Computational and Structural Biotechnology Journal 18 (2020) 482–500

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

The evolution of metabolism: How to test evolutionary hypotheses at the genomic level

Federico Scossa[a,b], Alisdair R. Fernie[a,c,*]

[a] Max-Planck-Institut für Molekulare Pflanzenphysiologie, 14476 Potsdam-Golm, Germany
[b] Council for Agricultural Research and Economics (CREA), Research Centre for Genomics and Bioinformatics (CREA-GB), Via Ardeatina 546, 00178 Rome, Italy
[c] Center of Plant Systems Biology and Biotechnology (CPSBB), Plovdiv, Bulgaria

ARTICLE INFO

Article history:
Received 1 September 2019
Received in revised form 12 February 2020
Accepted 13 February 2020
Available online 20 February 2020

Keywords:
Evolution
Metabolism
Plants
Genomics
Selection
Adaptation

ABSTRACT

The origin of primordial metabolism and its expansion to form the metabolic networks extant today represent excellent systems to study the impact of natural selection and the potential adaptive role of novel compounds. Here we present the current hypotheses made on the origin of life and ancestral metabolism and present the theories and mechanisms by which the large chemical diversity of plants might have emerged along evolution. In particular, we provide a survey of statistical methods that can be used to detect signatures of selection at the gene and population level, and discuss potential and limits of these methods for investigating patterns of molecular adaptation in plant metabolism.

Contents

Abbreviations: E&R, Evolve and Resequence; GWAS, genome wide association study; QTL, quantitative trait locus.

* Corresponding author at: Max-Planck-Institute for Molecular Plant Physiology (MPI-MP), Am Muehlenberg 1, 14476 Potsdam, Germany.
E-mail address: fernie@mpimp-golm.mpg.de (A.R. Fernie).

https://doi.org/10.1016/j.csbj.2020.02.009
2001-0370/

1. Introduction

Plants produce a vast array of metabolites during their lifecycle. Most of these compounds are called "secondary" or "specialized" metabolites, to distinguish them from primary metabolites, which are involved in the basic processes of growth and development. Secondary metabolites instead exert their functions during the interactions of plants with the surrounding environment: they are typically synthesized as defense compounds to deter predators (being toxic to plant pathogens), some may instead confer tolerance against abiotic stresses, and other serve as attractants for pollinators or as signals for plant-plant interactions [181]. Although recent research in plant metabolism has made this distinction less clear-cut than it was before [160], secondary metabolites – as a group – keep some distinctive characteristics: they show an amazing diversity of chemical structures and vastly exceed the number of primary metabolites. There are more than 40,000 reported structures for terpenoids [181], for example, and polyphenols are in the range of 8000–9000 known compounds [6]. Secondary metabolites also show marked qualitative and quantitative variation, both between tissues and developmental stages of a single plant, but also within different individuals of a species, and also across different species [273,181]. This chemical diversity originates from the activity of large and numerous families of enzyme-encoding genes which generally operate in highly branched (and often compartmentalized) metabolic pathways [147].

In the first part of this review (Sections 2 through 5) we try to answer the question about how these metabolic pathways emerged and were shaped during evolution, starting both from what is known about the "RNA world" and the non-enzymatic and **primordial metabolism**, to later look at the specific events driving the large diversification of plant secondary metabolism. In doing so, we will present the retrograde, Granick, patchwork and shell hypotheses for the evolution of metabolic pathways as well as reviewing current thinking as to their likelihoods. In an attempt to answer the focal question about the adaptive value of metabolic diversity, in the second part of this review (section 6), we will try to cover some of the strategies which may be used to understand whether genes carry signatures of selection, presenting examples about how these genomic approaches may illuminate the evolution of plant metabolism.

2. The RNA world

It is received wisdom that ancestral life forms inhabited an environment – commonly known as the primordial soup – rich in the spontaneously formed organic compounds of the prebiotic world [77]. This hypothesis – the **Oparin-Haldane theory** – predicts that some simple organic molecules could form in the highly reducing atmosphere of primitive Earth, simply with the supply of external sources of energy (e.g. UV, lightining). This theory remained speculative until 1953, when Stanley L. Miller, a graduate student at the University of Chicago, showed that glycine and alanine could form in an artificial system which mimicked the probable conditions of primitive Earth [176]. The primordial soup, in any case, was probably not the only environment in which simple organic molecules could be formed. The fall of meteorites and comets was common on primitive Earth: recent analysis of the content of the Murchison and other meteorites revealed an impressive diversity of organic chemicals, including the presence of ribose and other simple sugars [54,172,227,78]. If the formation of organic molecules in the primordial soup, or their "delivery" from falling meteorites is now widely accepted, these simple molecules did not yet represent life: "Life" is, as stated elegantly by Andreas Wagner [267], the *combination of metabolism* (a metabolic network connecting, in terms of reactions, simple building blocks into something more complex, along with the opposite process of breaking larger objects into simpler ones) and *replication* (the process of making more of itself). And this led the scientific community to face the first (of many) chicken and egg problems in the origin of life research: which came first, replication or metabolism (supporting "replication/genetics first": [10] and references therein; original formulation of "metabolism-first" came from [291], with later support from [263–265]).

With the discovery of the structure of DNA in 1953 [269], replication seemed to be the perfect process to occur first. DNA was the bearer of genetic information, and its double helix structure provided a perfect model for its replication (however, DNA can not replicate on its own). Later, when RNA was discovered to also possess catalytic properties [8,43], and some ability to perform primer extension on the basis of a RNA template [121], then the "RNA world hypothesis" acquired growing credit [83,243,125], and the idea of a "RNA replicator" surpassed metabolism as the most probable process to occur at the inception of life, with an RNA ribosome even being postulated [196]. Thus the "RNA world" [83] apparently preceded the DNA/protein/metabolite world.

More recently, however, this theory has been considerably challenged – not least by Markus Ralser who argues that most reactions in the model cell are protein-enzyme catalyzed with the remainder being driven by sunlight, free radical chemistry or metal ions [132]. Whilst the RNA world hypothesis postulates ribozyme-catalyzed reactions, there is a dearth of ribozymes identified that would be important for the core metabolism of any species. Moreover, many of the *in vitro*-selected ribozymes obtain their catalytic activity indirectly *via* their binding of metal ions such as zinc [211]. When taken together, these observations suggest that, if ribozyme-catalyzed metabolic reactions exist at all, their contribution to cellular metabolism as a whole is only marginal. From the perspective of evolutionary theory, it is additionally difficult to envision a scenario by which an RNA-catalyzed reaction system could evolve, as it could not have come into place one step at a time, but only as an operational entity [211]. Thus, the origin of metabolism as an RNA-based metabolic reaction system seems rather improbable. By contrast, RNA plays a dominant role in all steps of translation indicating that protein biosynthesis followed RNA in evolution and that this role was maintained [183].

3. Non-enzymatic (pre-biotic) metabolism

Credit for the "metabolism-first" hypothesis, as opposed to "replication-first", on the other hand, lagged behind with respect to the acceptance of the "RNA world". This was due to the initial lack of two main essential requirements for the metabolism-first theory to be supported: how could the presence of catalysts be explained and the need of small volumes of liquid to facilitate the interaction of molecules and occurrence of reactions [265,10]. These two conditions were met with the discovery of hydrothermal vents in the deep ocean, which represent fissures in the Earth's crust and harbor chemistries reminiscent of primitive Earth. Hydrothermal vents emit hot water and reactive gases at high temperatures, containing high concentrations of transition metals (Fe (II) and Mn(II)) as well as CO_2, H_2S, CH_4 and H_2. The mixture of hot fluid and gases, at alkaline pH, once they diffuse in the cold ocean water, form chimney-like porous deposits of carbonates. Through these pores, the components of the hydrothermal effluents can react together in a microenvironment characterised by large temperature, pressure and pH gradients. These microenvironments host today a rich microbial community, which represent the deepest branch in the tree of life and were probably the sites of the first metabolic reactions to occur on the planet [15,168]. The first reactions to take place could have included various prebiotic precursors of metabolism, including formamide, α-hydroxy and α-amino acids, fatty acids and pyruvate [264,114,50,221,84]. The formation of pyruvate and fatty acids, in particular, has been also confirmed recently under realistically simulated hydrothermal conditions [31,190].

Another recent series of experiments additionally demonstrated that the canonical pathways of glycolysis, the pentose phosphate (PP) pathway and the tricarboxylic acid (TCA) cycle possess highly similar non-enzymatic analogs [130,132]. Pyruvate and glucose, the end-products of glycolysis and gluconeogenesis, respectively, can be formed spontaneously in water, whilst metal ions such as ferrous iron and phosphate and sulfate radicals – which are abundant in hydrothermal vents – catalyze reactions resembling those of the PP and TCA cycles, respectively. These experiments revealed considerable specificity out of the vast chemical space of possible reaction products, providing strong indications that metabolic reactions similar to those used in modern cells did not necessarily originate from the evolutionary selection of complex catalysts [211]. Intriguingly, most modern enzymes of central carbon metabolism are independent on metallic co-factors. However, examples such as that of ribulose 5-phosphate epimerase – for which *E. coli* uses a ferrous ion for its catalytic function [237], while those of higher organisms do not [143], suggest that this may be the result of selection. Interestingly, in parallel to this loss is the evolution of complex iron transport systems to circumvent the problem of Fenton reactions causing superoxide, which results in oxidative stress and even fatal cellular toxicity [211]. In addition to the issue of toxicity, a number of other issues likely underlie the shift from non-enzyme based catalysis to enzyme-dominated catalysts. These include: (i) limited catalyst availability, (ii) improved substrate specificity of enzyme catalyzed reactions, (iii) the prevention of side reactions and (iv) the greater possibilities of metabolic regulation afforded by enzymatic catalysis [131]. It is important to bear in mind that whilst enzymatic catalysis dominates metabolism, non-enzymatic reactions still occur frequently within the metabolic network of all cell types. Indeed, the possibility of non-enzymatic catalysis is often one of the driving forces for keeping certain metabolites at low levels or compartmentalized within enzyme complexes. This fact notwithstanding, the evidence for an important role of non-enzymatic reactions in the evolution of metabolism is highly persuasive; this also seems to be consistent with the suggestion that metabolism could have emerged by exploitation of the rich chemistries present in hydrothermal vents. With the emergence of these prebiotic metabolisms, it is generally considered that the further step in the evolution of life, i.e., the transition to living matter, might have occurred with the encapsulation of these organic molecules within micelles or vesicles [34]. Although several mathematical models have been proposed to account for the development of evolvable systems from prebiotic chemistries [231], but see also [258,259], lipid micelles have been demostrated to form spontaneously through a self-assembly process starting from free fatty acids and minerals [98]. Also, with the availability of free ribonucleotides from the primordial non-enzymatic reactions (and this seems to be plausible, see [206,240,20], it was shown that RNA polymers can form in the presence of montmorillonite, a clay present in hydrothermal vents [75,126].

4. Primordial (biotic) metabolism

Thus, from a protocell, encapsulating a primordial metabolism and in the presence of a primitive, but catalytic, genetic material, the challenge next turned to understand how the complex biomolecular networks that underpin life took on the forms that we can observe today [37,77]. Indeed cellular metabolism is arguably one of the earliest biological networks and, as such, has been extensively studied and acts as a fantastic model system for studying network evolution. What has been elegantly described as the "complex fabric of (molecular) interconnections" is responsible for the functioning, survival and reproduction of cells [37]. Thousands of different biochemical processes are linked in highly tailored systems that have been acted upon by billions of years of evolution. These reactions are also highly compartmented – partially as a mechanism of metabolic regulation [245,4] – necessitating the evolution of complex transport systems. Indeed genome scale models of the model plant Arabidopsis suggest the need for a phenomenal 772 transporters in a 1200 reaction model [177]. The focus of this section (4) will be the evolution of cellular metabolism, starting from the current hypotheses made about the origin of life and the general theories to explain the assembly of metabolic pathways; in the next section (5), we will cover the emergence of primary and secondary metabolism of plants presenting some recent examples about how genetic diversification and catalytic promiscuity - two typical phenomena of secondary metabolism - may affect the evolution of novel metabolic traits and contribute to chemical diversity [157].

When life began on Earth, metabolism is believed to have centered on very few chemistries and simplified reaction pathways which likely formed in emergent replicating or organismal units and later developed as primitive cells [37]. It is important to note that the reactions of this primordial metabolism as well as the more ancient prebiotic chemistries are, essentially, non-existent today and we thus depend upon our knowledge concerning the conditions which prevailed on the primitive Earth. As such, they remain speculative, by contrast to chemistries that arose via enzymatic catalysis which can be explored at protein, RNA and DNA levels by coupling knowledge from biochemistry, structural biology, genomics and modern phylogenetic analyses. Despite having these tools at hand, we remain far from understanding the origin of life as several of the main problems associated with it remain unsolved. Ralser defined these eloquently in 2014 [211]. In essence they are (i) *how these early biomolecules form and reach life-compatible concentrations*; (ii) *how did genetics with its inheritable evolutionary selection come into place* and (iii) *how did metabolism evolve and facilitate the prototrophy of cells*.

Having briefly covered the hypotheses made on the first two points raised by Markus Ralser earlier in this review, we now turn to the theories made on the evolution of metabolic networks from the primordial metabolism of the universal common ancestor of all life forms.

Upon the biochemical routes of primordial metabolism, in fact, several evolutionary pressures may have acted to shape the genomes modifying the structure and regulation of the metabolic networks. Several hypotheses have been formulated so far - none of them mutually exclusive in the assembly of complex networks - to explain the emergence of novel metabolic traits. We will present these hypotheses below to later focus on the genome dynamics giving rise to the secondary metabolism of plants.

