rest of the environmentally-induced noise, the design of the biological tools to reprogram epigenetic alterations and, importantly, the development of instruments, such as epigenetic clocks, which will allow us to ascertain whether our interventions have an effect on our life expectancy with a healthy aging.

Introduction and general description

Epigenetic changes during lifetime (epigenetic drift vs epigenetic clocks)

Aging is a universal phenomenon in which biological functions gradually decline, ultimately leading to death. At the cellular level, these changes influence a wide array of molecular pathways and include both genetic and epigenetic alterations (López-Otín et al., 2013). Indeed, epigenetic mechanisms may, at least in part, mediate aging features such as genome instability or transcriptional noise (Pal and Tyler, 2016). These alterations have been described at all levels of epigenetic regulation, with DNA methylation and histone modification being those most widely studied, both in human and model organisms (Pal and Tyler, 2016). DNA methylation, in particular, has been investigated in humans, and this has led to some accepted notions regarding the gradual changes observed during lifetime: a global loss of DNA methylation at intergenic and repetitive regions, local gains at CpG-dense regions and, in general, an increase in intra- and inter-individual variability in DNA methylation patterns (Huidobro et al., 2013; Jones et al., 2015). This last observation, commonly referred to as "epigenetic drift", underscores the idea that the epigenetic changes observed during the aging process are comprised of a combination of functional and stochastic alterations, which in addition may or may not be significant in the regulation of gene expression (Tejedor and Fraga, 2017). Because epigenetic alterations can have innumerable origins, including environmental elements such as lifestyle factors, it is crucial to be able to separate specific and non-specific variations (Feil and Fraga, 2012).

Within this scenario, owing to the development of genome-scale microarray and NGS (next generation sequencing) technologies, the concept of the "epigenetic clock" has recently emerged in the field. Epigenetic clocks consist in mathematical models which use DNA methylation information to predict chronological age with an unprecedented level of precision (Horvath and Raj, 2018). Moreover, the consistent behavior across lifespan of the CpG sites involved sets them apart from those implicated in age-related epigenetic drift (Jones et al., 2015). Both aging-associated epigenetic changes and the recently characterized epigenetic clocks may help explain the molecular mechanisms involved in defining the phenotypic variability observed between individuals at the physiological as well as the pathological level. In order to do so, however, they need to be effectively disentangled from stochastic epigenetic variation.

Nutrigenomics and nutriepigenomics

Nutrition is considered one of the most impacting life style factors able to affect the genome and epigenome. Nutrigenomic describes the interaction between nutrition and genes to understand how specific food constituents or dietary regimes may affect human health. The novel discipline of nutriepigenetics integrate the knowlege of nutrigenomic and the effect of diet or dietary compounds in gene expression programs through epigenetic mechanisms. Diet, foods and its components affect the genome and epigenome and have the potential to modulate critical metabolic pathways influencing individual health and life expectancy. A link between certain dietary patterns and diverse non-communicable diseases (cancer, cardiovascular disease, obesity, diabetes) has been suggested. These associations constitute a starting point about how diets and ultimately foods might modify the genome and epigenome and, consequently, the proteome and metabolome. Therefore, each dietary pattern might provide diferent epigenetic and genetic signatures, which could be associated to a healthy status or disease susceptibility. Nevertheless, nutritional epigenomics is a quite recent research area and present data on the precise effects of diets, foods or food components on the epigenome are very limited. At present, most of these studies are descriptive and the epigenetic modifications as well as the underlying molecular mechanisms remain still largely unknown. In addition, most of nutritional epigenetic works have focused on DNA methylation, while post translational histone modifications and miRNA expression are even less analyzed.

Pollutants present in foodstuffs

Air and dust inhalation and food ingestion are considered the main routes of exposure to toxic chemical compounds for the general population. In the case of lipophilic compounds, the ingestion of food is the most important source of exposure. Many toxic environmental pollutants with endocrine-disrupting properties are lipophilic. Therefore, they bioconcentrate in living organisms and bioaccumulate through food webs making humans the organisms receiving the highest impacts. In addition, agrochemicals used for pest control, veterinary drugs employed on farming, migrating compounds related to food contact materials (FCMs), food additives (ranging from food colorings, preservatives, and stabilizing agents, to bioactive compounds), and contaminants introduced or formed during food storage (such as mycotoxins and biogenic amines) and processing (like acrylamide) reach humans through the diet. Many of these compounds have been demonstrated to be toxic and exert epigenetic modulation.

Precision nutrition in human health and disease prevention

Nutrition plays a central role in the prevention of many chronic pathologies, with diet being a key modifiable factor that may influence the incidence of highly prevalent metabolic disorders, both monogenic (e.g. celiac disease, lactose intolerance) and polygenic (e.g. type 2

diabetes, obesity, metabolic syndrome, cardiovascular diseases), or certain types of cancer. To complicate things fuerther, some diseases may be associated to both monogenic and polygenic risk factors. For instance, obesity, which incidence is steadily increasing and is associated to many comorbidities (diabetes, dyslipemia, hypertension, inflammation, etc.), may represent a symptom of up to 40 monogenic diseases and chromosomal abnormalities, but obesity may also depend on numerous genetic variants, with more than 600 genes and DNA regions associated to human obesity by GWAS. This shows the complexity of studying multifaceted polygenic traits related to numerous physiological pathways.

Besides, the potential therapeutic role of nutrition in the prevention of chronic degenerative diseases is multiple and complex. Not only nutrients in foods, but especially bioactive food components (polyphenols, carotenoids, phytosterols, isothiocyanates, glucosinolates, ω-3/ω-6, etc.) may have a marked effect on health. Many of these bioactive compounds will have actions at various molecular levels, from DNA expression, pre-transcriptional modifications, affecting protein functionality, metabolic processes, etc. These nutri(epi)genomic implications of food components are also accompanied by nutri(epi)genetic factors affecting crucial steps as eating preferences, ADME (absorption, distribution, metabolism and excretion), metabolic pathways and, in the end, the individual's phenotype and clinical response. In addition, the gut microbiota may be affected by these food components and modify their activity through catabolic modifications of the ingested molecules.

