The Integration of Metabolomics and Next-Generation Sequencing Data to Elucidate the Pathways of Natural Product Metabolism in Medicinal Plants

Authors
Federico Scossa\textsuperscript{1,2}, Maria Benina\textsuperscript{3}, Saleh Alseekh\textsuperscript{1}, Youjun Zhang\textsuperscript{1,3}, Alisdair R. Fernie\textsuperscript{1,3}

Affiliations
1 Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany
2 Consiglio per la Ricerca in Agricoltura e l’Analisi dell’Economia Agraria, Rome, Italy
3 Center of Plant Systems Biology and Biotechnology, Plovdiv, Bulgaria

Key words
Papaver somniferum, Papaveraceae, Catharanthus roseus, Apocynaceae, alkaloids, caffeine, cannabinoids, ginseng, artemisinin, taxol

ABSTRACT
Plants have always been used as medicines since ancient times to treat diseases. The knowledge around the active components of herbal preparations has remained nevertheless fragmentary: the biosynthetic pathways of many secondary metabolites of pharmacological importance have been clarified only in a few species, while the chemodiversity present in many medicinal plants has remained largely unexplored. Despite the advancements of synthetic biology for production of medicinal compounds in heterologous hosts, the native plant species are often the most reliable and economic source for their production. It thus becomes fundamental to investigate the metabolic composition of medicinal plants to characterize their natural metabolic diversity and to define the biosynthetic routes \textit{in planta} of important compounds to develop strategies to further increase their content. We present here a number of case studies for selected classes of secondary metabolites and we review their health benefits and the historical developments in their structural elucidation and characterization of biosynthetic genes. We cover the cases of benzisoquinoline and monoterpenoid indole alkaloids, cannabinoids, caffeine, ginsenosides, withanolides, artemisinin, and taxol; we show how the “early” biochemical or the more recent integrative approaches--based on omics-analyses--have helped to elucidate their metabolic pathways and cellular compartmentation. We also summarize how the knowledge generated about their biosynthesis has been used to develop metabolic engineering strategies in heterologous and native hosts. We conclude that following the advent of novel, high-throughput and cost-effective analytical technologies, the secondary metabolism of medicinal plants can now be examined under the lens of systems biology.

Introduction
The first archeological evidences of the use of herbal remedies date back to prehistory: Neanderthals, for example, who have been long considered mainly meat-eaters, had instead already a good knowledge of the surrounding vegetation and adopted sophisticated diets: their dental plaques contained residues of several herbs, indicating the early consumption of plants, perhaps already for self-medication purposes [1].

Written records of the use of medicinal plants--including recipes for preparing decocts and extracts--were also common in Ancient Egypt, Greece, Rome, China, and in the Middle East.
During the modern era, a more rational approach began to be applied to the study of herbal medicines, as was the case with the discovery of the properties of foxglove (Digitalis purpurea L., family Plantaginaceae) to treat edema and heart failures [2,3].

Since the beginning of the 19th century, in parallel with the development of the pharmaceutical industry, there was an impetus in the isolation of new compounds possessing a therapeutic or commercial potential. In 1805, morphine was isolated from the latex of opium poppies and went immediately into commercial production in Europe and the United States, where it soon reached widespread popularity as a pain relief medication [4]. After the discovery of morphine, many other compounds with therapeutic effects were isolated and purified from plants.

The botanical drugs we use today, like the ancient herbal remedies, are all examples of complex mixtures enriched in plant secondary metabolites. Along evolution, plants have in fact developed a vast array of chemical defenses to stand up against their enemies (herbivores, fungi) to attract pollinators or to disseminate various chemical signals in their surrounding environment. Secondary metabolites are present in all higher plants but display a large structural diversity: different taxa usually accumulate different classes of secondary metabolites, reflecting the adaptations to the various ecological niches plants colonized on Earth [5]. This is in contrast to the current knowledge about the role and distribution of primary metabolites (amino acids, organic acids, carbohydrates, etc.). Primary metabolites represent the intermediates of those metabolic pathways related to the basic processes of plant growth and development (e.g., glycolysis, TCA cycle, ATP (adenosine triphosphate) synthesis, Calvin-Benson cycle, etc.); as such, their presence is not confined to specific taxa, and the key metabolic steps for their biosynthesis and degradation are mostly conserved across the green lineage. So although most of the primary metabolic pathways have been well described in plants, both at the genetic and biochemical level, the elucidation of the pathways of secondary metabolites has lagged behind, due to their confined taxonomic distribution and inherent difficulties in purifying them from natural sources (due both to their low amounts and chemical complexity). The study of plant secondary metabolism is thus of interest not only for answering basic research questions—such as the evolution of metabolic pathways, the extent of natural metabolic diversity, and pathway regulation in relation to the environmental conditions—but also from an applied perspective, given that most of the natural products of medicinal importance are actually secondary metabolites.

Thus, although nearly 400,000 flowering plants have been classified so far, only a fraction of these, around 20,000, has been used since ancient times for medicinal purposes [5], and only a minority of these has been studied in detail with regard to the metabolic composition and biological effects of their crude extracts [6]. The Dictionary of Natural Products, for example, which is a curated database of various chemical entities isolated from plants and microbes, contains around 160,000 entries; this number is, however, considered a round-down approximation of the extent diversity of secondary metabolites in higher plants [7].

Today, almost 30% of the new chemical entities released by the FDA (Food and Drug Administration) are either (entirely) natural products, botanical drugs, or semisynthetic derivatives of a natural product [8]. The pharmaceutical industry has been, however, rather reluctant in investing in large-scale screening of natural products for drug discovery [6]. One of the reasons limiting the screening of small molecules in plants has been the inherent difficulties in the purification of known compounds in adequate yields, but also, as we have mentioned above, the incomplete knowledge of many of the biosynthetic pathways of secondary metabolites [9–11]. The full knowledge of the pathways of plant secondary metabolites is of course essential to develop alternative strategies of production in heterologous hosts (yeast, bacteria) for pharmaceutical applications [12].