Glossary

- **adaptation (signature of –)**: a specific sequence pattern, which can be detected at the DNA level, that distinguish a locus under selection from one evolving neutrally; the tests to identify these genomic footprints typically compare the frequency of sequence polymorphisms, the extent of linkage disequilibrium and/or population differentiation under the null hypothesis of neutrality.
- **balancing selection**: the process by which two alleles of a single gene are maintained in a population at intermediate frequency, higher than that predicted by genetic drift alone, due to heterozygote advantage.
- **compartmentation**: in the context of metabolic biology, the distribution of metabolites and enzymes (and thus, metabolic pathways) across different subcellular structures. Macrocompartmentation refers to the differential allocation of metabolites/enzymes between the cytosol and other membrane-bound organelles (as is often the case in cofactor metabolism) while microcompartmentation denotes the association of cytosolic enzymes with cytoskeleton (actin, tubulin) or their localization on the surface of organelles or endomembranes, [245].
- **duplicate loss**: the process by which duplicated copies of a gene are lost across evolutionary time. In plants, this is the common fate of most of the paralogs. A form of duplicate loss is pseudogenization, when one of the paralogs is retained in the genome, but is no longer functional due to a loss-of-function mutation.
- **evolve and resequence (E&R)**: the measurement of allele frequencies between an ancestral (base) and selected population with Pool-Seq within an experimental evolution setting.
- **experimental evolution**: the study of evolutionary processes over multiple generations in populations where the settings are set and controlled by the experimenter. It involves the application of selection pressures where only the individuals exceeding a certain phenotypic treshold are allowed to reproduce.
- **gene duplication**: a mutational event which implies the duplication of a particular genomic region (or whole genomes in case of polyploidization) into a different position in the genome. Several genetic mechanisms can originate gene duplicates (paralogs): in addition to whole genome duplications, paralogs may form following tandem/segmental duplications, or may derive from the activity of transposons. The contribution of each of these mechanisms on the amount of gene duplicates typically vary widely across various plant genomes.
- **gene fusion**: the process through which a new gene is formed through the fusion of multiple, previously separated ORFs (open reading frames, i.e. stretches of DNA bordered by a start and a stop codon). Fused genes may result from interstitial deletions (deletions of intergenic space) or from larger chromosome mutations (translocations, inversions, etc.).
- **genetic drift**: the change of allele frequencies due to random factors.
- **genome scan**: a survey of the genome-wide DNA polymorphisms across members of a population.
- **homology**: similarity of phenotypes in different lineages due to common ancestry. In evolutionary biology, the term homology can be used to define a common ancestral origin for any structure (e.g., organs, morphologies, genes, etc.): two genes are thus **homologs**, for example, if they derive from the same common ancestor [271].
- **linkage disequilibrium (LD)**: nonrandom association of alleles from different loci. It is usually measured as:

$$r^2_{AB} = \frac{(\pi_{AB} - \pi_A \pi_B)}{\pi_A \pi_B \pi_a \pi_b}$$

where: A and a are two alternative alleles at locus 1, B and b are two alternative alleles at locus 2, $\pi_{A/B/a/b}$ are their respective allele frequencies, π_{AB} is the haplotype frequency (for the AB allele combination) and r^2_{AB} is the LD between A and B. Mutation rate, selection, demography (e.g. migration, admixture, bottleneck), genetic drift and mating system (outcrossing Vs self-fertile species) are all factors influencing the extent of LD in natural populations [76].

- **maximum likelihood (ML)**: a statistical method to infer unknown parameters of a probability distribution. In phylogenetics, ML is both a method to infer tree topologies but can also be used to test evolutionary processes when the tree topology is known [288]. Typical parameters include, for example, the tree topology itself, branch lengths and substitution rates. Likelihood is simply defined as a quantity proportional to the conditional probability of observing the data (D), given the model (M), i.e., P(D|M). In the inference of phylogenetic trees, likelihood scores are calculated for each possible branching pattern; the maximum value (ML) is the highest score associated to the specific branching pattern which maximises the probability of observing the data (i.e., the individual site patterns of the multialignment) given the substitution model [72,116].
- **maximum parsimony (MP)**: a method to infer phylogenies based on minimising the number of character changes; historically it was the first approach to reconstruct ancestral sequences [239] and was later applied also for tests of neutrality in protein coding sequences [174]. MP provides relatively reliable reconstructions of ancestral states only in case of recent divergence [293].
- **mosaic origin**: as a consequence of the processes of endosymbiosis of cyanobacteria and α-proteobacteria, and subsequent gene transfer between the endosymbiont and the nuclear genome, some of the metabolic pathways in eukaryotes display signatures of mixed evolutionary origin. Typical examples of mosaics are the glycolysis and Calvin cycle pathways in plants [167], the pyrimidine biosynthesis [188] and the heme biosynthesis in photosynthetic eukaryotes [193]. In addition to endosymbiosis, also horizontal gene transfer has contributed to shape the mosaic structure of some specific pathways in photosynthetic eukaryotes (e.g. chlorophyll degradation, [192]).

- **Oparin-Haldane hypothesis**: a theory regarding the origin of life proposed independently in the 1920s by the Russian biochemist Aleksandr Oparin and British geneticist John B. S. Haldane. Both believed that if the conditions of the primordial atmosphere contained very low level of oxygen (''reducing atmosphere"), then organic compounds could have been formed directly from inorganic molecules with the supply of external energy sources (e.g. UV radiation, lightning). Haldane coined the term ''prebiotic soup" to denote the primordial aqueous environment, rich in ammonia and methane, in which the first organic compounds appeared; later, by further developments, these early organic metabolites could have acquired lipid membranes (Oparin's ''*coacervates*") to form the first living cells [151–152].
- **"one gene-one enzyme" theory:** the idea formulated in the 1940s by American scientists George W. Beadle and Edward L. Tatum that one gene specifies the synthesis of a single enzyme in a metabolic process. The idea spurred from Beadle & Tatum work with *Neurospora crassa*, a mold which could be grown easily in the laboratory with a simple growth medium (sugars, inorganic salts, vitamins). Beadle and Tatum irradiated spores of Neurospora and verified if the derived mutant strains were able to grow on complete and minimal media. Those mutants that failed to grow in the minimal medium were tested sistematically to identify which compound they were unable to synthesize [18]. They found that some mutant strains required the addition of specific aminoacids in order to grow in the minimal media: this result allowed them to associate the mutations with the production of enzymes in a specific metabolic pathway. The work of Beadle and Tatum represented the foundation of biochemical genetics and was awarded the Nobel prize in Medicine or Physiology in 1958 (shared with Joshua Lederberg). Although this theory was solidly verified, we now know it offers an oversimplistic view of molecular biology: not all genes encode for enzymes, and many enzymes act only as multipolypeptide complexes, where each single component is encoded by a distinct gene.
- **orthologs**: genes in different species originating from a single ancestral gene in the last common ancestor of the species under comparison [141,142].
- **paralogs**: genes in a single species originating from an event of gene duplication. A broader definition does not make any distinction whether paralogs reside in the same genome (i.e., in the same species) or with respect to when the duplication emerged [142].
- **positive selection (or directional, also known as darwinian selection)**: the increase in frequency of mutations (to higher prevalence in the population or even fixation) conferring higher fitness (advantageous mutations).
- **primary metabolism**: the ensemble of metabolites (and their pathways) which are essential for growth and reproduction of all organism. It usually includes the central (carbon) metabolic routes involving the four main classes of biological molecules (carbohydrates, proteins, nucleic acids and lipids). The number and structure of primary metabolites are largely conserved across the tree of life.

- **primordial metabolism**: the set of simplified reaction pathways which characterised the early heterotrophic living systems emerging from the primordial soup (or from the highly reactive microenvironments in hydrothermal vents). The term may be used in a more general sense to include also the ensemble of prebiotic chemistries that predated the emergence of the first life forms (i.e., the last universal common ancestor of all living forms, around 3.5 billion years ago). In its broader sense, primordial metabolism thus covers the chemistries of the prebiotic phase until the simple metabolism of the last universal common ancestor (from around 4.2 to 3 billion years ago).
- **purifying selection (or negative)**: the process by which deleterious mutations (decreasing fitness) disappear from a population.
- **relaxed selection**: the weakening or elimination of a regime of natural selection which previously acted to maintain the expression of a trait; it usually follows a change in environmental pressure or genetic make-up, such as following gene duplication [150].
- **secondary (or specialized) metabolism**: the ensemble of metabolites (and their pathways) which mediate the interactions of an organism with its surrounding environment. Secondary metabolites typically function in various ecological interactions (in response to pathogens, as attractants for pollinators, etc.), although some may regulate processes in growth and development [254,160], a role which was formerly typically assigned to primary metabolites only (see **primary metabolism**). Secondary metabolites occur in a much wider variety of structures and their number vastly exceed that of primary metabolites. In light of their ecological role in improving the fitness of the organisms, secondary metabolites may represent adaptive traits. Their fragmented distribution patterns across the tree of life could thus reflect adaptations to particular ecological niches. Pathways of secondary metabolism are present in microbes, fungi and plants. Many animals (both Invertebrates and Vertebrates) synthesize defensive metabolites with highly similar structures to the typical secondary metabolites of plants. Some quinazoline alkaloids, for example, which are widely accumulated in plants of the Rutaceae family (Citrus), are also synthesized, with similar structures, by Millipedes, Ascidians (sea squirts) and Amphibians. In all cases, they serve defensive purposes and protect the organism from predators [175]. The presence of very similar metabolites in distant taxa should consider the possibility that specific metabolic traits could have emerged initially to serve a particular function, but were later exapted [87] to serve a different one (e.g., sex attraction and defense).
- **selective sweep**: the decrease in nucleotide diversity within a population in regions flanking a locus under positive selection. *Hard sweeps* are usually defined as those following the emergence of a new mutation which is immediately beneficial (and thus, selected) in a population; *soft sweeps*, by contrast, are originated by selection on standing variation; as such, they emerge in neutral or nearly neutral loci whose derived phenotypes are subject to a change in selection pressures.
- **site frequency spectrum**: histogram representing distribution of allele frequencies in a population.
- **stabilizing selection**: in the context of population differentiation, the process by which the same allele is selected in different subpopulations.

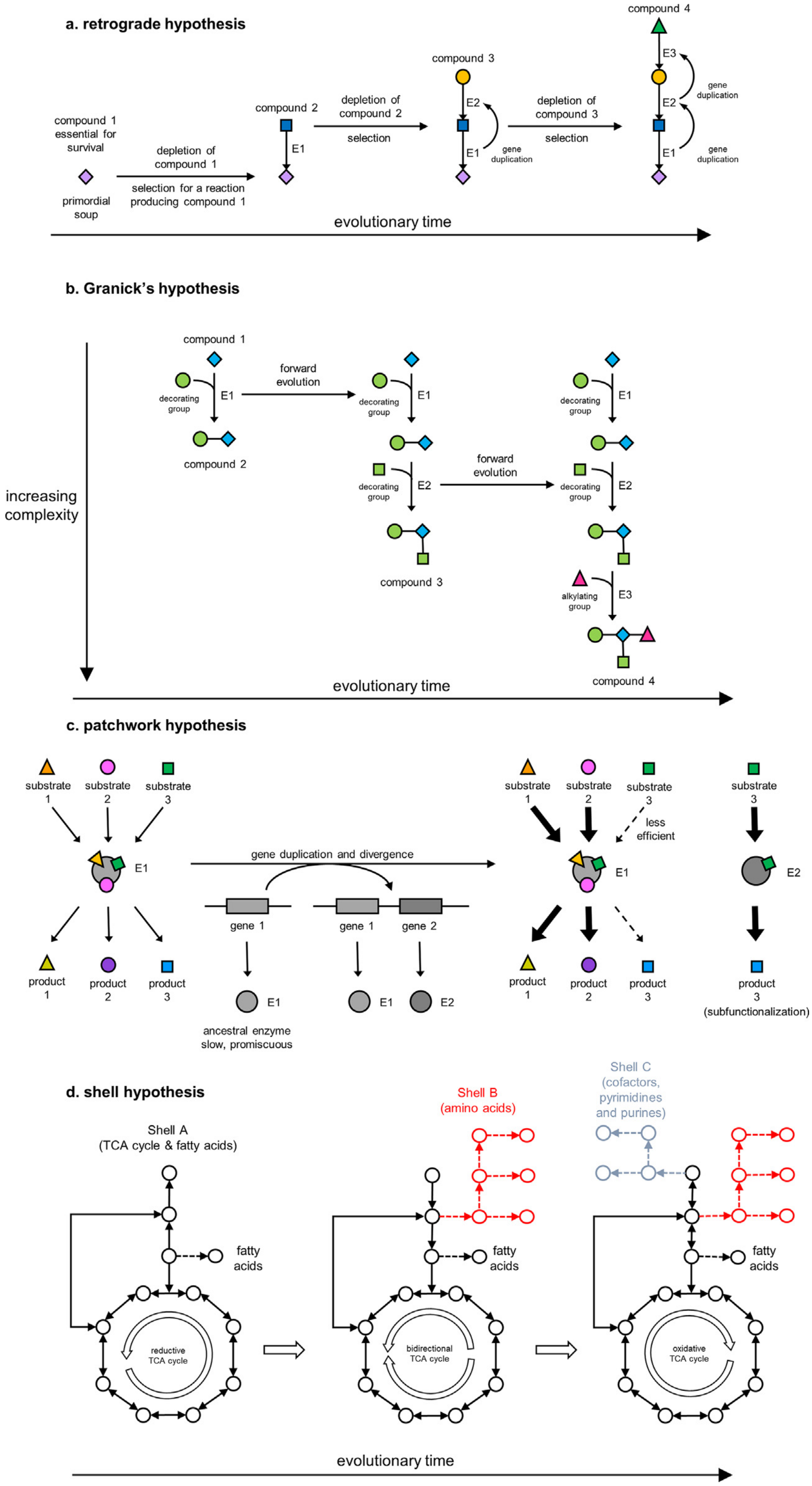

Fig. 1. Schematic diagrams representing the main hypotheses for the evolution of metabolic pathways. In the retrograde hypothesis (a), metabolic pathways are supposed to originate with sequential gene duplications starting from gene catalyzing the last step of current pathways. Depletion of the compounds present in the primordial soup may have originated a selection pressure leading to the survival and reproduction of the primordial cells able to produce the depleted compounds; this process could have then been repeated sequentially, in a backward direction, until the establishment of contemporary pathways. In Granick's hypothesis (b), pathways would have been assembled in a forward direction, from simple precursors to more complex products. Under this model, the older genes across evolutionary timescale would be represented by those catalyzing the earlier steps in contemporary pathways. In our diagram, the decoration and alkylating steps are examples of reactions adding complexity to the initial precursor and to the pathway intermediates. In the patchwork hypothesis (c), ancestral genes encoding promiscuous enzymes could have expanded the metabolic capabilities of primordial cells through gene duplication and subsequent divergence. A possible fate for an event of gene duplication is subfunctionalization (c), in which the catalytic activities of the ancestral gene are divided among the paralogs. In our example, following divergence of the duplicated genes, one of the ancestral reactions is taken on by one of the paralogs. In the shell hypothesis (d), evolution of metabolism can be traced back to the consecutive additions of distinct metabolic pathways. The core central pathway (reductive TCA, fatty acids biosynthesis), i.e., shell A, predated the addition of nitrogen metabolism in shell B; sulphur and cofactor metabolism were later added as shell C. As more pathways are added, the inner shells remain nested in the network as remnants of the earliest metabolisms. Abbreviations: E1, enzyme 1; E2, enzyme 2.