All this results in large variability in the individual response to diets and foods, and the need for an integrative molecular and -omic approach to nutrition. Like precision medicine, precision nutrition also requires an individual approach to the person based on the 4P principles (personalized, predictive, preventative, participative) for which genetics becomes essential.

The impact of diet–microbiota and interactions in precision nutrition

Nutrition plays a relevant role in human health. Molecular nutritional research has been defined as "the science that studies the effects of nutrients, food and its components, on the whole physiology and the state of good health at the molecular and cellular level". In the future, nutrition research should progress beyond the one-size-fits-all diet towards the study of the personalized host response to diet. This concept needs to integrate both biological (microbiome, epigenome, metabolome, genome, etc.) and environmental variables (diet, physical activity, drug intake, xenobiotics, infections, stress, etc.) to obtain detailed

predictions of the individual's responses to specific nutrients and other dietary bioactive compounds.

Physical activity

The beneficial effect of physical activity on health is well known (Fiuza-Luces et al., 2013; Neufer et al., 2015). A physically active lifestyle and regular exercise help to regulate blood pressure and metabolism, modulate homeostasis, and generally contribute to improving health and preventing diseases(Booth et al., 2012). Lack of physical exercise influences health throughout life, and is associated with higher premature mortality, coronary heart disease, type 2 diabetes, colon and breast cancer, as well as obesity (Booth et al., 2012; Nieman et al., 2019; Fernandez-Sanles et al., 2020). The benefits of exercise on memory and cognition are also known (Hillman et al., 2008; Fernandes et al., 2017), which highlights its importance in the maintenance of physical and mental health during life.

Although the biological mechanisms that regulate the beneficial effects of exercise on health are not fully understood, an important role for epigenetic marks has been proposed, because epigenetics can represent the molecular link explaining how the environment affects our genes (Feil and Fraga, 2012). Epigenetic changes associated with physical activity have been studied at different levels in terms of the tissue type and type of epigenetic mark analyzed, although they have mainly been studied in blood cells, in muscle and adipose tissue, and in brain tissue in murine models (Elsner et al., 2011; Seaborne et al., 2018; Nielsen et al., 2010). Within this context, DNA methylation is the most studied epigenetic mark, both at the global and the gene-specific level (Fernandez-Salnes et al., 2020; Seaborne et al., 2018a; Ronn et al., 2013; Schenk et al., 2019), followed by miRNAs and posttranslational histone marks (Elsner et al., 2011; Nielsen et al., 2010; Pandorf et al., 2009; Melo et al., 2015).

The emergence of -omic technologies in general, and epigenomics in particular, open up new avenues to study in more detail the effects of exercise on the epigenome, and especially the possible functional effects that these epigenetic changes may have on health.

Alcohol and drugs of abuse

In the same manner as our diet, during our lives we find ourselves exposed to many other types of compounds. The National Institute of Statistics in Spain estimated in 2017 that more than 20% of people aged 15 or older were daily smokers, and a similar proportion of citizens drank alcohol on a weekly basis (Instituto Nacional de Estadística, 2017). Alcohol and tobacco have well-known and wide-ranging consequences on our health, and thus it should

come as no surprise that their consumption has been linked to epigenetic changes. Because of their legal status and prevalence in the population, most of the research has been carried out on these two substances. Regarding tobacco, it has to be taken into account that the thousands of different hazardous compounds present in its smoke can affect a myriad of pathways (Talhout et al., 2011), while alcohol effects pertain to those of ethanol metabolism.

Impact in basic science panorama and potential applications

Epigenetic changes during lifetime (epigenetic drift vs epigenetic clocks)

The characterization of epigenetic changes with aging in the context of epigenetic drift is only starting to be addressed. A significant decrease in the cost of microarray technologies has allowed the development of large-scale studies, which can better capture more subtle epigenetic changes within the scenario of noisy inter-individual variability. In this same vein, large-cohort studies have revealed that DNA methylation changes in variability are linked to aging-associated molecular pathways (BIOS consortium et al., 2016), thus providing an explanation of the processes involved in aging, and have also served to identify particular genetic *loci* which may provide key avenues for anti-aging intervention (McCartney et al., 2020).

In addition, epigenetic clocks also show great potential as tools in the investigation of aging and aging-related disease. Epigenetic age has been shown to be accelerated in association with a wide range of pathologies, a significant observation which serves to show that the clocks can be used as biomarkers of disease (Horvath and Raj, 2018). The fact that same-age individuals manifest different epigenetic ages points towards the notion that these models are capturing, at least in part, "biological aging", which can be thought of as the general physiological status in relation to mortality risk and comorbidities (Bell et al., 2019). As such, a new generation of clocks focused on biological age is already being developed to better capture disease associations (Lu et al., 2019; Levine et al., 2018). Perhaps the most promising of their applications is the fact that they could be used to evaluate the success or failure of anti-aging interventions.

Nutrigenomics and nutriepigenomics

Future work in the field of nutrition and epigenetics has the potential to provide significant benefit to public health. Deciphering the epigenetic signatures triggered by bioactive food components might lead to personalized nutritional interventions that takes into account genetic/epigenetic information. Nutritional epigenetics represent a safe potential and

innovative strategy for the prevention or treatment of many prevalent chronic diseases that are close related to epigenetic modifications.

Pollutants present in foodstuffs

The presence of pollutants in foodstuffs is routinely controlled by the sanitary authorities, to ensure food safety. However, research on the effects of these substances at the epigenetic and genomic level is still rather limited, although mandatory for proper food legislation and regulation. An additional difficulty associated to the presence of industrial and agricultural pollutants and veterinarian drugs is that, when the toxicity of these chemicals is established, they are rapidly substituted by alternative products, making previous routine controls and toxicity data render obsolete and forcing the scientific community and food authorities to develop new analytical methodologies and strategies for their identification, determination, risk evaluation, and routine monitoring.