The advances in a number of systems-biology disciplines (genomics, transcriptomics, metabolomics, and computation biology), however, fueled by the decreasing costs for generating large-scale molecular data, are revolutionizing our research approaches also in the field of medicinal plants.

In the present review, we will present examples where the application of traditional biochemical and omic-based approaches contributed to new discoveries in the pathways of some secondary metabolites of medicinal importance. We will not cover in detail the knowledge acquired so far on the chemistry of natural products (but we refer the reader to recent excellent reviews on the subject: [13] for benzylisoquinoline alkaloids, [14] and [15] for monoterprenoid indole alkaloids (MIAs), [16] for cannabinoids, [17] for xanthine alkaloids, [18] for ginsenosides, [19] for withanolides, [20] for artemisinin, and [21] for taxol), and we instead focus on the historical developments and the advances made recently in completing the missing parts of the puzzle in the biosynthesis of some important natural products. In the first part of this review, we chose to focus on the cases of benzylisoquinoline alkaloids (BIAs), MIAs, cannabinoids, and caffeine, as they represent exemplary cases of how the application of several approaches, based on the integration of genomics and metabolomics, has helped clarify specific biochemical steps or entire pathway branches that had remained elusive. In the second part of this review, we will summarize the knowledge acquired so far on the biosynthesis of specific compounds (ginsenosides, withanolides, artemisinin, and taxol) from other important medicinal plants where we believe integrative approaches could help further the elucidation of their secondary metabolism with a view on the discovery of novel metabolites of medicinal importance. For each presented case study, we survey the health-related benefits and current medicinal use of these compounds and how traditional “reductionist” and integrative approaches are accelerating the development of metabolic engineering strategies (in heterologous and native hosts) for the production of secondary metabolites of pharmaceutical interest.

**Approaches for Pathway Discovery**

Traditionally, the first steps in the elucidation of plant metabolic pathways were based on the identification of a rather limited number of primary metabolites and on the use of radioactive labels to follow their fate. These were essentially the approaches that brought to the discovery of the reactions of the path of carbon in photosynthesis: the strategy was based on exposing a green algae to a stream of 14C-labeled CO2, followed by extrac-
tion, separation, and identification of metabolites with paper chromatography. The gradual decrease of the exposure time to labeled CO$_2$ allowed, for example, the identification of the product immediately downstream of the CO$_2$ fixation reaction (phosphoglyceric acid [22, 23]). Similarly, the remaining intermediates of the various reactions were identified, increasing the exposure time to labeled CO$_2$ [24]. With the advent of recombinant DNA technology, these initial labeling approaches were combined with the isolation of the respective genes and with the synthesis and purification of candidate enzymes. The advent of these new technologies was also accompanied by an increasing interest toward secondary metabolites, which were initially considered only as “waste” products of primary metabolites, with no physiological or ecological role [25]. The use of molecular biology techniques (i.e., molecular cloning and heterologous expression systems) along with classical protein biochemistry allowed, for example, to assess in vitro the catalytic properties, substrate specificities, and identity of the products for a large number of enzymes involved in secondary metabolism (and several examples from these early, targeted approaches for pathway discovery of medicinally important phytochemicals are reported in this review). In recent years, the leap of genomic technologies, with the relative ease in collecting large-scale sequence data, has bred new life into metabolism research [26]. The increasing number of available genome sequence is now frequently integrated with high-resolution/deep-coverage metabolomics approaches [27] not only to uncover structural and regulatory aspects of pathways of secondary metabolism, but also to go deeper into the evolution of metabolism across the diversification of land plants (and landmark examples in this area are the recent reconstructions of the synthesis of nicotine and caffeine [28, 29]). The case studies presented here thus represent successful examples of how targeted molecular approaches or, more recently, the combination of next-generation genomics with metabolic profiling are revolutionizing the field of medicinal plants with new knowledge concerning the synthesis of natural products.

Benzoisoquinoline Alkaloids

BIAs represent perhaps the oldest medicines humans have used to treat pain. These alkaloids belong to a large family with over 2500 known structures; they are mostly restricted to members of the order Ranunculales (in particular, they are present in the families Papaveraceae and Berberidaceae), Magnoliidae, and Laurales. Among the Papaveraceae, opium poppy (Papaver somniferum L.) has emerged as the model species to study metabolism of important BIAs, as this plant accumulates large amounts of different subgroups of these alkaloids [30]. The most abundant BIAs in roots of opium poppy are those of the benzophenanthridine-type (e.g., sanguinarine, a potent anti-inflammatory agent that has also showed antitumor properties [31, 32]), while the latex preferentially accumulates varying amounts of morphine and codeine (“morphinans”). Although the increasing use of opioid drugs (natural morphinans and their semisynthetic derivatives like oxycodone) in clinical practice is now raising concerns given their history of abuse, there is no doubt that morphine and codeine represent effective analgesics in the treatment of severe pain, at least in the short-term following an acute trauma [33]. The initial isolation of morphine from the latex of opium poppy stimulated further research into the elucidation of its biosynthesis in plants. The first studies were based on radiolabel incorporation of a few candidate substrates and established tyrosine and its derivatives as the precursors of morphine [34, 35]. We now know, after decades of research that have seen the application of more detailed tracer studies and biochemical characterization of the related enzymes, that the biosynthesis of BIAs involves a highly branched network of chemical transformations starting from two tyrosine derivatives, dopamine and 4-hydroxyphenylacetaldehyde (4-HPAA) [13]. These two metabolites condense to give rise to (S)-norcoclaurine, which is in turn modified by a number of O-, N-methyltransferases and oxidoreductases to produce (S)-reticuline, the precursor of almost all subgroups of BIAs. From (S)-reticuline, the pathway diverges into different branches, which may be active only in some species or tissues, resulting in the wide structural diversity of BIAs subgroups that has been observed in plants (Fig. 1).