The hypothesis of the origin of life from the pores of the deposits formed by hydrothermal vents posits that early organisms were heterotrophic and performed only minimal biosynthesis. The suggestion is that the increasing number of primordial cells would lead to an exhaustion of amino acids in the primordial soup thus imposing increasingly stronger selective pressure favoring those cells that had evolved the capacity to synthesize such molecules themselves. This, ultimately, led to the thousands of extant reactions and transport processes linked into pathways and networks that characterize life as we know it [77]. The presence of such highly complex metabolic networks raises the question as to how they originated starting from ancestral genomes that were likely only composed of a couple of hundred genes [187]. Different molecular mechanisms have clearly been behind both the expansion and the shaping of early genomes and metabolic pathways including, but not limited to, gene duplication, gene fusion and horizontal gene transfer. Whilst horizontal gene transfer (HGT) has been demonstrated in plants [212,25,192], it appears to be relatively uncommon and as such we will largely restrict discussion here to the other two mechanisms. A case of HGT, just to cite one example, involved phenylalanine ammonia lyase, the gene leading to the evolution of phenylpropanoid metabolism: the appearance of phenylpropanoids was in fact a key adaptation of plants to the life on terrestrial habitats [65,255].

Starting from Ohno's classical work, **gene duplication** is instead considered an immensely important force driving the evolution of life [194]. Comparative analysis of sequenced archaeal, bacterial and eukaryal genomes revealed that a very large proportion of genes are the outcome of duplication events [56,209] that either predate or follow from the appearance of the last universal common ancestor [119]. Moreover, due to the preponderance of whole genome duplications in the plant kingdom this is especially the case for plants [198,74,218,209]. It has been suggested that current genomes may be the result of DNA arrangements involving a limited number (20–100) of "starter types" [153], although how these originated remains unclear. The opposite phenomenon, i.e. **gene fusion**, wherein independent cistrons are fused to form bi- or multi-functional proteins is also a common evolutionary process. It provides a mechanism for the physical association of different proteins (be they catalytic or regulatory or both), and frequently involve genes encoding sequential steps of a pathway [66,103], with such fusions being reported for tryptophan and histidine pathways in prokaryotes [279,71], as well as the Calvin-Benson cycle [204], glycolysis [257], and secondary metabolic pathways in plants [277].

4.1. The retrograde hypothesis

As early as 1945, and based on the **Oparin-Haldane hypothesis** and the **one-to-one correspondence between genes and enzymes** [18], see glossary), Horowitz proposed that biosynthetic enzymes emerged via gene duplication that took place in the reverse order to that found in current pathways [109]. This theory posits that if compound A was limiting in the primordial soup, then its synthesis from its direct precursor would be the first (enzyme) catalyzed reaction to be required. Once this was established, this would likely lead, in turn, to selective pressure on the direct precursor, requiring the evolution of another reaction to synthesize this. Over many iterations, a pathway could thus be built from the final product(s) backwards up to the initial precursor [77] (Fig. 1a).

4.2. The Granick hypothesis

A direct alternative hypothesis, that is little discussed [70], is the proposal that biosynthetic pathways develop in the forward direction [90]. The central theme of this hypothesis is that biosynthesis of certain end-products could be explained by the forward evolution from relatively simple precursors. The model thus predicts that simple biochemical compounds predated more complex ones and that the enzymes catalyzing earlier steps of a metabolic role are older than those acting later in the pathway (Fig. 1b). For this hypothesis to hold it is imperative that each of the intermediates are of use for the organism [70]. This seems to hold true for heme and chlorophyll biosynthesis as well as isoprene biosynthesis, but problems arise for pathways such as purine and branched-chain amino acid biosynthesis for which the intermediates are of no apparent use [70].

4.3. The patchwork hypothesis

The patchwork hypothesis [120], by contrast, posits that metabolic pathways may have been assembled via recruitment of primitive enzymes that could react with a wide range of chemically related substrates [77]. Such relatively slow, non-specific enzymes enabled primitive cells containing small genomes to overcome their limited coding capacities. Gene duplication and neo-/subfunctionalization is the proposed mechanism underlying this recruitment of an ancestral enzyme to serve novel functions in emergent pathways (Fig. 1c). This hypothesis finds support from both the analysis of sequenced genomes and from directed evolution experiments [70]. It has, furthermore, been invoked to explain the evolution of several processes including the urea and TCA cycles as well as several pathways of amino acid biosynthesis and even more recently evolved pathways [77].

4.4. The shell hypothesis

A fourth hypothesis which needs to be considered is the shell hypothesis, put forward by Morowitz [184]. This postulates that the reductive citric acid cycle was the earliest pre-biotic self-replicating chemistry and it evolved in the absence of enzymes. This cycle is then believed to have led to an "energy amphiphile" core that enabled the discovery of new carbon-based chemistries upon which were built further chemistries (or shells). This hypothesis assumes that pre-biotic chemistries remain "imprinted" in modern metabolism as relics and suggests that biogenesis of metabolism manifested itself in a hierarchy of nested reaction networks of increasing complexity [37]. Indeed, it predicts the formation of the TCA cycle, glycolysis and fatty acid biosynthesis in shell A preceded that of the introduction of nitrogen via amino acids in shell B, sulphur in shell C, with ring closure giving rise to purines, pyrimidines and many other cofactors (Fig. 1d). Indeed the energy amphiphile core of this hypothesis is consistent with earlier proposals that life evolved on pyrite [264], although the gradual addition of shells, and in particular the late account for sulphur chemistry, are not consistent in light of the recent scenarios for a core organo-sulphur prebiotic metabolism [84].

5. The emergence of primary and secondary metabolism

Whilst all of the above theories have had their supporters, the patchwork recruitment scenario is arguably the best supported by accumulated evidence (see [226]) and [37] for details; we review additional support for the patchwork model with respect to other theories further below). To provide just a handful of examples here, enzymes with $(\beta\alpha)_8$-barrel fold structure have been found to catalyze similar reactions across pathways [53]. Similarly, analysis of the entirety of *E. coli* metabolism revealed a genuine mosaic with widespread recruitment of protein domains [249]. This is also clearly the case in many plants: the endosymbiotic events that gave rise to the mitochondrion and chloroplast (and possibly the peroxisome) during the evolution of the eukaryotic lineage duplicated enzyme functions, and unless there was a specific selective advantage, one of the duplicated enzymes was lost over evolutionary time. In the absence of a specific selective pressure, **duplicate loss** was essentially random, which has led to mixed **compartmentation** of metabolic pathways and a **mosaic evolutionary origin** for many pathways, irrespective of their compartmentation [245].

5.1. Gene duplication and neofunctionalization

Whilst the evolution of **primary metabolism** in plants is dominated by events of endosymbiosis, **gene duplication**, with the different fates of the resulting paralogs (including gene loss, see for example the fact that plants do not possess a complete urea cycle [5] is considered instead to be the main driver of the diversification of **secondary metabolism** [219,40]. Several models are usually associated to explain the emergence of novel gene functions following (or predating, as we will see below in case of the escape from adaptive conflict) gene duplication [194,117,52,59,96]). Despite the fact that the majority of gene duplicates is lost over evolutionary time, and that most of those that are retained are subject to strong **purifying selection**, a few retained paralogs may initially be instead under **relaxed selection** and may accumulate (potentially adaptive) mutations [165]. This model (neofunctionalization) thus imply that the original gene keeps its ancestral function, while a new function emerges in one of its derived paralogs, which can be maintained in the genome because of **positive selection** on the new function [104,39,275]. Within the context of metabolism, typical polymorphisms associated to the emergence of novel enzymatic functions may be located in the coding region, resulting in a shift in substrate preference and/or catalytic activity; in other cases, metabolic novelty may be driven by differences in transcript abundances of structural or regulatory genes, as a result of genetic polymorphisms or epimutations in the promoters or other regulatory regions [274,210].

Thus, accompanying gene duplications, these non-deleterious, extremely rare mutational events generating expansion of the existing pathways of primary metabolism could have emerged and be eventually fixed within populations [273] during the major expansion of plant metabolism, which occurred concomitantly with the colonization of terrestrial habitats, around 500 million years ago [65,255]. Many metabolic lineages of secondary metabolism could have emerged, at least initially, through such rare events. Then, following various modalities of gene duplications (either tandem, segmental and retroduplication), some of the retained, additional alleles may be occasionally subjected to relaxed constraints such that at least one copy is able to accrue considerable mutations, leading to greater mechanistic elasticity, entirely new substrate specificities or, more generally, to an alteration of enzyme activity [179]. This explains why expanded substrate recognition, flattened catalytic landscapes and multiple products from a single enzyme are common in specialized metabolism. Although the distinction between genes active in primary and secondary metabolism is, from a functional perspective, relatively indistinct, several evolutionary considerations support this classification. As we have said (but see also the examples further below), secondary metabolism genes originated from gene duplications of pre-existing genes (often from primary metabolism, [40]), have fragmented phylogenetic distribution [276], and are usually characterised by a higher rate of gene birth/loss with respect to the enzyme genes of primary metabolism [185]. Also, they may occur in metabolic clusters [191], are characterised by large expression variation (plasticity) and distinct correlation properties [137,97,253,278]. All these characteristics were used in a machine learning approach to train a model of the Arabidopsis genome which was able to predict whether a particular gene is part of primary or secondary metabolism. The model reached a high prediction accuracy integrating gene information related to their degree of conservation, genomic location, expression profiles, the presence/position of epigenetic marks and protein domain composition [182].

A classical example, which summarises the co-occurrence of several of the phenomena discussed above, comes from the study of methylthioalkylmalate synthase (MAM), an enzyme involved in the elongation steps of methionine, leading to the synthesis of alkylglucosinolates, a class of defense metabolites found in Cruciferae (Arabidopsis and related mustard species). Phylogenetic analyses support the origin of MAM from the duplication of the ancestral α-isopropylmalate synthase (α-IPMS), a gene active in the synthesis of leucine [55]. MAM duplicates evolved distinctive characteristics: they lack the C-terminal domains, typical of α-IPMS, which confer the capacity to be feedback-inhibited by the final product of the ancestral pathway, leucine; moreover, two amino acid changes in the active site shifted the specificity of the duplicates towards the MAM substrates, although both enzymes maintained marginal catalysis with their non-preferred substrates [23,55]. This case thus represents an example of how paralogs of genes of central metabolism may undergo: i) neofunctionalization involving extensive sequence variation, shifting substrate preference and ii) relaxation of ancestral functional constraints, with the loss of feedback inhibition. Both processes thus contributed to the evolution of MAM genes starting from a gene active in **primary metabolism**, giving rise to the synthesis of novel metabolic traits.

Another example, that of the gene encoding for homospermidine synthase (HSS), recapitulates well the processes leading to the appearance of evolutionary novelties in plant metabolism. Homospermidine is a widespread polyamine in the plant kingdom; in several families of Angiosperms it is used as the substrate for the synthesis of pyrrolizidine alkaloids (PAs), a class of secondary metabolites used as feeding deterrents by the plant. In Convolvulaceae, independently from the other families which also accumulate PAs, phylogeny strongly supports the origin of HSS genes from a single duplication of deoxyhypusine synthase (*dhs*), a gene of primary metabolism involved in the post-translational regulation of the eukaryotic initiation factor (elF5a). DHS transfers an aminobutyl moiety, derived from putrescine, to the lysine residues of elF5a, forming the rare aminoacid deoxyhypusine. In PA-free species of the Convolvulaceae, there were no apparent paralogs of *dhs*, or, when these were detected, gene duplication gave rise to non-functional HSS copies (pseudogenization). On the other hand, in species accumulating PAs (e.g., *Ipomoea neei*), duplication of the ancestral *dhs* generated paralogs which later accumulated both functional (i.e., non-synonymous changes in the aminoacid sequence) and regulatory divergence (i.e., tissue-specific expression), acquiring the novel catalytic activity typical of HSSs [127].

5.2. Gene duplication and the escape from adaptive conflict (EAC)

An alternative process which is invoked to explain the evolution of novel function (and the maintenance of the ancestral one) is the escape from adaptive conflict (EAC), which occurs when a new function emerges in a progenitor single-copy gene before its duplication. Under this model, both the ancestral and the novel function are maintained in the progenitor, but negative intragenic epistasis prevents improvement of both functions (advantageous mutations in the ancestral, bifunctional gene are removed by natural selection). Gene duplication thus resolves the conflict separating the fates of the paralogs, which can then accumulate mutations to improve separately the ancestral and the derived function [105].

An example of escape from adaptive conflict probably operated in Convolvulaceae, during the evolution of dihydroflavonol-4-reductase (DFR), a gene in the anthocyanin pathway which acts downstream of dihydroflavonols [59], but see also [12]. In *Ipomea purpurea*, *DFR* is present in a locus made of three tandem copies which are all expressed across several tissues. Analyses of sequence evolution and tests of substrate specificities of DFR enzymes from various Ipomea species partially support EAC as the process by which novel function emerged during the evolution of *DFR*. Before the first round of duplication, in fact, the lineages of single-copy *DFR* genes are subject to purifying selection, indicating that the gene is under strong functional constraints (mutations in this lineage are thus considered to be mainly deleterious). By contrast, these constraints were apparently released after the first round of gene duplication, when the paralogs show strong evidences of repeated positive selection, indicating the presence of adaptive evolutionary changes which improved, separately, both the ancestral and the derived function. Although there are no large-scale studies on the impact of EAC in plants, and alternative models may also account for the fate of *DFR* genes in Ipomea [12], EAC remains a possible model describing how new protein functions may evolve while maintaining, at the same time, their ancestral functions. More recently, biophysical methods, incorporating protein stability and population genetics data have been developed to model the most reasonable mode of protein evolution [233].