Therefore, the future of nutrigenomic goes through considering not only the beneficial effects of nutrients but also the possible harmful effects derived from either natural or anthropogenic contaminants present in foodstuffs. This is a complex challenge since many known contaminants have been associated with a high incidence and prevalence of different endocrine-related disorders in humans, but also because of the constant introduction in the food web of new substances that can also impact the (epi)genome.

Precision nutrition in human health and disease prevention

There is an increasing interest in precision nutrition for its potential in the prevention and treatment of chronic non-communicable diseases, both monogenic and polygenic. GWAS and other genetic studies have identified over 15000 SNPs associated with numerous pathologies and traits. Besides genetic polymorphisms, epigenetic modifications, which are tissue-specific, highly affected by environmental/lifestyle factors (including the diet), and reversible but also transgenerational inheritable, complicates the differential susceptibility and responsiveness of individuals to diet-related pathologies and dietary interventions. Along with these factors, ncRNA also play an important role, since lncRNA and specially miRNA shed in extracellular vesicles can have paracrine or endocrine-like effects in different target tissues and organs. In line with this, the role of ncRNA ingested with foods should also be considered as potentially important external epigenetic modifiers, since in general the impact of ncRNA on metabolic pathways and regulatory networks is still little known.

Bearing this in mind, the identification of genetic variants or epigenetic marks that predispose individuals to suffer from certain metabolic-related diseases is key for precision medicine and precision nutrition alike to estimate disease risk and design preventive strategies. Similarly, recognizing relevant gene-diet interactions will allow to identify responsive and non-responsive individuals for specific dietetic interventions, which in turn would permit designing personalized recommendations to maximize the benefit of nutritional interventions. This genotype-directed nutrition will be useful not only for individual personalized dietary advices, but will also improve public health recommendations and the design of nutrition solutions, including functional foods and nutraceuticals.

The impact of diet–microbiota and interactions in precision nutrition

Our diet and lifestyle, as well as other habits (e.g. sleep patterns) and exposures (e.g. stress, pollutants, blue light, food chemicals and plastics), have changed dramatically over the last few decades. Exposure to unhealthy food, inadequate dietary patterns, such as excess or deficiency of macro- and micronutrients, may increase the risk of non-communicable diseases (NCDs). This exposure may act through several potential mechanisms, including the modulation of the microbiome, but also affecting the metabolome and epigenetic regulation, cellular and physiological routes and the immune system, which affect host response and health, thereby increasing the risk of NCDs. Most of these mechanisms remain unclear, and thus there is an urgent need to generate scientific evidence through human studies.

A strong link has been observed between microbiota dysbiosis and the risk of NCDs, such as allergies, obesity, diabetes, immune-related problems and cardiovascular diseases, which have resulted in an increasing global burden that requires urgent action. According to the World Health Organization (WHO), NCDs are the cause of > 41 million deaths every year (accounting for 71% of all deaths worldwide) and thereby have a dramatic impact at the societal and economic levels.

Interestingly, the period comprising gestation and the first years of life is considered to be the most critical period in terms of the risk of developing NCDs. The impact of perinatal nutrition on infants' microbiome development has become the subject of significant interest but there is scarce information concerning women's microbiome and nutrition prenatally and during gestation. Maternal environmental exposures, including diet and microbes, can promote long-lasting or even induce permanent changes in foetal physiology, thereby exerting an impact on the risk of disease in later life. Nutrition also exerts both short- and long-term effects on human health through programming immunological, metabolic and microbiological

development. Furthermore, the interaction between nutrients-microbiota-host towards epigenetic regulation would impact foetal development and also, infant and maternal health outcomes.

Maternal microbiota represents the main microbial source for infants and specific microbial transference from mothers to their offspring occurs at birth and during lactation. Maternal microbial dysbiosis during pregnancy is transferred to the neonate, resulting in the inadequate microbial inoculum, immune and metabolic effects development with future unfavorable health outcomes. Currently, it is unclear whether transgenerational epigenetic inheritance or intrauterine exposures influence the offspring's health and disease susceptibility.

Diet plays a major role in shaping the gut microbiota, while nutrient–microbiota interactions influence the host's health outcomes, thereby having critical implications for health; indeed, "***We are what we eat***". Microbiota is involved in dietary digestion, nutrient absorption, immune system training, pathogen and toxin protection as well as production of specific compounds (SCFAs, vitamins, hormones, neurotransmitters, etc.). Microbiota interact with the metabolism of dietary carbohydrates, proteins, plant polyphenols, bile acids and vitamins. Additional studies are needed to identify which foods, macro- and micro-nutrients and specific dietary compounds influence the microbiota. Specific dietary nutrients, such as fiber, methyl donors (betaine, methionine, and choline), +folate and other group B vitamins (B2, B6 and B12), are linked to microbiota. Humans need these nutrients, and besides dietary sources, there is evidence that the gut microbiota are also a source of essential nutrients for the host (e.g., folate and other B-vitamins). Remarkably, most of these nutrients play a crucial role in epigenetic regulation and can have an impact on the human epigenome.

Physical activity

The application of new high-throughput technologies in the -omics era will not only allow the identification of exercise-induced molecular changes in the epigenome, but will also help to facilitate our understanding of the mechanisms that contribute to improving health. The value of analyzing in detail the changes or epigenetic alterations induced by exercise are unquestionable. It will help us to understand the beneficial effects of exercise on disease prevention and treatment and will offer new potential therapeutic targets. It will provide clues to how to combat cognitive diseases associated with aging such as Alzheimer's, given that the results obtained so far indicate the beneficial effect of exercise on memory and cognition (Hillman et al., 2008; Fernandes et al., 2017). A more in-depth knowledge of these alterations

will also generate information on molecular markers, indicators of optimal sports performance that are of great interest to professional athletes. And last but not least, it will help us clarify how exercise-induced epigenetic changes are sustained over time (epigenetic memory) (Sharples et al., 2016; Seaborne et al., 2018b), and even whether these epigenetic changes are inherited and thus have beneficial effects on the offspring (Segabizani et al., 2019; Spinder et al., 2019).