The early efforts in the elucidation of BIA biosynthesis were based on the purification of the putative enzymes and on the screening of cDNA libraries for the isolation of the corresponding genes; these initial studies allowed, for example, the characterization of norcoclaurine synthase, the enzyme responsible for the condensation of dopamine and 4-HPAA, producing (S)-norcoclaurine [36, 37]. Similar approaches have been followed in the elucidation of the remaining early steps of the BIA pathway: the synthesis of (S)-coclaurine, for example, by the action of a 6-O-methyltransferase (norcoclaurine 6-O-methyltransferase, 6OMT) [38, 39] or, analogously, the synthesis of (S)-N-methylcoclaurine by a N-methyltransferase (coclaurine N-methyltransferase, CNMT, [40]). The late steps of morphinan biosynthesis remained instead uncharacterized until the development of global gene expression resources for opium poppy. After screening a number of varieties and mutants differing in their accumulation of morphine, two candidate genes were eventually proposed on the basis of the correlation of their expression with the accumulation profiles of morphinans. The discovery of these two genes, thebaine-6-O-demethylase (DIOX1) and codeine-O-demethylase (DIOX3), was thus made possible by the development of cDNA microarrays from an opium poppy EST (expressed sequence tag) database [41]. The advent of these "early" global gene expression resources in P. somniferum (ESTs collections, microarray) heralded a new era in the study of BIA metabolism. Additional gene expression resources—based on next-generation sequencing—were developed and integrated with metabolomics and proteomics studies in order to identify novel gene candidates [42]. As an example of this approach, known cytochrome genes of the CYP80B3 and CYP82N3 subfamilies, responsible for hydroxylating (S)-N-methylcoclaurine and protopine, respectively, were used as queries in a co-expression analysis to discover additional BIA biosynthetic genes in several accessions of opium poppy [43].

More recently, integrative approaches based on the combination of gene expression analyses and metabolic profiling were also fundamental in unveiling the nature of a biochemical step in BIA biosynthesis that had remained elusive for a long time. The first step of the morphinan branch is the conversion of (S)-reticuline
into its R stereoisomer; the reaction is a two-step process involving the oxidation of (S)-reticuline to 1,2-dehydroreticuline and the subsequent reduction to (R)-reticuline. Although the reaction was supposed to be catalyzed by two different genes, in agreement with the reaction being a two-step process, screening of transcriptome libraries from opium poppy identified instead a single fused gene composed of two domains. This gene, named STORR (from S-to-R-reticuline), encodes a unique bifunctional protein containing a P450 monoxygenase at the N-terminus and an oxidoreductase at the C-terminus [44,45]. The genetic analysis of opium mutants with impaired synthesis of morphine and high accumulation of reticuline confirmed STORR as the causal locus for the epimerization of S-to-R-reticuline. Bifunctional genes like STORR, including monooxygenases fused with various additional domains (hydrolase, dioxygenase), have been found also in secondary metabolic pathways of other organisms (e.g., fungi, [46]); it is thus possible that the occurrence of these genes could represent a sort of metabolic channeling of higher efficiency, in which highly unstable intermediates—like those formed during an epimerization reaction—are converted into final products by the action of a single protein rather than by a multienzymatic assembly [44].

Another example of the application of integrative approaches to the metabolism of BIAs lies in the elucidation of the biosynthesis of noscapine. This alkaloid belongs to the phthalideisoquinoline subgroup of MIAs; it was already widely used for its antitussive properties but has recently been demonstrated to possess antitumor activity given its ability to bind tubulin and arrest cell division in a number of cancer cell lines [47]. It was later showed that noscapine specifically targets the NF-κB signaling pathway in tumor cells, repressing proteins involved in cell invasion and tumor proliferation [48]. Early radionlabeling experiments in the 1960s

![Fig. 1 BIAs biosynthetic pathways of P. somniferum (opium poppy) discussed in the text. All BIAs derive from (S)-norcoclaurine, the product of the condensation of two tyrosine derivatives, dopamine and 4-HPAA. After a series of O-, N-methyltransferase and hydroxylation reactions, (S)-norcoclaurine is converted into (S)-reticuline, the central precursor of all BIAs biosynthetic branches. NCS: norcoclaurine synthase; NMCH: (S)-N-methylcoclaurine 3′-hydroxylase; 4′-OMT 3′-hydroxy-N-methylcoclaurine 4′-hydroxylase; STORR: (S)-to-(R) reticuline (aka REPI, reticuline epimerase); P6H: protopine 6-hydroxylase; DBOX: dihydrobenzophenanthridine oxidase; SalSyn: salutaridine synthase; SalR: salutaridine reductase; SalAT: salutaridinol 7-O-acetyltransferase; T6ODM: thebaine 6-O-demethylase; CODM: codeine O-demethylase; COR: codeinone reductase; SOMT1: scoulerine 9-O-methyltransferase; CAS: canadine synthase; TNMT: tetrahydroprotoberberine N-methyltransferase; NOS: noscapine synthase. Dashed arrows indicate multiple steps.](image-url)
traced back the origin of noscapine to (S)-scoulerine [49], which is produced starting from (S)-reticuline by the action of a FAD-linked (FAD: flavin adenine dinucleotide) oxidoreductase (BBE, berberine-bridge enzyme). From (S)-scoulerine, the synthesis of noscapine requires at least six additional biosynthetic steps, including O- and N-methylations and several oxidations, but only recently could the complete pathway to noscapine be elucidated in detail. The clarification of the pathway was made possible thanks to the availability of opium poppy varieties accumulating different amounts of noscapine and morphinans. Stems and capsules of these varieties were subjected to RNA sequencing and metabolic profiling to identify genes specifically expressed by the high-noscapine variety (HN1). A number of O- and N-methyltransferases, along with several cytochrome P450s, were found to be highly expressed only in the high-noscapine variety. Genomic analysis showed that these genes were actually exclusive of HN1. A mQTL (metabolic quantitative trait loci) analysis for noscapine content in an F2 population identified a single locus that was found to be strongly linked to the high-noscapine phenotype in the segregating generation. The locus contained a cluster of 10 genes spanning 220 kbp; the clustered genes corresponded to those previously identified as being exclusively present in HN1 [50]. The reconstruction of the pathway was also supported by virus-induced silencing of the cluster genes, thus allowing to confirm the role of each gene and measuring the accumulation of the various intermediates [50]. The occurrence of the high-noscapine cluster was not a feature unique to BIA metabolism: cluster organization is in fact a recurrent feature in the genomic organization of pathway genes of secondary metabolism [51].

