CHALLENGE 3

ABSTRACT

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

KEYWORDS

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

CHALLENGE 3

THE TREE OF LIFE: INTERTWINING GENOMICS AND EVOLUTION

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

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

EXECUTIVE SUMMARY

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

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

1. INTRODUCTION AND GENERAL DESCRIPTION

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

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

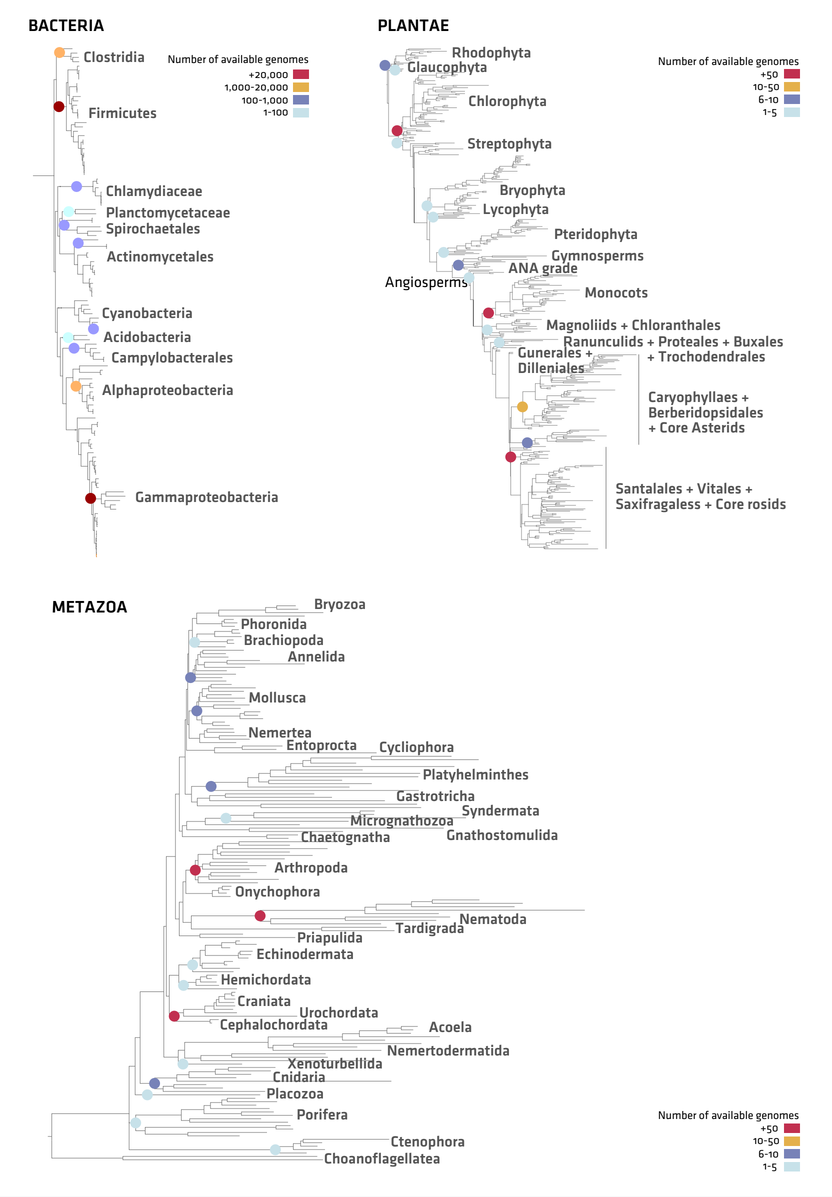
After the completion of the human genome, and in less than 20 years, evolutionary genomics has experienced an unprecedented momentum thanks to the continuous improvement of high-throughput sequencing technologies, which have steadily increased the yield of reads, decreased costs considerably, and improved the overall quality of results (Goodwin et al., 2016). At present, sequencing based on short reads is used widely to obtain transcriptomes, re-sequence genomes or generate thousands of phylogenetic markers through target enrichment. Moreover, it is now possible and increasingly affordable to obtain reference genomes with chromosome-level scaffolds through the combination of long reads (Amarashinge et al., 2020) and technologies that capture chromatin information.