Alcohol and drugs of abuse

Currently, there is substantial evidence that tobacco smoking leads to DNA methylation changes in human (Lee and Pausova, 2013). Microarray association studies have led to the identification of recurrent epigenetic markers such as the *F2RL3* gene, related to vascular functions, or the xenobiotic metabolism *AhR* gene (Breitling et al., 2011; Shenker et al., 2013; Sun et al., 2013). These and other observations suggest that the epigenetic alterations involved could be important in the molecular mechanisms associated to the tobacco adverse effects. On the other hand, the epigenetic effects of alcohol have been studied in a more mechanistic manner, especially because ethanol is known to disrupt one-carbon metabolism, and thus may influence methyl-group usage by DNA methyltransferases (Pérez et al., 2019; Ron and Messing, 2011). As in the case of tobacco, the most recent studies are starting to apply genome-wide technologies to screen for alcohol-associated genetic *loci* (Zhang and Gelernter, 2017). On the whole, genome-scale association studies have the potential to identify biomarkers of compound usage and potential pathways involved in the etiology of drug-related pathology, while mechanistic studies are particularly relevant to the latter application.

Key challenging points

I. Understanding of the epigenetic clock

What is the biological significance of the existence of CpG sites whose methylation status reflects chronological or biological aging? The *loci* which make up the clocks have been associated with specific genomic features, such as polycomb-associated sites (Raj and Horvath, 2020) but there are still no clear-cut links to any specific biological process. It remains to be clarified whether epigenetic clocks are “drivers” of aging or mere “passengers” reflecting the footprints of other processes, an observation which will be relevant to the design of anti-aging interventions.

II. Measuring and identifying variability

The characterization of epigenetic variability is key to the identification and separation of epigenetic drift from other functional or disease-associated changes. Large-scale studies facilitated by the integration of public datasets will throw light on this scenario. However, different interpretations and measurements of variability coexist in the field (BIOS consortium et al., 2016; Gentilini et al., 2015) and their biological significance will need to be better defined.

III. Delineating chronological and biological aging

Medical advances leading to increases in lifespan will make no sense without corresponding increases in healthspan. Both a more global and a more specific characterization of biological aging and its biomarkers is needed in order to be able to confront this challenge, and more-refined epigenetic clocks will surely be crucial in this scenario (Bell et al., 2019; Partridge et al., 2018).

IV. Investigating epigenetic transgenerational inheritance in mammals

It has been proposed that the epigenome susceptibility to adapt to lifestyle factors, including nutrition, is different through the lifespan of an organism, being more sensitive to changes at early stages (pre- and neonatal period) (Kanherkar et al., 2014). Also, epigenomic modifications seem to be reversible, although it has been postulated that these alterations can passed through generations. In this regard, experimental data in humans have demonstrated that metabolic disorders (undernutrition and maternal obesity) during early development periods (pregnancy) generate an abnormal developmental ambient which could modify the epigenome and predispose the offspring to metabolic diseases later in life (Tobi et al., 2014; de Rooij et al., 2006a,b). This has been explained by the so-called Developmental Origins of Health and Disease hypothesis in which a transgenerational epigenetic inheritance is proposed (Gluckman & Hanson, 2004) This hypothesis is also focussed on the effects of pre- and peri-conceptual nutrition in both parents which would point out to the possibility of providing a favorable nutritional status of development to obtain beneficial epigenetic changes (Fleming et al., 2018). However, at present it remains equivocal whether transgenerational epigenetic inheritance or intrauterine exposures influence offspring's health and disease susceptibility. These challenges deserve future investigations, and will require observation and tracking multiple generations.

V. Dissecting the role of diet-induced epigenetic modifications in human health and disease

Nutrition can also affect health and predispose to disease susceptibility later in life. Evidences from dietary intervention studies, as well as from researches in which the effect of food components have been assayed in humans and animal experimental models, have suggested that dietary components (nutrients and bioactive compounds) exert different biological activities that could render protective effects against different non-communicable diseases and lead to a healthier ageing. However, it is challenging to elucidate their molecular mechanisms and associated epigenetic modifications. Interestingly, this approach to evaluate the potential benefits of a nutrient or food component could be useful to identify epigenetic changes and define early biomarkers of disease. This will also be important to desing new effective nutrition-based preventative and therapeutic approaches, which will contribute to improve health and to reduce the risk of disease. Yet nutrigenomic studies are complex and proving a causation from an association is complicated. It is challenging to identify which components of a disease phenotype are related to nutrition, except for those diseases caused by a single gene defect. In addition, it is also complicated to understand how humans respond to specific diets or nutrients and to detemine the component/s in food responsible/s for an action. Indeed, there are many components in the diet and these substances interact causing multiple metabolic changes, which may even differ depending on the way of intake for the same food (Nicodemus-Johnson & Sinnott, 2017). In addition, it should be taken into account that the genome and the epigenome might interact, i.e. understand how epigenomic modifications could alter gene expression, and regulate the impact of nutrition constitutes also a great challenge that will enable the fisrt steps towards the precision nutrition in disease prevention.

VI. Understanding the impact of exogenous miRNA

miRNA have been identified in biological fluids (blood, human breast milk and milk from other species). However, the influence of these exogenous miRNAs has not been studied in depth despite these epigenetic elements are detected in most foodstuffs and are frequently conserved across species (Xia et al., 2011.; Ledda et al., 2020; Mal et al., 2018) Understanding the potential impact of exogenous miRNA on human epigenetics and their possible influence on health and disease susceptibility is decisive. Indeed, a proved benefitial or detrimental effect might lead to changes in food manufacturing, processing and/or cooking. In addition, because of their generalized presence in foods, miRNA have been suggested as potential biomarkers and/or communication elements (Benmoussa & Provost 2019). All this might enable a new therapy approach against non-communicable diseases and/or promotion of a healthier ageing.