The knowledge acquired so far on the biosynthesis of medicinally important BIs has of course allowed the transfer of partial or entire pathways into non-plant hosts. Strategies for chemical synthesis of morphinans (e.g., morphine, codeine, aka "opiates") have in fact been demonstrated not to be economically feasible; therefore, the licit cultivation of opium poppy is the only source of "opiates," from which several semisynthetic derivatives ("opioids") can be also obtained through semisynthesis (e.g., hydrocodone, [52]). Synthesis of thebaine and hydrocodone, for example, has been obtained in yeast starting from common precursors of primary metabolism. This has required the (over)expression of over 20 genes from yeast itself, plant (P. somniferum and Papaver bracteatum Lindl.), bacteria, and mammals. Many of the genes transformed into yeast were specifically engineered to increase their activity and stability (e.g., through site-specific mutagenesis to make the enzymes less sensitive to feedback inhibition or to modify their glycosylation patterns); although the fermentation titers for the production of thebaine and hydrocodone remained nevertheless low, especially when compared with the yields obtained with direct purification from opium or semisynthesis, the results obtained so far represent a starting point for further optimization of an alternative strategy for opioids production [53, 54]. Similar strategies have also been followed in Escherichia coli [55] and in yeast for the synthesis of dihydrosanguinarine, a BIA showing antitumor activity [56].

Monoterpenoid Indole Alkaloids
MIAs represent another class of important alkaloids whose biosynthesis has been studied in detail due to their diverse pharmacological effects. Vinblastine and vincristine, for example, two MIAs of the bisondole-type isolated from the plant Catharanthus roseus (L.) G. Don (Madagascar periwinkle, family Apocynaceae), show toxicity to white blood cells of mammals and are used today as effective medications to treat tumors like lymphoma and myeloma [57]. Other important MIAs include camptothecin, an inhibitor of DNA topoisomerase I isolated from the tree Camptotheca acuminate Decne (Nyssaceae) (irinotecan, a semisynthetic derivative of camptothecin, is one of the most diffused chemotherapeutic in the treatment of colon cancer) and quinine, an antimalarial isolated from the bark of the Cinchona trees (Cinchona spp.). Quinine is still in use today, although it has been replaced by artemisinin as the recommended first-line treatment for malaria.

MIAs constitute a large family, with over 3000 structures identified to date. They are mostly confined to plants of the order Gentianales, in the family of Apocynaceae, Loganiaceae, and Rubiaceae. The species C. roseus, which synthesizes over 150 different MIAs, has emerged in this case as the model plant for studying the biosynthesis and regulation of this important class of alkaloids [30].

The biosynthetic pathway of MIAs is complex. As an example, the complete biosynthesis of vinblastine in C. roseus proceeds through at least 30 enzymatic steps, which take place in several different tissues (phloem-associated parenchyma, epidermis, mesophyll, laticifer) and subcellular compartments (plastid, nucleus, ER [endoplasmic reticulum], and vacuole) [58–61]. The chemical complexity of most of the active MIAs hampered developments in chemical synthesis; this factor, combined with the general low number of MIAs recovered from plant sources, drove efforts toward the elucidation of biochemical pathways as a necessary step to develop metabolic engineering strategies. We will thus first summarize here the main branches of the MIA biosynthetic pathway to later focus on the recent discoveries made in the elucidation of the steps that were previously poorly characterized.

As the name suggests, all MIAs contain a terpenoid and an indole moiety. The terpenoid moiety derives from secologanin, a cyclic monoterpene formed from geraniol. The indole moiety of MIAs is instead coming from tryptamine, as a result of the decarboxylation of tryptophan. Tryptamine and secologanin then condense to give rise to strictosidine, the precursor of all MIAs. The whole pathway thus consists of four main parts:

- The first part is the synthesis of geraniol, through the plastid MEP (methylerythritol 4-phosphate) pathway. Although two different routes exist in plants for the synthesis of terpenoid precursors (the cytosolic mevalonate and the plastidic MEP pathway [62]), early labeling studies supported the origin of the terpene moiety of MIAs from the MEP pathway [63].
- The second part is the conversion of geraniol into secologanin in a series of eight steps that have been elucidated recently (ribo- doid pathway [64–68]) (Fig. 2).
- The “mid-pathway” then involves the formation of strictosidine starting from secologanin and tryptamine [69], its degly-
cosylation [70], and a series of downstream transformations whose steps have been clarified, in part, only recently [71, 72] (Fig. 3).

Finally, the “late-pathway” converts tabersonine, a downstream product of strictosidine, into vindoline, the immediate precursor of vinblastine [73–75] (Fig. 3).

The first approaches in the elucidation of the steps of MIA biosynthesis were mostly based on conventional strategies starting from the purification of the single enzymes, analysis of their AA (aminoacid) sequences, and cloning of full-length clones from cDNA libraries using degenerate primers. This was the approach followed, for example, for the identification of geraniol-10-hydroxylase (G10H), the enzyme responsible for the hydroxylation of geraniol, the first step of the iridoid pathway [76, 77] and for the purification and cloning of strictosidine beta-glucosidase [70, 78]. More recently, several transcriptome resources and databases have been developed in C. roseus, and these have been used for initial selection of candidate genes of MIA biosynthesis [79–85].