5.3. Evolution of protein functional specialization

Although the plant studies cited above, along with others [163,180,203,48,294], all afford extraordinary insights about the functional diversification of extant enzyme genes, and additonally provide support for the role of the patchwork hypothesis, the details about the historical trajectories leading to these functional specialization remain relatively unexplored. This is because we generally lack information about the ancestral genes, so that the progressive effects of the historical nonsynonymous substitutions are unknown. With the revolutionary works of Joseph Thornton [251,100,242] and the recent improvements in phylogenetic methods (summarised in [64]), however, the field of ancestral protein resurrection has made tremendous recent progresses. Indeed, it is now possible to resurrect the most probable sequences of the ancestral alleles from a well-supported phylogeny and characterise the function of their respective products.

In such evolutionary contexts, one of the first studies to assess the role of gene duplication on metabolic specialization was the analysis of fungal maltases. These enzymes, encoded by genes of the *MALS* family, are able to hydrolyse α-disaccharides (e.g., maltose, sucrose, etc.) into monosaccharides [262]. The members of this family underwent both recent and distant duplications and display today functional specialization in the hydrolysis of specific disaccharides. As a result of these duplications events and subsequent diversification of the paralogs, yeast species have variable numbers of *MALS* genes and, consequently, different activities towards α-disaccharides: if *S. cerevisiae* (baker's yeast), which hosts seven *MALS* genes, can hydrolyse most sugars, the related species *S. kluyveri,* for example, lacks the ability to hydrolyse melezitose, turanose and isomaltose, due to the absence of two *MALS* paralogs. The resurrection of the ancestral alleles and the functional characterization of the various maltases allowed them to investigate how from a promiscous and relatively inefficient ancestral glycosidase, the function to hydrolyse sugars was differentially partitioned into the descendants through various duplication events. Along these trajectories, some amino acid substitutions led to the increase of the catalytic efficiency of the ancestral enzyme, while others shifted the specificity toward a different disaccharide substrate. The high activity of MAL12 and MAL32 (two extant maltases) towards maltose-like substrates, for example, was the result of a gradual optimization of catalytic activities already present in the ancestral maltase. On the other hand, the capacity to hydrolyse isomaltose-like substrates, typical of a separate clade of the *MALS* genes (*IMA1-4*), resulted from a catalytic specialisation which became preponderant after one of the recent duplication events, although a marginal capacity to hydrolyse isomaltose-like sugars was already present in the ancestral maltase. The evolution of fungal glucosidase thus highlights the principle that various models of gene duplication (neo-/subfunctionalisation and EAC, [96]), rather than acting singularly to account for the specialization of extant enzymes, actually interweaved along evolutionary histories to give rise to metabolic diversification.

In plants, ancestral protein resurrection approaches have been pioneered by the group of Todd Barkman, who has studied the functional specialization of the SABATH (Salycilic Acid/Benzoic Acid/Theobromine) gene family of methyltransferases [112] and the convergent pathways leading to the synthesis of caffeine [111]. We will focus here on their first investigation, as an example of how resurrection and characterization of the SABATH ancestral proteins can provide insights about the historical changes at the basis of enzyme functional evolution. SABATH genes are present in several Angiosperms, have accumulated various duplications, and show today enzymatic activities towards benzoic acid and a wide range of its analogs (e.g., salycilic acid, dihydroxybenzoic acids, nicotinic acid, *o*-anisic acid etc.). All these enzymes are active today in the production of flower scents and in the synthesis of various defensive molecules against pathogens. The resurrected ancestral SABATH enzyme showed high activity towards benzoic acid, but several subsequent amino acid substitutions shifted substrate preference to salycilic acid. Of these changes, one substitution in the active site (the His to Met change at position 201) showed a clear signature of positive selection.

Moreover, along the other branches of the SABATH phylogeny, the resurrection of the intermediate ancestors uncovered how the specialization of extant enzymes emerged. In general, latent activities with non-preferred substrates in one ancestral enzyme became instead preponderant with the accumulation of specific amino acid changes in one of the daughter enzymes following duplication. In some cases, these activity shifts could be reconducted to signatures of positive selection on specific amino acid substitutions located in the active site. As with fungal maltases, it is difficult to classify the specialization of SABATH enzymes into a strict model for protein evolution under gene duplication. In this case, along the various branches, ancestral functions were either optimized, or shifted towards different substrates, and also partitioned among different descendant genes; with these three processes occurring concomitantly along evolution.

6. Testing evolutionary hypotheses

The studies of extant diversity in plant secondary metabolism necessarily raise the question of how such natural variation originated, but also whether if - and to which extent - this diversity might be considered as the results of adaptive processes. This is one of the key challenges in evolutionary genetics [147], part of the larger theme about the emergence of evolutionary innovations [57], and directly relates to our (largely) incomplete understanding of the consequences of mutations on the fitness of organisms [122]. In the next sections we will try to cover some of the approaches which can be used to detect **signatures of adaptation** at the genetic level. The statistical tests presented below have variously contributed to reinforce the patchwork hypothesis as perhaps the pervasive mechanism to explain the ramifications of secondary metabolism as we know it today. In plants, several cases of pathway ramification originated from the initial recruitment of neofunctionalized paralogs whose activity conferred some form of (higher) adaptive value. The evolutionary pressures acting on these genes were measured across the branches of phylogenetic timescale with various forms of test of selections involving the codon-based models of sequence evolution (d_N/d_S or related tests, see for example [23,62,127,102]. At a higher level, also genomic scans and approaches based on population subdivision have provided evidences of selection on a range of alleles variously involved in diversification/expansion of plant metabolism. Some of these evidence points to a role of metabolic genes as direct targets of selection during plant domestication and breeding, rather than being simply seen as "passive", neutral variants merely linked to the "real" target locus effectively under selection [281,45,22,295,284].

The success of the patchwork model relies perhaps on relaxing the assumptions the other hypotheses make with regard to the age of the enzymes involved during the assembly of a metabolic pathway. Both retrograde and Granick's forward-evolution hypotheses, for example, posit ancestry relationships in the enzymes along the steps of a metabolic pathway (e.g., in Horowitz's *retro*-evolution model, the last enzyme in the pathway is the ancestral one): these relationships are weakly supported today both in light of the mosaic distribution of protein folds across metabolic genes [36] and because of the unequal distribution of selective constraints in the genes along a biochemical pathway [195,51]. Also, in plants, genomic comparisons of the paralog copy number in several classes of metabolic gene families, across a large phylogenetic timescale, did not show a clear trend between the position of a specific gene in a pathway and its time of emergence [255]. More generally, it remains difficult, if not impossible, to assess the real contribution of retro- and forward evolution models in the assembly of biochemical pathways: these hypotheses were originally formulated in the context of the evolution of new functions starting from the primordial soup: these reactions are basically nonexistent today and can be re-created only through a synthetic biology/experimental evolution framework. As such, these theories may have some validity in the initial expansion of primordial metabolism, but have today limited explanatory application in providing support for the formation of a highly branched and compartmentalized metabolic network.

The patchwork hypothesis, on the other hand, is consistent with the various models of protein functional evolution following gene duplications [24,96,164,194,266], and with the evidences coming from directed evolution experiments on the evolution of promiscuous activities [1]. The picture emerging from the mechanisms underlying the assembly of metabolic pathways is thus adding increasingly solid support to the patchwork model of evolution, with some cases of specific pathway evolution originating instead from horizontal gene transfer [192].

As it will become clear in the examples below, tests of adaptation represent powerful statistical tools both to test evolutionary hypotheses or generate novel ones about the relationship between genotypes, phenotypes and environment; however, sequence analyses and phylogenetics, alone, are not sufficient and clearly need to be integrated within wider ecological experiments in natural settings to fully understand the genetic basis of adaptation [16,207].

The possibility to infer signatures of adaptation from sequence data derives from the work of Motoo Kimura and his neutral theory of molecular evolution [134,135]. This theory states that: "[...] *the overwhelming majority of evolutionary changes at the molecular level are not caused by selection acting on advantageous mutants, but by random fixation of selectively neutral or very nearly neutral mutants through the cumulative effect of sampling drift (due to finite population number) under continuous input of new mutations* [...]" [136]. One of the direct consequences of the neutral theory is that the largest part of sequence polymorphisms that we observe during evolutionary time are fixed by drift and confer no fitness advantage; hence, to distinguish neutral from adaptive loci the principle upon which tests of selection are based is to compare the distribution of empirical data against the null hypothesis of random genetic drift [189,261].

In contrast to the predictions of the neutral theory, however, the advent of whole **genome scans** has led many researchers to declare footprints of (putative) selection as being relatively common in many plant genomes [101,295,209]. In some plant genera (e.g., Helianthus [17], Capsella [235]), the proportion of adaptive substitutions (α) may well exceed values of 0.2–0.3 [171]. These figures certainly represent an overestimation of the proportion of loci which constitute true adaptive alleles [250]. It is known, for example, that specific DNA patterns may simply arise by chance, or derive from hitchhiking of false positive SNPs (neutral variants) in **linkage disequilibrium** with the polymorphism under selection or, also, by taking into account erroneous models of demographic history [252,292]. Also, even when a given locus carrying a signature of selection has been demonstrated to be causal for a certain phenotype, it is still necessary to rule out selection acting on its pleiotropic effects [197,222,16]. For these reasons, we here reserve the definition of "adaptive alleles" to those i) having a functional and causal relationship to phenotypes increasing fitness, and ii) whose frequency has been shown to change, in the expected direction, following selection on its focal trait(s) [16]. Along these lines, obtaining convincing evidences of natural selection acting at the sequence level has proven incredibly difficult; as such, true adaptive alleles can be considered rare in plant genomes [154,144,49,207,282]. Given the importance of secondary metabolites and their implications on plant fitness, it is not, however, surprising that structural and regulatory genes of plant metabolism often show signatures of selection. Indeed, the preponderance of such genes is comparable to other classically selected alleles, such as those of disease-resistance genes and life-history traits [138]. Below we provide a survey of sequence-based approaches to detect various forms of selection, along with a representative, non-exhaustive list of computer programs which can be used to infer phylogenies and perform tests of selection at the inter-specific and population level (Table 1).

6.1. Identifying selection on single loci

The **d_N/d_S ratio test** (also known as *K*a/*K*s, or ω) is one of the best known approaches to infer signature of selection in protein-coding genes from multiple-species alignments. The statistic is obtained as the ratio between the number of nonsynonymous substitutions per nonsynonymous site (d_N or *K*a) and the number of synonymous substitutions per synonymous site (d_S or *K*s). According to Kimura's theory, d_S will largely exceed d_N in the majority of protein-coding genes, because, statistically, nonsynonymous substitutions per nonsynonymous site will be mostly deleterious to protein function and will thus be eliminated by **purifying selection**; synonymous substitutions, on the other hand, will be instead mostly neutral, leading to a value of $d_N/d_S < 1$. This is frequently observed in large-scale alignments of several protein-encoding gene families at the macroevolutionary level. The opposite situation ($d_N/d_S > 1$) occurs instead when nonsynonymous substitutions per nonsynonymous site exceed synonymous substitutions, and this situation is indicative of the presence of repeated aminoacid changes which acted to favor novel protein structures and functions (**positive selection**). This type of selection, although extremely rare, especially when initial signatures of adaptation at the sequence level are then tested in ecological settings, is likely to act in coevolutionary processes, like during plant interactions with

Table 1
A representative, if inexhaustive, list of computer programs for phylogenetics and population genomics.