VII. Uncovering the genetic and epigenetic consequences of food chemicals

It is well known that thousands of toxic substances can be present in foodstuffs as a consequence of the production, transport, processing, packaging, and storage practices, but also due to the natural impact of the residual environmental contamination. The number of chemical substances that has been found to have epigenetic toxicity is continuously rising (Marczylo et al., 2016). On the other hand, our knowledge of the epigenetic and genetic effects of old and new chemicals is expected to increase significantly thanks to recent advances on the analytical techniques and methodologies. Besides, different diseases have shown to be transgenerationally transmitted in animal models (Guerrero-Bosagna & Jensen 2015). All this new knowledge pushes research toward studies focused on both the epigenetic changes observed nowadays and the transgenerational consequences of current human exposure to toxic chemicals related to the ingestion of food. To reach this objective it is essential to develop and transfer appropriate determination methodologies based on state-of-the-art instrumentation from the research to the official routine laboratories for the accurate, fast, and green determination of well-known regulated chemicals in food, but also of emerging and new toxics compounds that could be introduced in any of the steps of the food chain. To achieve this latter goal, non-target approaches must be implemented in routine controls, which makes mandatory the adaptation of the analytical methodologies in use to fulfill current demands regarding selectivity and sensitivity. Such approaches will allow achieving the comprehensive experimental data on food exposure that will allow proper correlation with epigenetic and genetic changes observed in the population.

VIII. Moving towards precision and personalized nutrition

The final objective of precision nutrition is being able to provide personalized dietary advice taking into account individual responses to maintain health and prevent disease. To this, advance in the following points is required:

- Large population studies based on well and thoroughly characterized populations (clinical anamnesis, sex, age, lifestyle, (epi)genetics, microbiota) are needed to advance in the identification of determinants of individual variability in the response to specific dietary interventions for the different diet-related traits. Clinical trials and cohort studies.
- Better knowledge of genetic determinants of ADME (for nutrients and bioactive food components) and food preferences, as well as dietary requirements influenced by SNPs. Comprehensive list of SNPs–DNA-nutrient database.

- Tackle the impact of ncRNA, both from the individual and diet-derived ncRNA, in target organs and their impact on different pathologies (e.g. obesity, cancer). Study the impact of dietary interventions on miRNA and their potential as biomarkers/therapeutic tools for precision nutrition in specific pathologies.
- Systems biology/Bioinformatics-Integration of clinical data, genetic background, microbiota, and other multiple -omic data (epigenomic, transcriptomic, proteomic, metabolomic, metagenomics…) in clinical trials and cohort studies requires powerful bioinformatic resources, including machine-learning algorithms able to predict response based on data integration.

IX. Reshaping microbiota through nutrition

Reshaping host–microbiota interactions through personalized nutrition would be a new tool for improving health towards disease control and prevention. It is important to understand the host response based on the microbiota profile in specific dietary intervention (e.g. responders and non-responders). When (circadian feeding patterns and intermittent fasting) and how (cooking processes) the specific dietary intake or food are consumed may exert a different impact on host physiology and microbiota, as well as in the host-microbiome interaction. To date, only limited research has been conducted to assess whether nutritional factors lead to changes in the microbiota and drive host-microbiota interactions mainly in the critical periods of life as early infancy and elderly. Such observations would explain the great interest in perinatal interventions and the potential use of probiotics, prebiotics and symbiotics to promote an "adequate microbiota" and, thus, to beneficially affect health. At the same time, there is significant interest in microbiota-related research aiming to establish the identity of specific microorganisms, microbial molecules and metabolites that contribute to the host's physiology, metabolisms and health. Understanding how the microbiome responds to dietary constituents and the subsequent biological impact, as well as clinical consequences, can be used for the development of precision-tailored dietary interventions.

- To investigate how specific dietary nutrients confers the organism with benefits that go beyond their nutritional input by helping to improve general well-being or reducing the risk of disease.
- Understand the host-microbiome-diet interactions and mechanisms behind
- To identify specific microorganisms, microbial molecules and/or metabolites contributing to the host's physiology, metabolisms and health.
- To understand the host response based on microbiota profile in different specific dietary intervention (responders and non-responders).

- To what extent human health is modulated bywhen (circadian feeding patterns and intermittent fasting) and how (cooking processes) dietary nutrients are consumed
- Develop mechanistic predictive models of the effect of dietary components and microbiota-based products on health outcomes

X. Towards a more global and mechanistic understanding of epigenetic consequences of physical exercise

There are several important questions that should be addressed in coming years regarding how epigenetic alterations due to physical activity can impact human health and disease:

- Design animal models to study the molecular mechanisms behind the beneficial or harmful effects of exercise that can be transferred to humans (e.g. in elite athletes).
- Implementation of new generation epigenetic technologies to study epigenetic marks at the genome-wide level.
- Integration of epigenomic data with data obtained from other -omics (i.e. transcriptomics and proteomics) to identify epigenetic changes with functional effects.
- The study of the effect of exercise on the epigenome of different cell types and tissues to identify common and specific changes.

XI. Defining the etiological role of epigenetic alterations induced by tobacco and alcohol in human disease.

Most of the current findings of epigenetic alterations are the result of association studies. However, this approach has limited power in clarifying whether these changes are causes or consequences of alcohol or tobacco-associated pathology. The signaling routes and molecular mechanisms involved remain to be defined by the development of mechanistic studies. Moreover, because of its accessibility, the majority of epigenetic changes have been described in peripheral blood, which is not the primary target tissue of alcohol nor tobacco. It is probable that, aside from biomarkers, more functional associations of epigenetic alterations and genetic expression will be detected by examining other tissues more directly related to these compounds.