As an example of this approach, transcriptome datasets from several tissues of a C. roseus plant [68, 86] have been screened to identify the gene responsible for an elusive step in iridoid biosynthesis, the cyclization reaction of 10-oxogeranial into iridodial (iridoid synthase). Since the reaction was known to occur in the presence of NADH (nicotinamide adenine dinucleotide [reduced])/NADPH (nicotinamide adenine dinucleotide phosphate [reduced]), the genes using these two cofactors were first selected from the entire transcriptome dataset; then only the transcripts showing a similar expression profile to that of G10H (an upstream gene in the same pathway) were retained and considered as candidates for iridoid synthase. The transcript showing the highest correlation to G10H was selected for functional validation. The expression of the enzyme in E. coli showed that it was able to convert 10-oxogeranial into cis-trans nepetalactol (which is in equilibrium with cis-trans iridodial), and VIGS (virus-induced gene silencing) of
the candidate gene in *C. roseus* confirmed downregulation of the transcript and the lower accumulation of several MIAs downstream of iridoid synthase (e.g., vindoline and catharanthine) [68].

Mining the expression databases from *C. roseus* and analysis of coregulation with additional known genes, proved to be useful also for the discovery of other genes involved in the remaining steps of iridoid biosynthesis [64].

One of the most interesting features of MIA biosynthesis is the spatial distribution of its enzymes. The various parts of the pathway operate in fact in different cell types: (i) the MEP reactions and the early reactions of iridoid biosynthesis occur in the phloem-associated parenchyma; (ii) the remaining steps of the iridoid pathway and the “mid” reactions take place in the epidermis, while (iii) the reactions of the late pathway occur in laticifers [75]. Adding to this complexity, the reactions taking place in the leaf epidermis are also compartmentalized at the subcellular level.

The condensation of tryptamine and secologanin to form strictosidine occur in事实 in the vacuole, while the downstream transformations of strictosidine occur in the nucleus and in the cytosol [87]. In particular, the physical separation between the synthesis of strictosidine (vacuole) and its immediate successive step, deglycosylation (nucleus), implies the existence of an export system from the vacuole. Transporter genes have long remained elusive in MIA biosynthesis, with only two systems characterized to date: the export of catharanthine (the immediate precursor of vinblastine) to the leaf surface [88] and the sequestration of vindoline inside the vacuole of mesophyll cells [89]. Also in this case, however, the recent developments of transcriptome resources, combined with functional studies *in planta*, allowed the elucidation of a transporter gene responsible for the export of strictosidine from the vacuole to the cytosol [90]. In order to identify transporter genes, self-organizing maps (SOMs) have been used to cluster all
transcript contigs according to the similarity of their expression profiles across a wide range of tissues and developmental stages. The high-quality nodes of the SOMs that contained known MIA biosynthetic genes were then retained and inspected for the presence of putative transporter genes. This led to the identification of a candidate transporter of the NPF (nitrate/peptide family) family (CrNPF2.9). Further analysis confirmed the role of this gene in the export of strictosidine from the vacuole. For example, transient silencing of CrNPF2.9 in leaf of C. roseus led to a necrotic phenotype, probably as a result of the increase in the vacuolar accumulation of strictosidine [90].

As in the case of BIAS, several strategies have been attempted also for production of MIAs in microbial hosts. The commercial production of vincristine and vinblastine, for example, which are powerful therapeutic agents for the treatment of several forms of blood cancer, relies entirely on extraction from plant sources. Most of the active MIAs, however, including vincristine and vinblastine, are produced in extremely low amounts, so their extraction from plant tissues is uneconomical and laborious for commercial production. The first attempt to produce MIAs in microbial hosts focused on the production of strictosidine in yeast. Strictosidine represents in fact the central precursor for a number of MIAs of medical importance (vincristine, vinblastine, quinine, strychnine, ajmalicine). Reconstitution of the pathway in Saccharomyces cerevisiae required the integration of a total of 21 genes; of these, 15 represented the entire known plant MIA pathway, while the remaining six were either duplication of yeast endogenous genes or animal-derived sequences. The transformed yeast strain also contained targeted deletions of endogenous genes to decrease the flux into competing routes. As reported already for opiate production in yeast, also in this case the final yields of strictosidine remained nevertheless low (around 0.5 mg/L) for commercial production [91]; the production of this yeast strain represents in any case the basis for further optimization of the flux toward strictosidine or as a starting point for the synthesis of non-natural products [92].

Cannabinoids

Cannabinoids constitute a group of terpenic alkylresorcinols found in Cannabis sativa L., a dioecious plant of the Cannabaceae family. They accumulate in the glandular cavity of specific types of trichomes (capitate sessile or stalked trichomes), which are particularly abundant in female flowers and, to a lesser extent, in other parts of the plant (e.g., leaves, shoots). More than 120 different cannabinoids have been isolated to date [16], although the study of their medical and pharmacological effects focused on the most abundant ones, tetrahydrocannabinol (THC) and cannabidiol (CBD) [93, 94]. Scientific studies on the medical effects of cannabinoids have helped to clarify additional biosynthetic steps [106–108]. Also, at least initially, the elucidation of the cannabinoid pathway was made difficult by the low incorporation of the label [109] and by the fact that cannabinoids occur in vivo as carboxylic acids but are then decarboxylated to neutral (active) forms during heating or smoking.

All phytocannabinoids are formed by an alkylresorcinol (phenolic) moiety coupled to a monoterpene (▶ Fig. 4). Labeling studies using 13C-glucose showed that the monoterpene moiety derived from the plastidial MEP pathway, while the alkylresorcinol was produced through the polyketide pathway [110]. The first step in the synthesis of THCA and CBDA is the condensation of oleic acid (OA, an alkylresorcinol) with geranylpyrophosphate (GPP), leading to cannabigerolic acid (CBGA), the immediate pre-
cursor of THCA and CBDA. The reaction is catalyzed by an aromatic prenyltransferase (geranyl pyrophosphate: olivetolate geranyltransferase, GOT), which was isolated in 1998 [111]. The gene (CsPT) was later cloned and shown to be expressed in leaves, flowers, and trichomes [112, 113].