Name	Description	References
Phylogenetic inference and evolutionary processes		
ape	the core R package for evolution and phylogenetics, includes all main distance methods for phylogeny estimation (neighbor-joining, MP, ML, bayesian methods)	[200]
BEAST2	estimates phylogenies using bayesian methods and tests evolutionary hypotheses with molecular clock models	[30]
Datamonkey	web-server for analyzing evolutionary signatures in sequence data, includes a wide range of tests for detecting recombination and selection at the level of genes, single sites or phylogenetic branches	[270]
HyPhy	comparative sequence analyses focusing on likelihood-based approach for inference of selection. Its recent version includes a tool to calculate relative evolutionary rates from protein and nucleotide data (LEISR)	[205,238]
MEGA (version X)	a widely used software for phylogenetics with an intuitive graphical interface. Includes all main methods for tree construction, calculation of evolutionary distances and some tests of selection	[149]
iMKT	website for performing various MK-derived tests for selection	[186]
MrBayes	bayesian phylogenetic analysis using Markov chain Monte Carlo (MCMC) methods	[214]
NOTUNG 2.9	a software for reconciliation of gene tree/species tree including homology inference for identification of gene duplications	[47,61]
PAML	one of the most widely used package for phylogenetics. Includes different programs for the analysis of DNA and protein sequences using maximum likelihood (ML). Also available with a graphical user interface (PAMLX). Mainly used to test evolutionary hypotheses (e.g., d_N/d_S ratio test) after importing trees from other programs. Performs ancestral sequence reconstruction	[285,289,283]
phangorn	R package for phylogenetic reconstruction, tree comparison and test of different phylogenetic models using MP and ML. May be used in conjuction with other R packages for phylogenetics (ape)	[223]
PHYLIP (v3.697)	Joe Felsenstein's original software, perhaps the first program to be distributed for inferring phylogenies; includes parsimony and likelihood methods	[73]
PhyML v3.0	a widely used phylogeny software based on the maximum-likelihood principle, also available as an online platform	[93,94]
RAxML v8.0, RAxML-NG, ExaML v3.0	phylogenetic inference based on ML, derived from the algorithms present in PHYLIP. Runs on Linus/Unix, ideally on a cluster. Also several graphical interfaces available. Additional bootstrap metrics implemented in the recent RAxML-NG. The ExaML code allows inference for extremely large datasets	[241,145,146]
Seaview v4.7	graphical user interface for molecular phylogeny. multiple alignments, calculates trees using PHYLIP's parsimony or PhyML algorithm for ML	[88]
Population genomics (site-frequency spectrum, selective sweeps, measures of diversity, pool-sequencing)		
adegenet	R package for multivariate analyses of marker data, usually the entry-point to other specific packages for population genomics (e.g., pegas)	[123,124]
Arlequin	widely used software for population genomics, includes all major methods for population diversity and tests of neutrality at the population level (SFS, LD, Tajima's D, F_{st}, etc.)	[67]
CMS (Composite of Multiple Signals)	Sensitive approach for inference of selection based on the combination of different test statistics (haplotype frequency, linkage disequilibrium and population differentiation). Generally more robust with respect to the use of separate statistics and less dependent on demographic processes	[91,92]
DnaSP version 6.12.03	analysis of polymorphisms from single or multiple loci. calculates measures of DNA sequence variation within and between populations; several neutrality tests implemented: e.g., the HKA, Tajima's D and the McDonald and Kreitman test (MKT)	[217]
GenAlEx 6.5	Excel add-in for basic analyses of population genetics. Includes calculation of indices for population structure	[202]
Genepop	R package for general population genetic methods. Includes exact tests for independence, measures of population differentiation and disequilibrium among pairs of loci	[216]
hierfstat	R package for estimation of population structure using F-statistics (tests for population differentiation)	[85]
iSAFE	a coalescent-based method for identification of the specific mutations favored by selection in a selective sweep	[2]
pegas	from the developers of ape, another R package for the analysis of population genetic data. includes calculation of nucleotide diversity (π), SFS, LD scans and F_{ST} for population differentiation. With adegenet and ape constitutes a unique working environment for a wide range of phylogenetics and population genomics analyses	[199]
poolfstat	R package for the analysis of Pool-Seq data and estimation of F_{ST} (degree of differentiation between populations)	[106]
poolseq	R package for the analysis and simulation of Pool-Seq time series data	[248]
PoPoolation	Analysis of sequence data from pooled individual	[140]
PoPoolation2	Calculates significant differences in the allele frequencies between Pool-Seq datasets. Accepts data from multiple populations. Can be used for GWAS and E&R experiments	[139]
rehh	R package for detecting selection footprints based on haplotype homozigosity (e.g., "Extended Haplotype Homozygosity", EHH-based Tests)	[80,81]
S/HIC	A machine learning approach to infer position of hard and soft sweeps. Integrates nine different selection statistics and their degree of variation across the chromosomes. Robust under various demographic scenarios	[228–229]
SweeD	calculates theoretical SFS under a given demographic model and detects selective sweeps using a likelihood test, based on the SweepFinder algorithm	[201]
SweepFinder 2	performs genomic scans for detection of selective sweeps: this and SweeD are based on modelling changes with respect to the neutral SFS	[58]
SWIF(r)	Machine-learning inference of selective sweeps sites through refiniment of composite methods (e.g. CMS)	[244]
VariScan version 2.0.3	calculates all main population genetic parameters (e.g. number of segregating sites, nucleotide and haplotype diversity, LD-based statistics, neutrality tests)	[260]

pollinators and predators. Statistical evidence for positive selection is thus relatively common in secondary metabolism, and has been provided for glucosinolate [23,55,102,148], terpenoid [46] and pyrrolizidine alkaloid metabolism [127] and in the SABATH methyltransferase genes, a family involved in the synthesis of floral volatiles with functions in pollinator attraction and defense [13,14]. The results from these studies provide support to the patchwork model for the evolution of metabolic novelty: they show the recruitment, under a regime of positive selection, of enzymes originating from the duplication of genes already active in existing pathways.

Initially, approximate methods were used to calculate the rate of d_N and d_S, based simply on classification of the sites in sequence alignments and counting the number of nonsynonymous and synonymous substitutions; these methods are considered largely inadequate today as they did not take into account the transition/-transversion and the codon usage bias [161]. Later, starting from the work of Messier and Stewart, the method has been used to estimate d_N/d_S across each branch of a phylogenetic tree, identifying episodes of positive evolution along single lineages which were previously undetected simply by pairwise comparison of extant sequences. The innovative aspect brought about by the approach of Messier and Stewart was to reconstruct the sequences in all ancestral nodes of the phylogeny in order to test where, along phylogeny, episodes of selection occurred [174]. In their case, given the recent divergence of the species under study (primates), **maximum parsimony** (MP) was a reliable criteria to reconstruct the ancestral states; today the d_N/d_S ratio test is based instead on **maximum likelihood** (ML) approaches which calculate, for each internal node, all possible alternative character states and ancestral sequences, assigning a weight to all ancient alleles according to their probability of occurrence [286,290,287]. ML approaches allow to extend sequence reconstruction to ancestors in the distant past [99].

The **McDonald-Kreitman test (MKT)** represents an extension of the d_N/d_S ratio test taking into account the polymorphisms within species (e.g., between different individuals of the same species, or different accessions) and comparing it to the divergence between species for a protein-coding gene. Under neutrality, the null hypothesis is that the d_N/d_S within species (i.e., $\underline{p}_N/\underline{p}_S$ for the polymorphisms within species) equals the $\underline{d}_N/\underline{d}_S$ ratio (which measures the divergence between species). Positive evolution can be detected by a between-species ratio greater than the within-species ratio. Data in the MKT test take the form of a 2 × 2 table which be used to obtain statistical significance with a G-test for independence [169,63]. The McDonald-Kreitman test should be used only when comparing recent divergence (i.e., closely-related genes), so that all alleles share the same evolutionary history (no chance of recombination); the test is also unable to distinguish positive selection from other cases where slightly deleterious mutations might have been fixed due to population bottlenecks, a common outcome during speciation. For a detailed review of the potential and limits of the MKT test, see [118].

The methods summarised above may be used to detect signatures of adaptation in protein-coding genes, especially when comparing sequence differences between different species; however, a large part of adaptive phenotypic variation also exist at the microevolutionary level (within species) and may arise as a consequence of changes in non-coding parts of the genome, and, consequently, as part of variation in gene expression. In these cases, the approaches which may be used to detect selection shift from tests of neutrality on single loci to population genetics methods which detect, in general, regions of reduced or increased variability across whole genome sequences. These approaches thus identify regions in the genome that deviate from neutrality, on the basis of the assumption that high sequence conservation is suggestive of negative selection and may indicate the presence of a functional constraint [189,107,272]. We provide a brief overview of these population-genomics methods below. However population genetics methods cannot provide unambiguous identification of the relevant/causal loci under selection, and need to be integrated with functional assays of molecular function and validation of the candidate polymorphisms in natural settings to identify a true adaptive allele [178].

6.2. Identifying selection at the population level

6.2.1. Linkage disequilibrium (LD)

A significant deviation from the neutral model can be detected as an extension of the level of **linkage disequilibrium** (LD) along defined regions in the genome [236]. While it increases in prevalence in the population, the polymorphism in the selected allele is in strong association (disequilibrium) with other neighboring variants; this association, which extends over a physical region of variable length (haplotype) is maintained by the lack of recombination within the region. Various genome scans techniques can therefore measure the physical size of these haplotypes as a proxy for the extent of LD. An unusual large extension of LD in a particular region in the genome, with respect to other regions which are known, or suspected, to evolve neutrally, may indicate the presence of selection. Also in this case, there are several approaches which can be used to detect regions of extended LD in the genome. The extended haplotype homozygosity (EHH) has been initially applied in the form of a genome scan to the human genome [220] but derived approaches – always based on measuring extended LD (in combination with measures of population subdivision, see below) – have also recently been used in plants [27]. EHH is first based on the identification of “core haplotypes” across the genome (which correspond to putative selected loci), followed by the calculation of the LD decay along defined distances, propagating in both directions, from the core haplotype. Clearly, as one travels away from the selected allele, haplotype homozygosity decreases in the population, as recombination may increase polymorphisms in the region and reduce the size of the homozygous haplotype. Large haplotype homozygosity, when correlated to high values of haplotype frequency in the population, indicates the presence of directional selection and deviation from neutrality. LD tests are limited in their power to detect historical signatures of selection, however, as once the selected allele is fixed, the size of the haplotype can be rapidly reduced by recombination (the regions flanking the selected polymorphisms, once fixation is reached, are under relaxed selection); they represent powerful approaches, in any case, for detecting recent or ongoing selection events [170,234].

6.2.2. Population subdivision

The second family of approaches to detect genomic regions under selection isbased on the measure of population subdivision. Over time, in fact, under the influence of drift, demography and natural selection, populations of animals and plants may differentiate, both at the genotypic and phenotypic level, to form several distinct subpopulations. The magnitude of population differentiation can be measured by Wright’s group of F-statistics [280]). Of these, the fixation index, F_{ST}, is one of the most commonly used statistics to describe the partitioning of genetic variation within and among populations. There is a well-developed framework of population genetics theory behind the mathematical definition of F_{ST} and its estimators [108]; here, suffice to say that F_{ST}, in its simplified form, may be expressed as:

$$F_{ST} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}} \qquad (1)$$

where $\sigma^2_{between}$ is the total genetic variation in a defined locus between populations while σ^2_{within} is the total genetic variation in the same locus within the population [156]. Both measures of genetic variation are obtained from the variance of allele frequencies within and between populations. F_{ST} values vary from 0 to 1; a value of 0 indicates that allele frequencies are equally partitioned among populations; a value of 1, on the other hand, indicates the extreme case in which one of the alleles has a frequency of 1 in one of the subpopulations and is thus absent from the others. Such a condition, or, indeed, whenever F_{ST} approaches values close to 1, is indicative of a high level of genetic differentiation between populations. This condition may be reached by selective forces (in case of local adaptation), although several demographic processes, other than selection, may also influence the value of F_{ST} [108]. Especially in the case of genome-wide scans, which mitigate the effect of demographic processes, high values of F_{ST}, in comparison to those obtained from a neutral locus, are suggestive of directional selection acting in one of the subpopulations. On the other hand, low values of F_{ST} may indicate stabilizing selection or directional selection on all subpopulations [19]. One of the earliest methods to test population subdivision was developed by Lewontin and Krakauer [158]; the method is based on the comparison of empirical F_{ST} values against those obtained under a simulated model of neutral evolution. Over the years, the approach has been refined and applied to large collections of SNPs collected across entire genomes [3], and further improved taking into account demographic history through inclusion of the kinship matrix of the subpopulations [28]. The detection of F_{ST} outliers in genome-wide scans has been frequently applied in plants, starting from the use of a small number of microsatellites in sorghum (*S. bicolor* [41]) and sunflower (*H. annuus,* [128]) to more recent studies using SNPs or whole-genome sequencing in Arabidopsis and its relatives [115,268] and wheat [42]. A further expansion of the approaches based on population subdivision lies in the comparison between Q_{ST} and F_{ST} [156]. Q_{ST} is the counterpart of F_{ST}, but it is based, rather than on the measure of genetic variation at marker loci, on genetic variance in quantitative traits. If F_{ST} is measured on neutral marker loci, and Q_{ST} is based on quantitative traits with an additive genetic basis, then the comparison between F_{ST} and Q_{ST} assumes relevance with respect to the causes of trait divergence among subpopulations. If $F_{ST} \approx Q_{ST}$, for example, then the divergence measured from the trait among subpopulations is comparable to that of a neutral locus, and could have therefore been the result of genetic drift; **directional selection** could have instead acted in cases where $F_{ST} < Q_{ST}$ which indicates a trait divergence between-populations higher than that obtained from neutral loci: this case is suggestive of directional selection acting on the trait.

Q_{ST}-F_{ST} comparisons have been applied in plants, with a main focus on morphological, reproductive and stress tolerance traits [173,33,215,213]. Despite the potential of this approach in detecting selection at the level of quantitative traits, however, very few studies have made use of large-scale molecular phenotypic data (e.g., from transcript/protein profiling or metabolomics) to estimate Q_{ST}; this is perhaps due to the assumptions of the method, which need to be verified on a case-by-case basis, requiring F_{ST} to be calculated from strictly neutral loci and Q_{ST} to represent variance from purely additive phenotypic traits; also, Q_{ST}-F_{ST} comparisons are framed in a population genetic framework, therefore it is necessary to collect genotypic and phenotypic data from a large number of individuals. In a recent study, however, the divergence of metabolic traits (primary metabolites) and neutral marker loci was measured in a collection of three subpopulations representing the stages of durum wheat domestication. The approach allowed to detect signatures of directional selection directly in molecular phenotypic traits (metabolites), marking the transition from wild accessions to primary domesticates to the subsequent diversification as a result of recent selective breeding [22].

6.2.3. Site-frequency spectrum (SFS)

One of the consequences of **positive selection**, at the population level, is the increase of the frequency of the selected allele, which can rapidly reach fixation (100% frequency). This process is accompanied by a related decrease of sequence diversity, due to hitchhiking effects, in the regions around the beneficial selected allele. These regions (haplotypes) of marked reduction of genetic diversity leave a distinctive hallmark in the genome (**selective sweep**) and can be detected with a simple plot of the level of genetic diversity or, alternatively, looking at the relative proportion of the mutations in the population. Either plots can be obtained on specific genomic regions (comparing, for example, a locus suspected to be under selection with another locus known to evolve neutrally) or through whole genome scans with a sliding-window approach. Approaches based on the calculation of nucleotide diversity (π) were initially introduced by Tajima [246], but are still frequently used today in plant genome studies to identify sweep regions [29,38,113,162,295]. Tajima later extended his concept of nucleotide diversity into the formulation of a neutrality test, the Tajima's D, which can be used to assess significance of SFS. The test is based on the mathematical relation between two measures of genetic variation: 1) the number of polymorphic (or segregating) sites in *n* sequences (S); 2) the average number of nucleotide differences, for all possible pairwise combinations, in the same set of *n* sequences (k). Tajima demonstrated that the difference between these two measures represents the effect of selection [247]. In fact, when a selective sweep emerges, new mutations may accumulate in any case in the proximity of the selected locus [32]. These mutations are initially rare, and influence the parameter S (which is independent from allele frequencies), driving it to larger values with respect to parameter k (which instead depends from allele frequencies). Thus, large negative values of Tajima's D, which, in a simplified form, can be expressed as:

$$D = \frac{k-S}{std\ dev\ (k-S)} \tag{2}$$

are indicative of an excess of rare alleles and may suggest the presence of a recent selective sweep. Statistical significance of D can be obtained by comparing its value to the confidence limits of a beta distribution [247]. As with other neutrality tests, Tajima's D value can also be driven by several demographic processes (bottlenecks, migrations, population subdivisions, etc.) independently of natural selection, therefore caution must be taken in the interpretation of the results, especially when rejecting the null hypothesis of the population in mutation-drift equilibrium.