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

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

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

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



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

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

As the sequencing of the human genome produced a paradigm shift in medical research, the sequencing of reference genomes representing life diversity on Earth should bring about a revolution in evolutionary biology in the coming years (Richards, 2015). It will be possible to address many of the current challenges in the field including reconstruction of robust phylogenetic relationships, improved determination of orthologous and paralogous relationships, characterization of the tempo and mode of gene family evolution, understanding of genome dynamics, identification of the genomic targets of natural selection, exhaustive detection of evolutionary innovations, recognition of causal connections between genotype and phenotype, recognition of the genomic regions and functions responding to environmental change, etc. Altogether, the availability of reference genome should set the basis to globally enhance biological research (and conservation) of previously neglected groups, as many latest technologies are only applicable if genomic data are available (Richards, 2015). A genomic approach to the ToL has also important applied deliverables and for instance, it should accelerate the discovery in related species within the ToL of a variety of highly efficient and specific metabolites and natural drugs, which are the products of millions of years of evolution and selection (see also Challenge 2 and Challenge 7).

2. IMPACT IN BASIC SCIENCE PANORAMA AND POTENTIAL APPLICATIONS

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

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

One reason to unveil unknown microbial diversity is to broaden the general knowledge that has been gained from model organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*. As more microbial species are identified, it is becoming clear that the existing diversity of molecular, cellular, and functional biology in nature goes far beyond what has been learnt from model organisms (e.g., Rivas-Marín et al., 2016). Another important reason to enhance microbial species discovery is to help discern the relative contribution of Archaea and Bacteria to the endosymbiotic origin of Eukarya (see also Challenge 1). For instance, a new phylum of Archaea, the Lokiarchaea, was recently identified through metagenomic analysis of deep marine sediments. The genomes of these cells contained the highest number of previously considered eukaryotic-specific features, although the cells lacked the eukaryotic-like cellular organization (Imachi et al., 2020). Moreover, various phylogenies consistently placed eukaryotes within the Lokiarchaea phylum, although the debate remains open. Finally, an applied reason to promote wide microbial prospecting is that they are likely the source of many new metabolic routes of importance in global cycles and many interesting products of biotechnological utility that could be detected as more efficient bioinformatics tools are developed.

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

One relevant source of unresolved nodes in the ToL is constituted by highly diversified taxa originated through rapid radiations, which are commonplace at different taxonomic levels. In addition to the difficulty of inferring phylogenetic relationships due to the intrinsic short internal branches, the challenge has been to determine whether shared polymorphism between radiated taxa is due to recent divergence and incomplete lineage sorting, or partly the result of hybridization and gene flow during speciation. Whole-genome assemblies will not only provide increased power to definitively resolve phylogenetic relationships in rapid radiations, but also to address the role of hybridization in promoting, not preventing, speciation (Stryjewski and Sorenson, 2017). In this sense, comparative genomics is and will be expanded enormously to the population level (i.e., population genomics) aiming at understanding the relative role of natural selection, genetic drift, migration, hybridization, incomplete lineage sorting, and demography on the diversification of species. The continuum of speciation can be comprehensively studied at an unprecedented resolution from population fragmentation and ecological divergence to lineage split and species formation. This is of paramount importance to understand how organisms interact with the biotic and abiotic components of landscape heterogeneity, which has major implications to forecast their future responses to global change.

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

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

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

3. KEY CHALLENGING POINTS

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

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

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

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

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

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

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

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

Horizontal gene transfer. One of the biggest challenges of the current decade is the evaluation of the extent of horizontal gene transfer (HGT) occurring among Archaea and Bacteria in their natural environments, and how that would affect our view of the real diversity of these taxa. It has been proposed that prokaryote taxa evolution cannot be fully described without HGT (Palmer et al., 2019), and that genetic exchanges are so rampant that would blur the ToL at least for these taxa. The latter may be too extreme, as the most recent phylogeny based on almost 100k genomes fairly mirrored the reconstructed trees based on the 16S rRNA gene, which is assumed to be inherited only vertically (Parks et al., 2018). At present, there are different platforms such as the MiGA database (Rodriguez-R et al., 2019), which acquire and organise high quality genomes from new microbial isolates, as well as MAGs from environmental samples and species microbiomes. Therefore, this will provide access to statistically sound datasets to both reconstruct the Archaea and Bacteria ToL and trace HGT. In eukaryotes, detection and validation of HGT have demonstrated to be far more complex, and often require high coverage of the genomes as well as strict controls for bacterial contamination. That said, there are plenty of examples of HGT to eukaryotes (Husnik and McCutcheon, 2018). In eukaryotes, when a complete gene is transferred from Archaea or Bacteria, the gene function can be retained, widening the functional complements of the organism, including nutritional improvements, toxin delivery, adaptation to extreme environments, and protection from archaeal or bacterial pathogens. There is no doubt that as genome quality improves and more neglected taxa are sequenced, the extent and diversity of HGT in eukaryotes will be finally unveiled.