CSIC advantage position and multi/inter-disciplinarity

Aging and recreational drugs

Within CSIC, the *Interdisciplinary Thematic Platforms (PTIs) program has established transversal research initiatives that connect research groups across its different institutes, in*

parallel with the European Commission's missions. In this setting, the HEALTH-AGING PTI provides the perfect framework for the development of synergies between research groups in the quest to tackle the modern challenges regarding healthy aging and the prevention of age-associated diseases. This will only be made possible by the conjunction of multi-area experienced groups at institutes such as the Cajal Institute (IC-CSIC) and the Institute of Biomedical Research of Barcelona (IIBB-CSIC). On the other hand, the collaborations needed to fully characterize and understand drug-associated epigenetic alterations (e.g. tobacco and ethanol) can also be framed within the HEALTH-AGING PTI.

Nutrition and diet

The Spanish National Research Council is an international reference institution in which multidisciplinary investigations are hold. Understanding the regulatory effect of nutrition on critical metabolic pathways to influence individual health and prevent diseases by elucidating its molecular pathways involved in genomic and epigenomic modifications is an emerging research area. This challenge will require a multidisciplinary approach (nutritionists, toxicologists, endocrinologists, developmental biologists, clinicians, epidemiologists, etc.) and it will provide a great benefit for the public health. Indeed, a deeper understanding of nutrient-induced genetic and epigenetic changes may provide a great opportunity to explore therapies based on these mechanisms, which constitutes a chance to prevent non-communicable diseases and promote a healthier ageing. To achieve these goals, it is essential to have access to -omics technologies (genomic, epigenomic, transcriptomic, proteomic, metabolomic, metagenomics, bioinformatics, etc.), which currently are available only for certain groups, while many others work outsourcing these analyses to private companies/technologic parks or relaying in external collaborations. Furthermore, collaboration between research groups in these disciplines and also with clinical researchers at hospitals is also essential to advance in resolving some of the identified key challenging points. However, so far there is little collaboration of this kind. In any case, the impact of these nutritional investigations is clear. In this line, several consolidated research groups in CSIC develop lines of research addressing the main diseases and health problems of the population from a multidisciplinary perspective. Due to the social impact of these research studies undoubtedly will contribute to the maintenance and enhancement of internationally good reputation and recognision of CSIC.

Pollutants

The CSIC is and will continue to be a reference institution at national and international levels in the determination of pollutants in foodsttufs and their impact in human epigenetics. Studies in these emerging research areas will certainly continue, mainly in collaboration with highly renowned national and international groups, and new challenges will be successfully addressed and effectively solved in the next decades so contributing to maintain and enhance the international visibility and heft of CSIC.

Physical exercise

To successfully understand the impact that physical exercise and recreational drugs have on our epigenome and, thus, on human health and disease, it is essential to foster the collaboration of different multidisciplinary research groups. Within CSIC there are research groups with great experience in the design of animal models to study the effect of exercise at the physiological level (Cajal Institute) as well as the epigenetic level (CINN-CSIC). Synergies with Institutes hosting strong computational analysis units such us CNB-CSIC and CIC-IBMCC, USAL/CSIC will be also essential to analyze and integrate the large amount of data generated.

Plan and resources

While there is capacity within the different institutes for the development of research regarding the interplay between our lifestyle and our (epi)genome, there is still a general lack of groups specialized in this area. Because of this, two parallel strategies should be developed to: reinforce the existent groups with personnel, equipment and computational resources (the latter is particularly necessary in the -omics field) and stimulate the creation of new research teams. More specifically, we highlight the following needs:

(i) experienced researchers with new complementary (or lacking) knowledge to many of the considered areas of expertise.

(ii) highly specialized technicians who facilitate the maintenance of all this novel knowledge and technological development and the efficient transfer to new-comers to the research groups.

(iii) due to the multidisciplinarity of the studies involved in these investigations and the requirement from -omic technologies to understand the impact of the genomic and epigenomic changes, incorporation of professionals from different research areas will also be recommendable (bioinformatics, biochemists, nutricionists, etc.).

(iv) the access to –omic technologies could be improved by creating platforms that ensure the accesibilty to the required infrastructure and that posses dedicated and qualified technicians that can offer (centralized) internal services. – 1-2 years.

(v) the generation of –omic data will also require the endorsement of a bioinformatics unit with experts in data mining, mathematicians, etc, which will necessarily demand the incorporation of qualified personnel.

(vi) the acquisition and maintenance of novel analytical instruments that allow retaining current levels of competence of the CSIC research groups in a highly competitive scenario.

(vii) Implementation/improvement of facilities for large-scale human intervention trials and cohort studies, with specific support from dietitians, nurses, etc.

On the other hand, research in this field requires the establishment of internal communication channels between groups with the potential to carry out this type of projects, and to promote multidisciplinary collaborations. Moreover, CSIC should also promote the synergies between institutes in the biomedical field with those in the social sciences, such as the Institute of Economics, Geography and Demography (IEGD-CSIC), which have an established capacity to gather demographic material of use in this field of research.

References

Bell, C.G., Lowe, R., Adams, P.D., Baccarelli, A.A., Beck, S., Bell, J.T., Christensen, B.C., Gladyshev, V.N., Heijmans, B.T., Horvath, S., et al. (2019). DNA methylation aging clocks: challenges and recommendations. Genome Biol 20, 249.

Benmoussa, A. & PROVOST, P. (2019). Milk microRNAs in health and disease". Compr. Rev. Food Sci. Food Safety 18, 703-722.

Biesiekierski JR, Jalanka J, Staudacher HM. Can Gut Microbiota Composition Predict Response to Dietary Treatments?. Nutrients. 2019;11(5):1134. Published 2019 May 22. doi:10.3390/nu11051134.

BIOS consortium, Slieker, R.C., van Iterson, M., Luijk, R., Beekman, M., Zhernakova, D.V., Moed, M.H., Mei, H., van Galen, M., Deelen, P., et al. (2016). Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol 17, 191.

Booth, F.W., Roberts, C.K., and Laye, M.J. (2012). Lack of exercise is a major cause of chronic diseases. Compr Physiol 2, 1143-1211.

Breitling, L.P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. (2011). Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. The American Journal of Human Genetics 88, 450–457.