6.3. Machine learning approaches in population genomics

Machine learning is a set of methods to infer functional relationships existing within the input data without making any *a priori* assumptions. In their essence, machine learning approaches find a mathematical function, and elaborate a predictive model, between a set of multidimensional datasets (input) and the response variables of the system [9]. Supervised machine learning approaches "train" on empirical datasets (or datasets generated by numerical simulations) to predict the response of a specific output variable, whose response values are unknown [230]. These approaches were initially developed in computer science, but have been also applied to many areas of computational biology, and we refer the readers to excellent reviews and recent applications [60,82,89,110,296] to focus here on some recent developments of machine/deep learning in evolutionary and population genetics.

One of the major scientific challenges in population genetics is discriminating between marks of selection and demographic processes. This has proven to be extremely difficult: bottlenecks and selection, for example, leave very similar signatures in the genomes, and the two phenomena – to add even more complexity – often occur simultaneously in natural populations (e.g., when a population colonizes a new environment, it usually experiences a demographic bottleneck; at the same time the selection pressures present in the new environment may lead to adaptation). The main argument adopted so far to disentangle these contributions – basically that selection occurs in targeted regions, while demographic processes affect genomes globally – was found to be relatively inaccurate, given the large impact that natural selection has apparently had in shaping plant genomes [281,95,159,101,69].

Several approaches have been followed to address the confounding contribution of demography and selection: one of the most commonly used today is based on the combination of several statistics from various selection tests, with the objective of lessening the dependence of selection signatures from the demographic history of the populations. One of these approaches, called "composite of multiple signals" (CMS), integrates measures of LD (e.g., extended haplotype homozygosity, EHH) with population differentiation (F_{ST}) and shows a greater power in detecting selection, under different demographic scenarios, also in cases where the single test statistics failed to identify targets of selection [91,92].

More recently, machine and deep learning approaches have contributed to the development of other tools which were generally more robust to the confounding effects that demography has on selection (e.g., S/HIC, see below), or that allow simultenous estimation of demographic processes and selection [232]. In one of these approaches, S/HIC (Soft/Hard Inference through Classification, [228]), nine different statistics (based on nucleotide diversity, haplotype homozigosity, etc.) are calculated with a sliding-window approach to identify the location of soft and hard sweeps. Trained under proper datasets simulating various demograhic events, S/HIC showed high specificity and sensitivity in detecting both types of sweeps. Other tools recent tools further refine the limitations of composite approaches in detecting selection (for example, when one of the component statistic is undefined, see SWIFr [244]); while some other approaches allow to point specifically at the favored mutation within the selective sweep, without any prior knowledge of the underlying demography (iSAFE, [2]).

6.4. Whole genome sequencing of pools of individuals (Pool-Seq)

Research in population genetics requires collection of a large amount of data about the polymorphisms existing among individuals; although the costs associated with DNA sequencing are constantly decreasing, estimating allele frequencies, from single individuals at the population scale, is still rarely feasible. Pool-Seq approaches were developed to overcome these limitations. Essentially, Pool-Seq implies that DNAs extracted from multiple individuals are pooled before preparing the library for sequencing. The allele frequencies thus obtained can be compared between multiple populations, sampled, for example, from different environments or geopgraphical locations to infer signatures of adaptation. The approach has been shown to be particularly accurate in the estimation of allele frequencies, and also cost-effective with respect to sequencing of single individuals, provided a few requirements regarding the experimental settings are met (e.g., pool size, [224]). In the study of local adaptation, Pool-Seq has been used in *Arabidopsis lyrata* to identify the polymorphisms associated to the adaptation to serpentine soils [256], in teosinte, between lowland and highland populations [79] and also, on a smaller scale, to define clinal patterns and selection in a collection of *Solanum chinense* accessions (a wild tomato) from Chile and Peru [26]. In a recent study, Pool-Seq has been used to characterize the genomic differences driving ecotype differentiation of *Mimulus gattus* across coastal and inland-adapted populations [86]. This study added further support to the importance of selection on chromosomal inversions for determining speciation [155]. In approaches based on **Evolve & Resequence (E&R)**, Pool-Seq is used in combination with **experimental evolution** [129] to identify the differences in allele frequencies between the ancestral and the selected population: this allows to identify the causal allele(s) driving the phenotypic change in response to the selective agent set by the experimenter [225]. E&R studies have afforded exceptional insights about understanding genetic basis of adaptation, but few have focused on plants (given the difficulties inherent to their longer generation times, [133,21]); E&R studies have instead been popular within the Drosophila community [35,11,166]. Typically, allele frequencies between the base (ancestral) and selected populations are calculated across the whole genome with a sliding-window approach, and then tests of selection are perfomed on each window separately. We refer the readers to excellent surveys of the test statistics used in E&R approaches [225] and provide a list of software tools for the analysis of Pool-Seq data in Table 1.

7. Summary and outlook

In this review we have discussed the evolution of metabolism and the hypotheses made to explain the assembly of the biosynthetic pathways starting from non-enzymatic and primordial metabolism of the earliest living forms. The explosion of chemodiversity of plant secondary metabolites represented a further step where different evolutionary forces acted to expand the metabolic routes of primary metabolism reaching the pathway structures we observe today. Perhaps the greatest research challenge in the study of this metabolic expansion is to understand which part of chemical diversity emerged as a result of selection, i.e., as adaptations to natural environment, and which part, on the other hand, was instead derived from non-selective processes (demographic history of the populations, but also hitchhiking, pleiotropy, epistasis). Also, related to this, and since selection necessarily shaped the structure of metabolism as we observe it today, do selective pressures affected metabolic steps sequentially, as they emerged along evolutionary time, as in the Granick or retrograde hypothesis, or does selection acted globally on whole novel pathways which could have emerged after whole genome duplications?

These questions remain largely unanswered. With the relative ease with which we collect sequence data, and the various approaches for detecting selection, however, we have at hand today catalogues of molecular footprints of selection from a large number of plant species; this however should not lead us to infer molecular adaptation to be a pervasive phenomenon in plant genomes. We often assume, for example, that proteins are the result of a long history of functional optimization, but ancestral resurrection has largely showed that historical contingency and random chance have often driven the fixation of suboptimal forms [242]. Also, very few alleles showing molecular signatures of selection have been tested in an ecological background; and especially in the context of secondary metabolism, many true adaptive alleles may easily go undetected by current methods to detect selection. In fact, metabolic traits generally show medium to high heritabilities and have a genetic architecture mostly based on many loci of small effects [44,7]: under these conditions, it is possible that adaptation could have proceeded through subtle changes in the frequencies of several unlinked alleles [208], without leaving a clear signature at the sequence level. In plants, one approach to study polygenic adaptation would be to observe a concurrent shift in the frequencies of many unrelated alleles under some form of selective pressure. This has been done in Arabidopsis (thale cress, the model organism in plant biology), measuring the differential survival of a collection of geographical accessions under extreme drought, a trait resulting from an underlying metabolic phenotype [68]. The adaptive alleles in this case did not show reduction of haplotype diversity, and would have gone undetected by conventional approaches based on detection of hard sweeps. The adaptive nature of metabolic diversity is a complex issue, and its study brings conceptual and empirical challenges which can be faced only through the integration of approaches from multiple disciplines. At the enzymatic level, ancestral resurrection may clarify the epistatic relationships that different mutations may have on selection; progresses are also needed, however, on the computational side, to develop models which could better take into account the confounding effects that pleiotropy, demography and polygenic adaptation have on selection. Also, when possible, the large number of polymorphisms "putatively" under selection, detected from genome-wide scans, should be also tested in natural environments to provide measures of selection in an ecological setting. Combining all these approaches holds great promise for understanding the consequences of genetic variation on fitness and adaptation.

Acknowledgments

A.R.F. acknowledges the financial support of the EU Horizon 2020 Research and Innovation Programme, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620). A.R.F. is supported by the Max-Planck Society. The authors sincerely thank the anonymous reviewers for critically reading the manuscript and suggesting substantial improvements on the earlier draft.

References

[1] Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nat Genet 2005;37:73–6.

[2] Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, et al. Identifying the favored mutation in a positive selective sweep. Nat Methods 2018;15:279–82.

[3] Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 2002;12:1805–14.

[4] Alam MT, Olin-Sandoval V, Stincone A, Keller MA, Zelezniak A, Luisi BF, et al. The self-inhibitory nature of metabolic networks and its alleviation through compartmentalization. Nat Commun 2017;8:16018.

[5] Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, et al. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 2011;473:203–7.

[6] Alseekh S, Fernie AR. Metabolomics 20 years on: what have we learned and what hurdles remain?. Plant J 2018;94:933–42.

[7] Alseekh S, Tohge T, Wendenberg R, Scossa F, Omranian N, Li J, et al. Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. Plant Cell 2015;27:485–512.

[8] Altman S. Nobel lecture. Enzymatic cleavage of RNA by RNA. Biosci Rep 1990;10:317–37.

[9] Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol 2016;12:878.

[10] Bada JL, Lazcano A. Origin of life. Some like it hot, but not the first biomolecules. Science 2002;296:1982–3.

[11] Barghi N, Tobler R, Nolte V, Schlotterer C. Drosophila simulans: a species with improved resolution in evolve and resequence studies. G3 (Bethesda) 2017;7:2337–43.

[12] Barkman T, Zhang J. Evidence for escape from adaptive conflict?. Nature 2009;462:E1. discussion E2-3.

[13] Barkman TJ. Evidence for positive selection on the floral scent gene isoeugenol-O-methyltransferase. Mol Biol Evol 2003;20:168–72.

[14] Barkman TJ, Martins TR, Sutton E, Stout JT. Positive selection for single amino acid change promotes substrate discrimination of a plant volatile-producing enzyme. Mol Biol Evol 2007;24:1320–9.

[15] Baross JA, Hoffman SE. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. Origins Life Evol B 1985;15:327–45.

[16] Barrett RD, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 2011;12:767–80.

[17] Baute GJ, Kane NC, Grassa CJ, Lai Z, Rieseberg LH. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. New Phytol 2015;206:830–8.

[18] Beadle GW, Tatum EL. Genetic control of biochemical reactions in neurospora. Proc Natl Acad Sci U S A 1941;27:499–506.

[19] Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 2004;13:969–80.

[20] Becker S, Feldmann J, Wiedemann S, Okamura H, Schneider C, Iwan K, et al. Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. Science 2019;366:76–82.

[21] Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR, et al. A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. Genetics 2014;196:829–40.

[22] Beleggia R, Rau D, Laido G, Platani C, Nigro F, Fragasso M, et al. Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. Mol Biol Evol 2016;33:1740–53.

[23] Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J. Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci U S A 2006;103:9118–23.

[24] Bergthorsson U, Andersson DI, Roth JR. Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A 2007;104:17004–9.

[25] Bock R. Witnessing genome evolution: experimental reconstruction of endosymbiotic and horizontal gene transfer. Annu Rev Genet 2017;51:1–22.

[26] Bondel KB, Lainer H, Nosenko T, Mboup M, Tellier A, Stephan W. North-South colonization associated with local adaptation of the wild tomato species Solanum chilense. Mol Biol Evol 2015;32:2932–43.

[27] Bonhomme M, Boitard S, San Clemente H, Dumas B, Young N, Jacquet C. Genomic signature of selective sweeps illuminates adaptation of Medicago truncatula to root-associated microorganisms. Mol Biol Evol 2015;32:2097–110.

[28] Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics 2010;186:241–62.

[29] Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR, Hu TT, et al. Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. Proc Natl Acad Sci U S A 2007;104:12057–62.

[30] Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2019;15:e1006650.

[31] Braakman R. Mapping metabolism onto the prebiotic organic chemistry of hydrothermal vents. Proc Natl Acad Sci U S A 2013;110:13236–7.

[32] Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995;140:783–96.

[33] Brouillette LC, Mason CM, Shirk RY, Donovan LA. Adaptive differentiation of traits related to resource use in a desert annual along a resource gradient. New Phytol 2014;201:1316–27.

[34] Budin I, Bruckner RJ, Szostak JW. Formation of protocell-like vesicles in a thermal diffusion column. J Am Chem Soc 2009;131:9628–9.

[35] Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 2010;467:587–90.

[36] Caetano-Anolles G, Kim HS, Mittenthal JE. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci U S A 2007;104:9358–63.

[37] Caetano-Anolles G, Yafremava LS, Gee H, Caetano-Anolles D, Kim HS, Mittenthal JE. The origin and evolution of modern metabolism. Int J Biochem Cell Biol 2009;41:285–97.

[38] Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 2007;3:1745–56.

[39] Carretero-Paulet L, Fares MA. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol Biol Evol 2012;29:3541–51.

[40] Carrington Y, Guo J, Le CH, Fillo A, Kwon J, Tran LT, et al. Evolution of a secondary metabolic pathway from primary metabolism: shikimate and quinate biosynthesis in plants. Plant J 2018.

[41] Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, et al. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theor Appl Genet 2005;111:23–30.

[42] Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci U S A 2013;110:8057–62.

[43] Cech TR. Nobel lecture. Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena. Biosci Rep 1990;10:239–61.

[44] Chan EK, Rowe HC, Hansen BG, Kliebenstein DJ. The complex genetic architecture of the metabolome. PLoS Genet 2010;6:e1001198.

[45] Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, et al. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (Helianthus annuus). Plant Cell 2008;20:2931–45.

[46] Chen H, Li G, Kollner TG, Jia Q, Gershenzon J, Chen F. Positive Darwinian selection is a driving force for the diversification of terpenoid biosynthesis in the genus Oryza. BMC Plant Biol 2014;14:239.

[47] Chen K, Durand D, Farach-Colton M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 2000;7:429–47.

[48] Christ B, Xu C, Xu M, Li FS, Wada N, Mitchell AJ, et al. Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. Nat Commun 2019;10:3206.