CBGA is then the substrate of two different FAD oxidases: the tetrahydrocannabinolic acid synthases (THCA synthase) and the cannabidiolic acid synthase (CBDA synthase), which produce, respectively, THCA and CBDA. The two genes, which share 84% similarity, are encoded by different loci [114]. Both THCA and CBDA synthase were purified through enzymatic assays from crude extracts and their respective genes cloned using degenerate PCR primers (THAS: [115]; CBDA: [116, 117]).

The steps leading to the synthesis of the alkyresorcinol precursor of cannabinoids, OA, have, however, remained elusive, and it was not until recently that these biosynthetic steps have been clarified. OA was long supposed to be synthesized starting from hexanoyl-CoA through successive condensations with three molecules of malonyl-CoA, in a series of steps catalyzed by a type III polyketide synthase (PKS, [118, 119]). A type III PKS cloned from cannabis leaves (named tetraketide synthase, TKS), however, did not produce OA and was instead shown to accumulate, among other by-products, α-pyrones [120]. These metabolites were typical downstream products of polyketide pathways in bacteria lacking polyketide cyclase activity [121]. On the basis of this, candidates with structural similarity to polyketide cyclases were selected from an EST library of cannabis trichomes, leading to the identification of a member of the dimeric α+β barrel protein superfamily (DABB superfamily). This gene, which was distantly related to type II polyketide cyclases of bacteria (Streptomyces), was able to convert, in the presence of TKS, hexanoyl-CoA and malonyl-CoA into OA, acting effectively as a noncanonical polyketide cyclase [107]. A similar approach, based on mining the same EST database from cannabis trichomes, was also used to identify the acyl-activating enzyme responsible for the synthesis of hexanoyl-CoA, the first step of the polyketide pathway in cannabinoid biosynthesis [108].

The elucidation of the steps in the biosynthesis of the main phytocannabinoids opened the possibility to transfer the pathway to heterologous hosts for commercial production of THCA/THC and CBDA/CBD. These two cannabinoids have in fact several pharmacological effects. THC, the neutral psychoactive form of THCA, targets mainly the CB1 receptor in the central nervous system and has analgesic and antispastic activities. Its consumption is, however, associated to well-known side effects (memory loss, de-
increased coordination, and, in some individuals, anxiety [122]). CBD, on the other hand, may reduce the side effects of THC and has shown pharmacological potential to reduce inflammation and symptoms of epilepsy [123]. Sativex, which is the only cannabinoid-based drug approved so far in 27 countries, is a mouth spray of THC and CBD. This drug is used today to treat the spasticity associated to multiple sclerosis [94]. Given the potential shown by THC and CBD, various strategies have been attempted in metabolic engineering of cannabinoids. Cell cultures of cannabis, even in the presence of elicitors, have resulted in limited yields, probably due to the lack of compartmentalization required by the high toxicity of cannabinoids [124,125]. A more promising approach might be represented by the production of THCA synthase in Pichia pastoris and its use in a cell-free two-liquid phase reactor to drive the synthesis of THCA. Also, this system, however, achieved relatively low yields (0.121 g·L⁻¹·h⁻¹ of THCA), probably as a consequence of the sensitivity of THCA synthase to be inhibited by its substrate [126,127].

Today, the regulations around the use of cannabis, and the research around it, are becoming less strict. Several European countries and the United States have exemptions for the medical use of marijuana; other U.S. states have legalized cannabis consumption, in moderate amounts, for personal use. Canada and Israel have funding bodies and programs specific for cannabis research. As the regulations in cannabis research will ease, we anticipate the development of additional genomic and metabolomics resources in cannabis. The integration of these resources will aid the elucidation of the full biosynthetic pathways of cannabinoids, opening the way to the discovery of novel compounds of potential medicinal importance.

Caffeine

Caffeine (1,3,7-trimethylxanthine) is a xanthine (purine) alkaloid found in guarana, yerba mate, cacao, and several species used to make tea. Traditionally, it is called guaranine when it comes from the guarana plant (Paullinia cupana Kunth, family Sapindaceae), theine when it comes from the tea plant (Camellia sinensis (L.) Kuntze, family Theaceae), and mateine in mate infusions; however, they all are the same compound. In addition, cacao, which accumulates only trace amounts of caffeine, contains the similar compound theobromine, which has similar, albeit less potent, bioactivities to caffeine. Of the species listed above, genome sequences for coffee [128], tea [129], and cacao [130] have been published indicating that at least three metabolic pathways for caffeine biosynthesis evolved independently co-opting genes from different gene families. The appearance of at least three pathways for caffeine biosynthesis in higher plants is thus an example of recurrent convergent evolution: the presence of caffeine per se in species from multiple plant orders (Malvales, Sapindales, Ericales, and Gentianales) did not always imply the recruitment of homologous genes [29] (Fig. 5). Intriguingly, this study, which relied on sequence information from five flowering species, revealed that caffeine biosynthesis was characterized by an even greater degree of convergent evolution than was previously thought, with citrus, chocolate, and guarana plants containing two previously unknown pathways of caffeine synthesis using either caffeine synthase or xanthine methyltransferase-like enzymes. Moreover, ancestral sequence reconstruction revealed that these pathways would have arisen rapidly since the ancestral enzymes were co-opted from their previous biochemical roles to those of caffeine biosynthesis. As such, this seminal paper provides a fantastic blueprint for studies into the evolution of natural product biosynthesis.