De Rooij, S. R., et al. (2006a). Impaired insulin resistance secretion after prenatal exposure to the Dutch famine. Diabetes Care 29, 1897-1901.

De Rooij, S. R., et al. (2006b). Glucose tolerance at age 58 and the decline of glucose tolerance in comparison with age 50 in people prenatally exposed to the Dutch famine. Diabetologia 49, 637-643.

Elsner, V.R., Lovatel, G.A., Bertoldi, K., Vanzella, C., Santos, F.M., Spindler, C., de Almeida, E.F., Nardin, P., and Siqueira, I.R. (2011). Effect of different exercise protocols on histone acetyltransferases and histone deacetylases activities in rat hippocampus. Neuroscience 192, 580-587.

Feil, R., and Fraga, M.F. (2012). Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet 13, 97-109.

Fernandes, J., Arida, R.M., and Gomez-Pinilla, F. (2017). Physical exercise as an epigenetic modulator of brain plasticity and cognition. Neurosci Biobehav Rev 80, 443-456.

Fernandez-Sanles, A., Sayols-Baixeras, S., Castro, D.E.M.M., Esteller, M., Subirana, I., Torres-Cuevas, S., Perez-Fernandez, S., Aslibekyan, S., Marrugat, J., and Elosua, R. (2020). Physical Activity and Genome-wide DNA Methylation: The REgistre GIroni del COR Study. Med Sci Sports Exerc 52, 589-597.

Fiuza-Luces, C., Garatachea, N., Berger, N.A., and Lucia, A. (2013). Exercise is the real polypill. Physiology (Bethesda) 28, 330-358.

Fleming, T.P., et al. (2018). Origins of lifetime health around the time of conception: causes and consequences. Lancet 391, 1842-1852.

Gentilini, D., Garagnani, P., Pisoni, S., Bacalini, M.G., Calzari, L., Mari, D., Vitale, G., Franceschi, C., and Di Blasio, A.M. (2015). Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. Aging 7, 568–578.

Gluckman, P.D., and Hanson, M.A. (2004). Developmental origins of disease paradigm: a mechanistic and evolutionary perspective. Ped. Res. 56, 311-317.

Guerrero-Bosagna, C. and Jensen, P. (2015). Globalization, climate change, and transgenerational epigenetic inheritance: will our descendants be at risk? Clinical Epigenetics 7, 8. DOI 10.1186/s13148-014-0043-3.

Hillman, C.H., Erickson, K.I., and Kramer, A.F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. Nat Rev Neurosci 9, 58-65.

Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet 19, 371–384.

Huidobro, C., Fernandez, A.F., and Fraga, M.F. (2013). Aging epigenetics: Causes and consequences. Molecular Aspects of Medicine 34, 765–781.

Instituto Nacional de Estadística (2017). Encuesta Nacional de Salud 2017.

Jones, M.J., Goodman, S.J., and Kobor, M.S. (2015). DNA methylation and healthy human aging. Aging Cell 14, 924–932.

Kanherkar, R.R., Bhatia-Dey, N., Csoka, A.B. (2014). Epigenetics across the human lifespan. Front. Cell Develop. Biol. 2, 1-19.

Kolodziejczyk AA, Zheng D, Elinav E. Diet-microbiota interactions and personalized nutrition. Nat Rev Microbiol. 2019;17(12):742-753. doi: 10.1038/s41579-019-0256-8.

Ledda, B., et al. (2020). Small RNAs in eucaryotes: new clues for amplifying microRNA benefits. Cell Biosci. 10, DOI: 10.1186/s13578-019-0370-3.

Lee, K.W.K., and Pausova, Z. (2013). Cigarette smoking and DNA methylation. Front. Genet. 4.

Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) 10, 573–591.

Mal, C., Aftabuddn, M., Kundu, S. (2018). IIKmTA: Inter and intra kingdom miRNA-target analyzer. Interdiscip. Sci. Comput. Life Sci. 10, 538-543.

Marczylo, E.L., Jacobs, M.N., Gant, T.W. (2016). Environmentally induced epigenetic toxicity: potential public health concerns. Crit. Rev. Toxicol. 46, 676-700.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The Hallmarks of Aging. Cell 153, 1194–1217.

Lu, A.T., Quach, A., Wilson, J.G., Reiner, A.P., Aviv, A., Raj, K., Hou, L., Baccarelli, A.A., Li, Y., Stewart, J.D., et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging 11, 303–327.

McCartney, D.L., Zhang, F., Hillary, R.F., Zhang, Q., Stevenson, A.J., Walker, R.M., Bermingham, M.L., Boutin, T., Morris, S.W., Campbell, A., et al. (2020). An epigenome-wide association study of sex-specific chronological ageing. Genome Med 12, 1.

Melo, S.F., Barauna, V.G., Junior, M.A., Bozi, L.H., Drummond, L.R., Natali, A.J., and de Oliveira, E.M. (2015). Resistance training regulates cardiac function through modulation of miRNA-214. Int J Mol Sci 16, 6855-6867.

Mills, S.; Lane, J.A.; Smith, G.J.; Grimaldi, K.A.; Ross, R.P.; Stanton, C. Precision Nutrition and the Microbiome Part II: Potential Opportunities and Pathways to Commercialisation. Nutrients 2019, 11, 1468.

Neufer, P.D., Bamman, M.M., Muoio, D.M., Bouchard, C., Cooper, D.M., Goodpaster, B.H., Booth, F.W., Kohrt, W.M., Gerszten, R.E., Mattson, M.P., et al. (2015). Understanding the Cellular and Molecular Mechanisms of Physical Activity-Induced Health Benefits. Cell Metab 22, 4-11.

Nielsen, S., Scheele, C., Yfanti, C., Akerstrom, T., Nielsen, A.R., Pedersen, B.K., and Laye, M.J. (2010). Muscle specific microRNAs are regulated by endurance exercise in human skeletal muscle. J Physiol 588, 4029-4037.