[49] Coberly LC, Rausher MD. Pleiotropic effects of an allele producing white flowers in Ipomoea purpurea. Evolution 2008;62:1076–85.

[50] Cody GD, Boctor NZ, Hazen RM, Brandes JA, Morowitz HJ, Yoder HS. Geochemical roots of autotrophic carbon fixation: hydrothermal experiments in the system citric acid, H2O-(+/- FeS)-(+/- NiS). Geochim Cosmochim Acta 2001;65:3557–76.

[51] Cole CT, Ingvarsson PK. Pathway position constrains the evolution of an ecologically important pathway in aspens (Populus tremula L.). Mol Ecol 2018;27:3317–30.

[52] Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 2008;9:938–50.

[53] Copley RR, Bork P. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. J Mol Biol 2000;303:627–41.

[54] Cronin JR, Moore CB. Amino acid analyses of the Murchison, Murray, and Allende carbonaceous chondrites. Science 1971;172:1327–9.

[55] de Kraker JW, Gershenzon J. From amino acid to glucosinolate biosynthesis: protein sequence changes in the evolution of methylthioalkylmalate synthase in Arabidopsis. Plant Cell 2011;23:38–53.

[56] de Rosa R, Labedan B. The evolutionary relationships between the two bacteria Escherichia coli and Haemophilus influenzae and their putative last common ancestor. Mol Biol Evol 1998;15:17–27.
[57] de Vladar HP, Santos M, Szathmary E. Grand views of evolution. Trends Ecol Evol 2017;32:324–34.
[58] DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics 2016;32:1895–7.
[59] Des Marais DL, Rausher MD. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 2008;454:762–5.
[60] Dobos O, Horvath P, Nagy F, Danka T, Viczian A. A deep learning-based approach for high-throughput hypocotyl phenotyping. Plant Physiol 2019;181:1415–24.
[61] Durand D, Halldorsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol 2006;13:320–35.
[62] Dutartre L, Hilliou F, Feyereisen R. Phylogenomics of the benzoxazinoid biosynthetic pathway of Poaceae: gene duplications and origin of the Bx cluster. BMC Evol Biol 2012;12:64.
[63] Egea R, Casillas S, Barbadilla A. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res 2008;36:W157–62.
[64] Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. Robustness of reconstructed ancestral protein functions to statistical uncertainty. Mol Biol Evol 2017;34:247–61.
[65] Emiliani G, Fondi M, Fani R, Gribaldo S. A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. Biology Direct 2009;4:7.
[66] Enright AJ, Ouzounis CA. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2001;2:34.
[67] Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 2010;10:564–7.
[68] Exposito-Alonso M, Vasseur F, Ding W, Wang G, Burbano HA, Weigel D. Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. Nat Ecol Evol 2018;2:352–8.
[69] Exposito-Alonso M, Genomes Field Experiment T, Burbano HA, Bossdorf O, Nielsen R, Weigel D. Natural selection on the Arabidopsis thaliana genome in present and future climates. Nature 2019;573:126–9.
[70] Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev 2009;6:23–52.
[71] Fani R, Brilli M, Fondi M, Lio P. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol Biol 2007;7 (Suppl 2):S4.
[72] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76.
[73] Felsenstein J. PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–6.
[74] Fernie AR, Tohge T. The genetics of plant metabolism. Annu Rev Genet 2017;51:287–310.
[75] Ferris JP, Hill Jr AR, Liu R, Orgel LE. Synthesis of long prebiotic oligomers on mineral surfaces. Nature 1996;381:59–61.
[76] Flint-Garcia SA, Thornsberry JM, Buckler EST. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 2003;54:357–74.
[77] Fondi M, Emiliani G, Fani R. Origin and evolution of operons and metabolic pathways. Res Microbiol 2009;160:502–12.
[78] Furukawa Y, Chikaraishi Y, Ohkouchi N, Ogawa NO, Glavin DP, Dworkin JP, et al. Extraterrestrial ribose and other sugars in primitive meteorites. Proc Natl Acad Sci U S A 2019;116:24440–5.
[79] Fustier MA, Brandenburg JT, Boitard S, Lapeyronnie J, Eguiarte LE, Vigouroux Y, et al. Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. Mol Ecol 2017;26:2738–56.
[80] Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 2012;28:1176–7.
[81] Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour 2017;17:78–90.
[82] Gazestani VH, Lewis NE. From genotype to phenotype: augmenting deep learning with networks and systems biology. Curr Opin Syst Biol 2019;15:68–73.
[83] Gilbert W. Origin of life: the RNA world. Nature 1986;319:618.
[84] Goldford JE, Hartman H, Marsland 3rd R, Segre D. Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. Nat Ecol Evol 2019;3:1715–24.
[85] Goudet J. hierfstat, a package for r to compute and test hierarchical F-statistics. Mol Ecol Notes 2005;5:184–6.
[86] Gould BA, Chen Y, Lowry DB. Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. Mol Ecol 2017;26:163–77.
[87] Gould SJ, Urba ES. Exaptation - a missing term in the science of form. Paleobiology 1982;8:4–15.
[88] Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;27:221–4.
[89] Grabowski P, Rappsilber J. A primer on data analytics in functional genomics: how to move from data to insight?. Trends Biochem Sci 2019;44:21–32.
[90] Granick S. Speculations on the origins and evolution of photosynthesis. Ann N Y Acad Sci 1957;69:292–308.
[91] Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 2010;327:883–6.
[92] Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying recent adaptations in large-scale genomic data. Cell 2013;152:703–13.
[93] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.
[94] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–21.
[95] Hahn MW. Toward a selection theory of molecular evolution. Evolution 2008;62:255–65.
[96] Hahn MW. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 2009;100:605–17.
[97] Hanada K, Sawada Y, Kuromori T, Klausnitzer R, Saito K, Toyoda T, et al. Functional compensation of primary and secondary metabolites by duplicate genes in Arabidopsis thaliana. Mol Biol Evol 2011;28:377–82.
[98] Hanczyc MM, Mansy SS, Szostak JW. Mineral surface directed membrane assembly. Orig Life Evol Biosph 2007;37:67–82.
[99] Hanson-Smith V, Kolaczkowski B, Thornton JW. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol Biol Evol 2010;27:1988–99.
[100] Harms MJ, Thornton JW. Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol 2010;20:360–6.
[101] Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet 2013;45:891–8.
[102] Heidel-Fischer HM, Kirsch R, Reichelt M, Ahn SJ, Wielsch N, Baxter SW, et al. An insect counteradaptation against host plant defenses evolved through concerted neofunctionalization. Mol Biol Evol 2019;36:930–41.
[103] Henry CS, Lerma-Ortiz C, Gerdes SY, Mullen JD, Colasanti R, Zhukov A, et al. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. BMC Genomics 2016;17:473.
[104] Hileman LC, Baum DA. Why do paralogs persist? Molecular evolution of CYCLOIDEA and related floral symmetry genes in Antirrhineae (Veronicaceae). Mol Biol Evol 2003;20:591–600.
[105] Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 2007;449:677–81.
[106] Hivert V, Leblois R, Petit EJ, Gautier M, Vitalis R. Measuring genetic differentiation from pool-seq data. Genetics 2018;210:315–30.
[107] Hohenlohe PA, Phillips PC, Cresko WA. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. Int J Plant Sci 2010;171:1059–71.
[108] Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 2009;10:639–50.
[109] Horowitz NH. On the evolution of biochemical syntheses. Proc Natl Acad Sci U S A 1945;31:153–7.
[110] Hoyal Cuthill JF, Guttenberg N, Ledger S, Crowther R, Huertas B. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Sci Adv 2019;5:eaaw4967.
[111] Huang R, O'Donnell AJ, Barboline JJ, Barkman TJ. Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. Proc Natl Acad Sci U S A 2016;113:10613–8.
[112] Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, et al. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. Proc Natl Acad Sci U S A 2012;109:2966–71.
[113] Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature 2012;490:497–501.
[114] Huber C, Wachtershauser G. Activated acetic acid by carbon fixation on (Fe, Ni)S under primordial conditions. Science 1997;276:245–7.
[115] Huber CD, Nordborg M, Hermisson J, Hellmann I. Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. Mol Biol Evol 2014;31:3026–39.
[116] Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 1997;28:437–66.
[117] Hughes AL. Adaptive evolution after gene duplication. Trends Genet 2002;18:433–4.
[118] Hughes AL. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity (Edinb) 2007;99:364–73.
[119] Jaillon O, Aury JM, Wincker P. "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. C R Biol 2009;332:241–53.
[120] Jensen RA. Enzyme recruitment in evolution of new function. Annu Rev Microbiol 1976;30:409–25.
[121] Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 2001;292:1319–25.
[122] Joly-Lopez Z, Flowers JM, Purugganan MD. Developing maps of fitness consequences for plant genomes. Curr Opin Plant Biol 2016;30:101–7.

[123] Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 2008;24:1403–5.
[124] Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 2011;27:3070–1.
[125] Joyce GF. The antiquity of RNA-based evolution. Nature 2002;418:214–21.
[126] Joyce GF, Szostak JW. Protocells and RNA self-replication. Cold Spring Harb Perspect Biol 2018;10.
[127] Kaltenegger E, Eich E, Ober D. Evolution of homospermidine synthase in the convolvulaceae: a story of gene duplication, gene loss, and periods of various selection pressures. Plant Cell 2013;25:1213–27.
[128] Kane NC, Rieseberg LH. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, Helianthus annuus. Genetics 2007;175:1823–34.
[129] Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC. Experimental evolution. Trends Ecol Evol 2012;27:547–60.
[130] Keller MA, Turchyn AV, Ralser M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. Mol Syst Biol 2014;10:725.
[131] Keller MA, Piedrafita G, Ralser M. The widespread role of non-enzymatic reactions in cellular metabolism. Curr Opin Biotechnol 2015;34:153–61.
[132] Keller MA, Kampjut D, Harrison SA, Ralser M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. Nat Ecol Evol 2017;1:83.
[133] Kelly JK, Koseva B, Mojica JP. The genomic signal of partial sweeps in Mimulus guttatus. Genome Biol Evol 2013;5:1457–69.
[134] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[135] Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
[136] Kimura M. The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet 1991;66:367–86.
[137] Kliebenstein DJ. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. PLoS ONE 2008;3:e1838.
[138] Kliebenstein DJ, Osbourn A. Making new molecules – evolution of pathways for novel metabolites in plants. Curr Opin Plant Biol 2012;15:415–23.
[139] Kofler R, Pandey RV, Schlotterer C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 2011;27:3435–6.
[140] Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS ONE 2011;6: e15925.
[141] Koonin EV. An apology for orthologs - or brave new memes. Genome Biol 2001;2:1005.
[142] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 2005;39:309–38.
[143] Kopp J, Kopriva S, Suss KH, Schulz GE. Structure and mechanism of the amphibolic enzyme D-ribulose-5-phosphate 3-epimerase from potato chloroplasts. J Mol Biol 1999;287:761–71.
[144] Korves TM, Schmid KJ, Caicedo AL, Mays C, Stinchcombe JR, Purugganan MD, et al. Fitness effects associated with the major flowering time gene FRIGIDA in Arabidopsis thaliana in the field. Am Nat 2007;169:E141–57.
[145] Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 2015;31:2577–9.
[146] Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 2019.
[147] Kroymann J. Natural diversity and adaptation in plant secondary metabolism. Curr Opin Plant Biol 2011;14:246–51.
[148] Kumar R, Lee SG, Augustine R, Reichelt M, Vassao DG, Palavalli MH, et al. Molecular basis of the evolution of methylthioalkylmalate synthase and the diversity of methionine-derived glucosinolates. Plant Cell 2019;31: 1633–47.
[149] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.
[150] Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, et al. Relaxed selection in the wild. Trends Ecol Evol 2009;24:487–96.
[151] Lazcano A. Historical development of origins research. Cold Spring Harb Perspect Biol 2010;2:a002089.
[152] Lazcano A. Alexandr I. Oparin and the origin of life: a historical reassessment of the heterotrophic theory. J Mol Evol 2016;83:214–22.
[153] Lazcano A, Miller SL. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. Cell 1996;85:793–8.
[154] Le Corre V, Roux F, Reboud X. DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol 2002;19:1261–71.
[155] Lee CR, Wang B, Mojica JP, Mandakova T, Prasad K, Goicoechea JL, et al. Young inversion with multiple linked QTLs under selection in a hybrid zone. Nat Ecol Evol 2017;1:119.
[156] Leinonen T, McCairns RJ, O'Hara RB, Merila J. Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet 2013;14:179–90.
[157] Leong BJ, Last RL. Promiscuity, impersonation and accommodation: evolution of plant specialized metabolism. Curr Opin Struct Biol 2017;47:105–12.
[158] Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 1973;74:175–95.
[159] Li J, Li H, Jakobsson M, Li S, Sjodin P, Lascoux M. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol 2012;21:28–44.
[160] Li J, Schuman MC, Halitschke R, Li X, Guo H, Grabe V, Hammer A, Baldwin IT. The decoration of specialized metabolites influences stylar development. Elife 2018:7.
[161] Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 1993;36:96–9.
[162] Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, et al. Genomic analyses provide insights into the history of tomato breeding. Nat Genet 2014;46:1220–6.
[163] Liu Z, Tavares R, Forsythe ES, Andre F, Lugan R, Jonasson G, et al. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. Nat Commun 2016;7:13026.
[164] Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science 2000;290:1151–5.
[165] Lynch M, Conery JS. The evolutionary demography of duplicate genes. J Struct Funct Genomics 2003;3:35–44.
[166] Mallard F, Nolte V, Tobler R, Kapun M, Schlotterer C. A simple genetic basis of adaptation to a novel thermal environment results in complex metabolic rewiring in Drosophila. Genome Biol 2018;19:119.
[167] Martin W, Herrmann RG. Gene transfer from organelles to the nucleus: how much, what happens, and Why?. Plant Physiol 1998;118:9–17.
[168] Martin W, Baross J, Kelley D, Russell MJ. Hydrothermal vents and the origin of life. Nat Rev Microbiol 2008;6:805–14.
[169] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991;351:652–4.
[170] McVean G. The structure of linkage disequilibrium around a selective sweep. Genetics 2007;175:1395–406.
[171] Mei W, Stetter MG, Gates DJ, Stitzer MC, Ross-Ibarra J. Adaptation in plant genomes: Bigger is different. Am J Bot 2018;105:16–9.
[172] Meierhenrich UJ, Munoz Caro GM, Bredehoft JH, Jessberger EK, Thiemann WH. Identification of diamino acids in the Murchison meteorite. Proc Natl Acad Sci U S A 2004;101:9182–6.
[173] Merilä J, Crnokrak P. Comparison of genetic differentiation at marker loci and quantitative traits. J Evol Biol 2001;14:892–903.
[174] Messier W, Stewart CB. Episodic adaptive evolution of primate lysozymes. Nature 1997;385:151–4.
[175] Michael JP. Quinoline, quinazoline and acridone alkaloids. Nat Prod Rep 2008;25:166–87.
[176] Miller SL. A production of amino acids under possible primitive earth conditions. Science 1953;117:528.
[177] Mintz-Oron S, Meir S, Malitsky S, Ruppin E, Aharoni A, Shlomi T. Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proc Natl Acad Sci U S A 2012;109:339–44.
[178] Mitchell-Olds T, Willis JH, Goldstein DB. Which evolutionary processes influence natural genetic variation for phenotypic traits?. Nat Rev Genet 2007;8:845–56.
[179] Moghe GD, Last RL. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol 2015;169:1512–23.
[180] Moghe GD, Leong BJ, Hurney SM, Daniel Jones A, Last RL. Evolutionary routes to biochemical innovation revealed by integrative analysis of a plant-defense related specialized metabolic pathway. Elife 2017:6.
[181] Moore BD, Andrew RL, Kulheim C, Foley WJ. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. New Phytol 2014;201:733–50.
[182] Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, et al. Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A 2019;116:2344–53.
[183] Moore PB, Steitz TA. The involvement of RNA in ribosome function. Nature 2002;418:229–35.
[184] Morowitz HJ. A theory of biochemical organization, metabolic pathways, and evolution. Complexity 1999;4:39–53.
[185] Mukherjee D, Mukherjee A, Ghosh TC. Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in Arabidopsis thaliana. Genome Biol Evol 2015;8:17–28.
[186] Murga-Moreno J, Coronado-Zamora M, Hervas S, Casillas S, Barbadilla A. iMKT: the integrative McDonald and Kreitman test. Nucleic Acids Res 2019;47:W283–8.
[187] Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 1996;93:10268–73.
[188] Nara T, Hshimoto T, Aoki T. Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. Gene 2000;257:209–22.
[189] Nielsen R. Molecular signatures of natural selection. Annu Rev Genet 2005;39:197–218.
[190] Novikov Y, Copley SD. Reactivity landscape of pyruvate under simulated hydrothermal vent conditions. Proc Natl Acad Sci U S A 2013;110:13283–8.
[191] Nutzmann HW, Scazzocchio C, Osbourn A. Metabolic gene clusters in eukaryotes. Annu Rev Genet 2018;52:159–83.
[192] Obata D, Takabayashi A, Tanaka R, Tanaka A, Ito H. Horizontal transfer of promiscuous activity from nonphotosynthetic bacteria contributed to evolution of chlorophyll degradation pathway. Mol Biol Evol 2019;36:2830–41.