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

The accurate discrimination of factors that lead to tree discordance is one the major challenges in phylogenomics. Because ILS and introgression between closely related taxa may produce similar discordant phylogenetic trees, distinguishing between both is complex, and requires developing appropriate statistical approaches that use whole-genome data in reduced taxon datasets (Durand et al., 2011). These tests are in their infancy and their improvement (e.g., by inferring the direction of gene flow in large taxon datasets) represents an open and active field.

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

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

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

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

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

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

CHALLENGE 3 REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

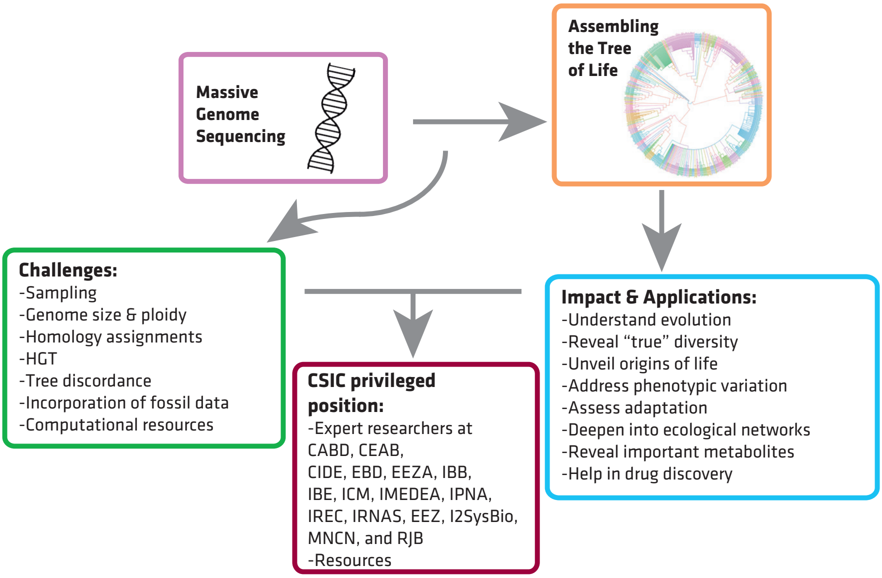
Srivastava, D.S., Cadotte, M.W., MacDonald, A.A.M., Marushia, R.G. and Mirotchnick, N. (2012). Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters 15,* 637–648.

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

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

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

SUMMARY FOR EXPERTS



Massive Genome Sequencing

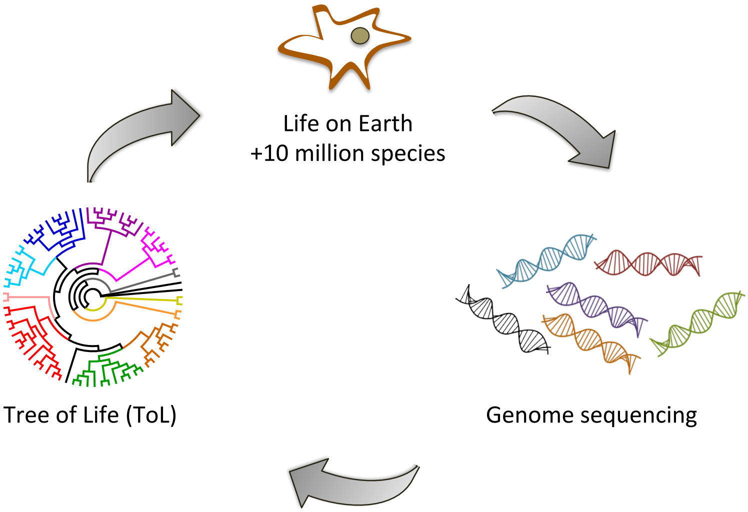
Assembling the Tree of Life

Challenges:
-Sampling
-Genome size & ploidy
-Homology assignments
-HGT
-Tree discordance
-Incorporation of fossil data
-Computational resources

CSIC privileged position:
-Expert researchers at CABD, CEAB, CIDE, EBD, EEZA, IBB, IBE, ICM, IMEDEA, IPNA, IREC, IRNAS, EEZ, I2SysBio, MNCN, and RJB
-Resources

Impact & Applications:
-Understand evolution
-Reveal "true" diversity
-Unveil origins of life
-Address phenotypic variation
-Assess adaptation
-Deepen into ecological networks
-Reveal important metabolites
-Help in drug discovery

SUMMARY FOR THE GENERAL PUBLIC


Life on Earth
+10 million species
Tree of Life (ToL)
Genome sequencing