Nieman, D.C., and Wentz, L.M. (2019). The compelling link between physical activity and the body's defense system. J Sport Health Sci 8, 201-217.

Nicodemus-Johson, J., and Sinnott, R.A. (2017). Fruit and juice epigenetic signatures are associated with independent immunoregulatory pathways. Nutrients 9, E752. DOI: 10.3390/nu9070752.Pal, S., and Tyler, J.K. (2016). Epigenetics and aging. Sci. Adv. 2, e1600584.

Pandorf, C.E., Haddad, F., Wright, C., Bodell, P.W., and Baldwin, K.M. (2009). Differential epigenetic modifications of histones at the myosin heavy chain genes in fast and slow skeletal muscle fibers and in response to muscle unloading. Am J Physiol Cell Physiol 297, C6-16.

Partridge, L., Deelen, J., and Slagboom, P.E. (2018). Facing up to the global challenges of ageing. Nature 561, 45–56.

Pérez, R.F., Santamarina, P., Fernández, A.F., and Fraga, M.F. (2019). Epigenetics and Lifestyle: The Impact of Stress, Diet, and Social Habits on Tissue Homeostasis. In Epigenetics and Regeneration, (Elsevier), pp. 461–489.

Raj, K., and Horvath, S. (2020). Current perspectives on the cellular and molecular features of epigenetic ageing. Exp Biol Med (Maywood) 153537022091832.

Rinninella E, Cintoni M, Raoul P, et al. Food Components and Dietary Habits: Keys for a Healthy Gut Microbiota Composition. Nutrients. 2019;11(10):2393. Published 2019 Oct 7. doi:10.3390/nu11102393.

Ron, D., and Messing, R.O. (2011). Signaling Pathways Mediating Alcohol Effects. In Behavioral Neurobiology of Alcohol Addiction, W.H. Sommer, and R. Spanagel, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 87–126.

Ronn, T., Volkov, P., Davegardh, C., Dayeh, T., Hall, E., Olsson, A.H., Nilsson, E., Tornberg, A., Dekker Nitert, M., Eriksson, K.F., et al. (2013). A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLoS Genet 9, e1003572.

Schenk, A., Pulverer, W., Koliamitra, C., Bauer, C.J., Ilic, S., Heer, R., Schier, R., Schick, V., Bottiger, B.W., Gerhauser, C., et al. (2019). Acute Exercise Increases the Expression of KIR2DS4 by Promoter Demethylation in NK Cells. Int J Sports Med 40, 62-70.

Seaborne, R.A., Strauss, J., Cocks, M., Shepherd, S., O'Brien, T.D., Someren, K.A.V., Bell, P.G., Murgatroyd, C., Morton, J.P., Stewart, C.E., et al. (2018a). Methylome of human skeletal muscle after acute & chronic resistance exercise training, detraining & retraining. Sci Data 5, 180213.

Seaborne, R.A., Strauss, J., Cocks, M., Shepherd, S., O'Brien, T.D., van Someren, K.A., Bell, P.G., Murgatroyd, C., Morton, J.P., Stewart, C.E., et al. (2018b). Human Skeletal Muscle Possesses an Epigenetic Memory of Hypertrophy. Sci Rep 8, 1898.

Segabinazi, E., Spindler, C., Meireles, A.L.F., Piazza, F.V., Mega, F., Salvalaggio, G.D.S., Achaval, M., and Marcuzzo, S. (2019). Effects of Maternal Physical Exercise on Global DNA Methylation and Hippocampal Plasticity of Rat Male Offspring. Neuroscience 418, 218-230.

Sharples, A.P., Stewart, C.E., and Seaborne, R.A. (2016). Does skeletal muscle have an 'epi'-memory? The role of epigenetics in nutritional programming, metabolic disease, aging and exercise. Aging Cell 15, 603-616.

Shenker, N.S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M.A., Belvisi, M.G., Brown, R., Vineis, P., and Flanagan, J.M. (2013). Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum. Mol. Genet. 22, 843–851.

Spindler, C., Segabinazi, E., Meireles, A.L.F., Piazza, F.V., Mega, F., Dos Santos Salvalaggio, G., Achaval, M., Elsner, V.R., and Marcuzzo, S. (2019). Paternal physical exercise modulates global DNA methylation status in the hippocampus of male rat offspring. Neural Regen Res 14, 491-500.

Sun, Y.V., Smith, A.K., Conneely, K.N., Chang, Q., Li, W., Lazarus, A., Smith, J.A., Almli, L.M., Binder, E.B., Klengel, T., et al. (2013). Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum. Genet. 132, 1027–1037.

Swann JR, Rajilic-Stojanovic M, Salonen A, Sakwinska O, Gill C, Meynier A, Fança-Berthon P, Schelkle B, Segata N, Shortt C, Tuohy K, Hasselwander O . Considerations for the design and conduct of human gut microbiota intervention studies relating to foods. Eur J Nutr. 2020 Apr 3. doi: 10.1007/s00394-020-02232-1

Talhout, R., Schulz, T., Florek, E., Van Benthem, J., Wester, P., and Opperhuizen, A. (2011). Hazardous Compounds in Tobacco Smoke. IJERPH 8, 613–628.

Tejedor, J.R., and Fraga, M.F. (2017). Interindividual epigenetic variability: Sound or noise? BioEssays 39, 1700055.

Tobi, E. W., et al. (2014). DNA methylation signatures link prenatal famine exposure to growth and metabolism. Nature Commun. 5, 5592. DOI: 10.1038/ncomms6592.

Xia, J.H., He, X.P., Bai, Z.Y., and Yue, G.H. (2011). Identification and characterization of 63 microRNAs in the Asian Seabass Lates calcarifer. PLoS One 6, e17537. DOI:10.1371/journal.pone.0017537.

Zhang, H., and Gelernter, J. (2017). Review: DNA methylation and alcohol use disorders: Progress and challenges: DNA Methylation in Alcohol Addiction. Am J Addict 26, 502–515.