[193] Obornik M, Green BR. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. Mol Biol Evol 2005;22:2343–53.
[194] Ohno S. Evolution by gene duplication. Berlin, Heidelberg: Springer; 1970.
[195] Olson-Manning CF, Lee CR, Rausher MD, Mitchell-Olds T. Evolution of flux control in the glucosinolate pathway in Arabidopsis thaliana. Mol Biol Evol 2013;30:14–23.
[196] Orgel LE. Prebiotic chemistry and the origin of the RNA world. Crit Rev Biochem Mol Biol 2004;39:99–123.
[197] Otto SP. Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc Biol Sci 2004;271:705–14.
[198] Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. Plant Physiol 2016;171:2294–316.
[199] Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 2010;26:419–20.
[200] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004;20:289–90.
[201] Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol 2013;30:2224–34.
[202] Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics 2012;28:2537–9.
[203] Peng M, Shahzad R, Gul A, Subthain H, Shen S, Lei L, et al. Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. Nat Commun 2017;8:1975.
[204] Petersen J, Teich R, Becker B, Cerff R, Brinkmann H. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol Biol Evol 2006;23:1109–18.
[205] Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics 2005;21:676–9.
[206] Powner MW, Gerland B, Sutherland JD. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature 2009;459:239–42.
[207] Prasad KV, Song BH, Olson-Manning C, Anderson JT, Lee CR, Schranz ME, et al. A gain-of-function polymorphism controlling complex traits and fitness in nature. Science 2012;337:1081–4.
[208] Pritchard JK, Di Rienzo A. Adaptation – not by sweeps alone. Nat Rev Genet 2010;11:665.
[209] Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol 2019;20:38.
[210] Quadrana L, Almeida J, Asis R, Duffy T, Dominguez PG, Bermudez L, et al. Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat Commun 2014;5:3027.
[211] Ralser M. The RNA world and the origin of metabolic enzymes. Biochem Soc Trans 2014;42:985–8.
[212] Richards TA, Soanes DM, Foster PG, Leonard G, Thornton CR, Talbot NJ. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell 2009;21:1897–911.
[213] Rifkin JL, Liao IT, Castillo AS, Rausher MD. Multiple aspects of the selfing syndrome of the morning glory Ipomoea lacunosa evolved in response to selection: A Qst-Fst comparison. Ecol Evol 2019;9:7712–25.
[214] Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 2012;61:539–42.
[215] Roschanski AM, Csillery K, Liepelt S, Oddou-Muratorio S, Ziegenhagen B, Huard F, et al. Evidence of divergent selection for drought and cold tolerance at landscape and local scales in Abies alba Mill. in the French Mediterranean Alps. Mol Ecol 2016;25:776–94.
[216] Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 2008;8:103–6.
[217] Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 2017;34:3299–302.
[218] Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, et al. Revisiting ancestral polyploidy in plants. Sci Adv 2017;3:e1603195.
[219] Ruprecht C, Mendrinna A, Tohge T, Sampathkumar A, Klie S, Fernie AR, et al. FamNet: a framework to identify multiplied modules driving pathway expansion in plants. Plant Physiol 2016;170:1878–94.
[220] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature 2002;419:832–7.
[221] Saladino R, Botta G, Pino S, Costanzo G, Di Mauro E. Genetics first or metabolism first? The formamide clue. Chem Soc Rev 2012;41:5526–65.
[222] Scarcelli N, Cheverud JM, Schaal BA, Kover PX. Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci U S A 2007;104:16986–91.
[223] Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics 2011;27:592–3.
[224] Schlotterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. Nat Rev Genet 2014;15:749–63.
[225] Schlotterer C, Kofler R, Versace E, Tobler R, Franssen SU. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity (Edinb) 2015;114:431–40.
[226] Schmidt S, Sunyaev S, Bork P, Dandekar T. Metabolites: a helping hand for pathway evolution?. Trends Biochem Sci 2003;28:336–41.
[227] Schmitt-Kopplin P, Gabelica Z, Gougeon RD, Fekete A, Kanawati B, Harir M, et al. High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. Proc Natl Acad Sci U S A 2010;107:2763–8.
[228] Schrider DR, Kern AD. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet 2016;12:e1005928.
[229] Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol 2017;34:1863–77.
[230] Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends Genet 2018;34:301–12.
[231] Segre D, Lancet D, Kedem O, Pilpel Y. Graded Autocatalysis Replication Domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. Orig Life Evol Biosph 1998;28:501–14.
[232] Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol 2016;12:e1004845.
[233] Sikosek T, Chan HS, Bornberg-Bauer E. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. Proc Natl Acad Sci U S A 2012;109:14888–93.
[234] Siol M, Wright SI, Barrett SC. The population genomics of plant adaptation. New Phytol 2010;188:313–32.
[235] Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 2013;45:831–5.
[236] Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res 1974;23:23–35.
[237] Sobota JM, Imlay JA. Iron enzyme ribulose-5-phosphate 3-epimerase in Escherichia coli is rapidly damaged by hydrogen peroxide but can be protected by manganese. Proc Natl Acad Sci U S A 2011;108:5402–7.
[238] Spielman SJ, Kosakovsky Pond SL. Relative evolutionary rate inference in HyPhy with LEISR. PeerJ 2018;6:e4339.
[239] Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA. The ribonuclease from an extinct bovid ruminant. FEBS Lett 1990;262:104–6.
[240] Stairs S, Nikmal A, Bucar DK, Zheng SL, Szostak JW, Powner MW. Divergent prebiotic synthesis of pyrimidine and 8-oxo-purine ribonucleotides. Nat Commun 2017;8:15270.
[241] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.
[242] Starr TN, Picton LK, Thornton JW. Alternative evolutionary histories in the sequence space of an ancient protein. Nature 2017;549:409–13.
[243] Strobel SA. Repopulating the RNA world. Nature 2001;411:1003–6.
[244] Sugden LA, Atkinson EG, Fischer AP, Rong S, Henn BM, Ramachandran S. Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun 2018;9:703.
[245] Sweetlove LJ, Fernie AR. The spatial organization of metabolism within the plant cell. Annu Rev Plant Biol 2013;64:723–46.
[246] Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics 1983;105:437–60.
[247] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–95.
[248] Taus T, Futschik A, Schlotterer C. Quantifying selection with pool-Seq time series data. Mol Biol Evol 2017;34:3023–34.
[249] Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol 2001;19:482–6.
[250] Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps?. Genome Res 2006;16:702–12.
[251] Thornton JW, Need E, Crews D. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 2003;301:1714–7.
[252] Thornton KR, Jensen JD. Controlling the false-positive rate in multilocus genome scans for selection. Genetics 2007;175:737–50.
[253] Tohge T, Fernie AR. Co-expression and co-responses: within and beyond transcription. Front Plant Sci 2012;3:248.
[254] Tohge T, Fernie AR. Specialized metabolites of the flavonol class mediate root phototropism and growth. Mol Plant 2016;9:1554–5.
[255] Tohge T, Watanabe M, Hoefgen R, Fernie AR. The evolution of phenylpropanoid metabolism in the green lineage. Crit Rev Biochem Mol Biol 2013;48:123–52.
[256] Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. Nat Genet 2010;42:260–3.
[257] Unkles SE, Logsdon Jr JM, Robison K, Kinghorn JR, Duncan JM. The tigA gene is a transcriptional fusion of glycolytic genes encoding triose-phosphate isomerase and glyceraldehyde-3-phosphate dehydrogenase in oomycota. J Bacteriol 1997;179:6816–23.
[258] Vasas V, Szathmary E, Santos M. Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. Proc Natl Acad Sci U S A 2010;107:1470–5.
[259] Vasas V, Fernando C, Szilagyi A, Zachar I, Santos M, Szathmary E. Primordial evolvability: Impasses and challenges. J Theor Biol 2015;381:29–38.
[260] Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 2005;21:2791–3.
[261] Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet 2013;47:97–120.

[262] Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, et al. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol 2012;10:e1001446.

[263] Wachtershauser G. Before enzymes and templates – theory of surface metabolism. Microbiol Rev 1988;52:452–84.

[264] Wachtershauser G. Evolution of the first metabolic cycles. Proc Natl Acad Sci U S A 1990;87:200–4.

[265] Wachtershauser G. Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog Biophys Mol Biol 1992;58:85–201.

[266] Wagner A. Selection and gene duplication: a view from the genome. Genome Biol 2002;3:1012.

[267] Wagner A. Arrival of the fittest: solving evolution's greatest puzzle. Penguin Publishing Group; 2014.

[268] Wang B, Mojica JP, Perera N, Lee CR, Lovell JT, Sharma A, et al. Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in Boechera stricta. Genome Biol 2019;20:126.

[269] Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953;171:737–8.

[270] Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Biol Evol Mol 2018.

[271] Webber C, Ponting CP. Genes and homology. Curr Biol 2004;14:R332–3.

[272] Weigel D, Nordborg M. Population genomics for understanding adaptation in wild plant species. Annu Rev Genet 2015;49:315–38.

[273] Weng JK, Philippe RN, Noel JP. The rise of chemodiversity in plants. Science 2012;336:1667–70.

[274] Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet 2007;3:1687–701.

[275] Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 2015;32:820–32.

[276] Wink M. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry 2003;64:3–19.

[277] Winzer T, Kern M, King AJ, Larson TR, Teodor RI, Donninger SL, et al. Plant science. Morphinan biosynthesis in opium poppy requires a P450-oxidoreductase fusion protein. Science 2015;349:309–12.

[278] Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell 2017;29:944–59.

[279] Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 2001;11:356–72.

[280] Wright S. Evolution in Mendelian Populations. Genetics 1931;16:97–159.

[281] Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects of artificial selection on the maize genome. Science 2005;308:1310–4.

[282] Wu Y, Guo T, Mu Q, Wang J, Li X, Wu Y, et al. Allelochemicals targeted to balance competing selections in African agroecosystems. Nat Plants 2019;5:1229–36.

[283] Xu B, Yang Z. PAMLX: a graphical user interface for PAML. Mol Biol Evol 2013;30:2723–4.

[284] Xu G, Cao J, Wang X, Chen Q, Jin W, Li Z, et al. Evolutionary metabolomics identifies substantial metabolic divergence between maize and its wild ancestor, teosinte. Plant Cell 2019;31:1990–2009.

[285] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997;13:555–6.

[286] Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998;15:568–73.

[287] Yang Z. Inference of selection from multiple species alignments. Curr Opin Genet Dev 2002;12:688–94.

[288] Yang Z. Computational molecular evolution. Oxford University Press; 2006.

[289] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.

[290] Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000;15:496–503.

[291] Ycas M. A note on the origin of life. P Natl Acad Sci USA 1955;41:714–6.

[292] Zeng K, Charlesworth B. The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics 2010;186:1411–24.

[293] Zhang J, Nei M. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol 1997;44(Suppl 1): S139–46.

[294] Zhou F, Pichersky E. The complete functional characterization of the terpene synthase family in tomato. New Phytol 2020.

[295] Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, et al. Rewiring of the Fruit Metabolome in Tomato Breeding. Cell 2018;172(249–261):e212.

[296] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51:12–8.