Caffeine’s exact function in planta is unclear, and two main roles, which are by no means mutually exclusive, have been proposed. In the first of these, sometimes called the chemical defense theory, caffeine is believed to protect young leaves and fruit from predators [131,132]. In keeping with this, Uefuji et al. [133] demonstrated that leaves of transgenic tobacco (Nicotiana tabacum L. [Solanaceae]) plants, engineered to produce caffeine, were less susceptible to insect feeding than leaves that did not contain caffeine. In the second, sometimes known as the allelopathic theory, caffeine is believed to be released by the seed coat to prevent germination of other seeds [134]. Evaluation of the cacao genome, the first of the three caffeine-containing species to be sequenced, suggested that cacao harbors a rich repertoire of homologs of secondary metabolism-associated genes, including pathways for oils, storage lipids, flavonoids, and terpenes as well as the alkaloid class to which caffeine belongs. The analysis of multiple metabolomics studies of this species suggests that functional prediction of the gene repertoire mentioned above was indeed largely correct [135]. The evolution of caffeine and indeed its metabolic precursor theobromine was, however, looked at in more detail following publication of the coffee and tea genomes [128,129]. Intriguingly, coffee was characterized to contain several species-specific gene families expanding including that of the xanthine N-methyltransferases (XMTs) involved in caffeine production and revealed that these genes expanded through sequential tandem duplications independently of genes from cacao and tea. As for cacao, a large number of metabolomics studies have been performed on coffee and tea identifying high contents of caffeine, quinate, and chlorogenic acid in the former [136–138] and catechins, terpenes, and caffeine in the latter [139–141]. Since there is also an increasing amount of transcriptomics data available for these species [142–148], it would appear likely that evaluating the dynamic behavior of transcripts related to caffeine biosynthesis in comparison to other unknown genes (and to the levels of the metabolites themselves) will greatly enhance our understanding as to how these pathways are controlled. One study of particular interest is the long read sequencing of the coffee bean transcriptome since this provided more and longer transcript variants specifically allowing the identification of a further 10 transcripts likely to encode key enzyme isoforms of caffeine biosynthesis [142]. This information thus greatly extends the number of candidate genes that are potentially important determinants of the final caffeine level within plant cells, and their study will thus prove instrumental in allowing rational design of metabolic engineering strategies aimed at modifying caffeine content. In addition, two other studies, this time in tea, have been highly informative in analyzing the regulation of caffeine biosynthesis. The first of these built gene regulatory networks for secondary metabolism of a wide range of tea tissues implicating a large number of transcription factors in the regulation of caffeine
biosynthesis [149]. The second article used a comparative transcriptomic and metabolomics analysis of tea and oil tea that does not produce caffeine, indicating higher expression of the key phenylpropanoid enzymes flavanone-3-hydroxylase, dihydroflavonol reductase, and anthocyanidin reductase in tea but lower levels of phenylalanine ammonia-lyase and chalcone isomerase; however, the exact link between this and the levels of caffeine is not apparent from this study [150]. Thus, these studies offer hints to the regulation; however, due to the genetic recalcitrance of the species, it will likely be several years before these can be confirmed at the molecular level.

Caffeine is a compound whose medicinal properties are at least in part offset by its addictive properties [151, 152], and as such, it remains very much debated as to how healthy it actually is. That said, a lot of the idea that coffee is dangerous springs from work in the 1970s and 1980s in which its consumption was linked to higher incidence of cancer and heart disease [153, 154]; however, much of this early research should be disregarded since it did not take into account other health-detrimental habits in the cohorts such as cigarette smoking. More recent analyses evaluating health and diet data of a cohort of 400,000 adults over a period of 13 years revealed no evidence that coffee consumption increased death from either these diseases or indeed any others with anything but a minor drop in mortality rate among regular coffee drinkers [155]. Coffee has additionally been linked to lower rates of type 2 diabetes [155], reduced risk for some cancers [156], and...
Ginsenosides have been classified according to their chemical skeleton in two different types: dammarane- and oleanane-type ginsenosides. Based on the glycosides attached, the dammarane ginsenosides are further divided into three different subgroups: PPD-type (protopanaxadiol), PPT-type (protopanaxatriol), and ocottiol-type (Fig. 6).

Recent studies showed that the molecular structure of the ginsenosides is important in defining their medical properties. The anticancer activities of these saponins depend on the number of sugar molecules and on their attachment position [162]. Protopanaxadiol and protopanaxatriol ginsenosides with no sugar residues or PPT and PPD ginsenosides containing up to three sugar residues inhibited different types of cancer, while others containing a higher amount of sugar residue showed none or very weakly antiproliferative effects [166–168]. Furthermore, it has been shown that the biological response of different types of ginsenosides is also related to the number and positions of the hydroxyl groups, which reflects the polarity of these molecules and thus facilitates the interaction with the cell membrane [169–172]. Also, differences in stereochemistry were demonstrated to produce different pharmacological effects [173].

The biosynthetic pathway of ginsenosides is not entirely characterized and many steps still need to be elucidated. The studies so far show that the main precursor used for the triterpene ginsenosides is squalene, which is formed from the condensation of two farnesyl pyrophosphate (FPP) molecule. The synthesis of each FPP requires the condensation of one dimethylallyl pyrophosphate (DMAPP) with two molecules of isopentenyl pyrophosphate (IPP). IPP can be produced in the cytosol through the mevalonic acid (MVA) pathway or in the chloroplast from the methylerythritol (MEP) pathway [62]. The role of the plastidial IPP is still unclear since ginsenoside biosynthesis mainly relies on the pool of cytosolic IPP [174], although a certain degree of compensation was observed in case of inhibition of either MEV or MEP [175].

The crucial steps in the generation of ginsenoside diversity are the cyclization of 2,3-oxidosqualene by oxidosqualene cyclases (OSC) and the subsequent hydroxylations and glycosylations [176, 177] (Fig. 6). Dammarenediol synthase (DDS) is a member of the family of OSCs, which is specifically found only in Panax species [18]. Its encoding gene has been characterized as the very first step in ginsenoside biosynthesis [178, 179].

The product of this enzymatic conversion is dammarenediol, which is the precursor of three of the four types of ginsenosides: PPD-, PPT-, and ocottiol-type. In the next subsequent reactions, the dammarenediol is hydroxylated in two consecutive reactions to protopanaxadiol and protopanaxatriol by protopanaxadiol and protopanaxatriol synthases (PPDS and PPTS, members of the cytochrome P450 family). Both protopanaxadiol and protopanaxatriol are further glycosylated by uridine diphosphate (UDP)-dependent glycosyltransferases (UGTs), whose genes remain to be identified. Extensive additional glycosyl decorations give rise to the diversity of all detected ginsenosides [180]. Recent studies provided a better understanding of a part of PPT-type biosynthetic pathway by characterization of four P. ginseng UGTs catalyzing protopanaxatriol glycosylations [181].

The biosynthesis of the oleanane-type ginsenosides starts always from 2,3-oxidosqualene, which is then cyclized to β-amyrin.
by β-amyrin synthase and converted to oleanolic acid by the action of oleanane acid synthase, member of the cytochrome P450s family. The remaining reactions, leading to glycosylated oleanane-type ginsenosides, are catalyzed by additional UGT genes that have not been identified so far (▶Fig. 6).

In the last years, a novel dammarenediol-type ginsenoside (compound K) has been isolated from human blood after oral administration of ginseng [182]. Interestingly, compound K has never been detected in Panax plants. The authors suggested that this novel ginsenoside could actually represent a minor component whose biosynthesis may actually occur in Panax plants, since the transcripts encoding two of the fundamental enzymes (CYP716A47 and UGTPg1) responsible for its conversion are present in P. ginseng tissues. Compound K could possess a number of beneficial effects for human health, given its anticancer, anti-diabetes, and anti-inflammatory properties tested in vitro [183, 184]. Currently, compound K is synthesized from deglycosylation of PPD-type ginsenosides [185].

Given the medicinal importance of ginsenosides, a number of bioengineering strategies have been developed in order to increase their production and to compensate the time required for field cultivation, which generally involves four to six years. Four different main strategies have been undertaken to synthesize ginsenosides in native and heterologous hosts: (i) developing cell and tissue culture methods [186]; (ii) adventitious root cultures [187]; (iii) transgenic plants [188]; and (iv) engineered yeast systems [189].

The first tissue culture of ginseng was reported in 1964 [190], and many other successful studies followed afterward [191,192]. The effects of different growth regulators on the final product formation have been evaluated, including sucrose (used as the most common carbon source in ginseng cultures), phosphate, copper, and nitrate. These investigations showed that the rate of biomass growth and the respective ginsenoside content correlated directly with the medium sugar concentration (up to 60 g L$^{-1}$). Higher sugar concentrations inhibited cell growth and had a negative im-
impact on ginsenoside production [193]. Phosphate, copper, and nitrates in different concentrations improved the ginsenoside yield and thus stimulated ginsenoside production in cell cultures [194, 195].

An example of the tissue culture approach is using adventitious roots as high biomass producers and studying the effect of different treatments or chemical elicitors [189, 196, 197]. As the major physiological role of the ginsenosides is related to plant defense [198, 199], stress-inducible factors have been used in order to improve their production. Treatments with methyl jasmonate and salicylic acid generally induced oxidative stress and increased ginsenoside content, as well as gamma-irradiation, which enhanced the final product up to 16-fold [200, 201].

In addition to the cell and tissue culture methods, genetic engineering methods have been used successfully to up- and down-regulate key genes involved in ginsenoside biosynthesis, such as 3-hydroxy-3-methylglutaryl coenzyme A, squalene synthase (SS), cytochrome p450 (CYPs), and DDS. Transgenic plants overexpressing these genes showed an increased amount of ginsenosides [188, 202–204].

Successful achievements of producing PPD, PPT, oleanolic acid, and compound K have been also made by using engineered yeast strains [185, 203, 205].

All these works provide an insight into the complex mechanisms of ginsenoside biosynthesis and explore new methods for large-scale production of these important pharmacological compounds. Nevertheless, many efforts still need to be done in order to further elucidate the biochemical pathways leading to ginsenoside formation, as well as to clarify the events responsible for their diversification in Panax species. Further studies are needed to improve the current available platforms and resources, as well as to advance the knowledge about their clinical applications.

Withanolides

Withanolides are a group of naturally occurring C-28 oxygenated steroidal lactone triterpenoids that have been found in at least 15 genera of Solanaceae (e.g., Withania, Tubocapsicum, Lycium, Datura, to mention few). Their presence has been reported also in Fabaceae (legumes) and Lamiaceae (the family to which most aromatic plants belong) [206]. Within Solanaceae, the shrub Withania somnifera (L.) Dunal (“Indian ginseng” or “Ashwagandha”) has been the focus of several pharmacological studies, given its wide use in Ayurveda (the major system of Indian traditional medicine) as a general tonic to increase vigor and memory and lessen the symptoms associated to rheumatism, fatigue, and dehydration [207]. On the basis of the anecdotal reports from the Ayurvedic practices, W. somnifera extracts were subjected to intense pharmacological scrutiny and showed to possess promising anti-tumor and anti-inflammation properties [208–210].

Despite the growing relevance of withanolides in medical research (which we will cover in detail further below), information about their biosynthetic routes and pathway regulation in planta remain scarce. Over the past years, more than 200 different withanolides have been isolated from roots, berries, and leaves of W. somnifera [19]; the focus of most of the pharmacological research was placed, however, almost exclusively on Withaferin A (Fig. 7), the first withanolide to be isolated from W. somnifera [211]. In general, we now know that the C28-steroidal lactones are biosynthesized from the C5-terpenoid precursors IPP and DMAPP. As in the case of ginsenosides, the key step in the synthesis of withanolides is the cyclization of 2,3-oxidosqualene. In the biosynthesis of withanolides, the product of this reaction is cycloartenol, which is then converted to 24-methylenecholesterol, the precursor of all withanolides (Fig. 7). Methylenecholesterol is then subjected to a series of hydroxylations, elongations, glycosylations of the carbocyclic skeleton, and further cyclization of its side chain, resulting in compounds with complex structural features [212–216]. According to the difference in the substituted groups of C-17 side chain, withanolides can be divided into two types: type A with a δ-lactone or δ-lactol and type B with γ-lactone or γ-lactol side chain [217]. Some recent investigations have identified putative regulatory and structural genes involved in withanolide biosynthesis [218–221].

As we have already mentioned, in the past few decades, withanolides attracted considerable research attention, and several studies were carried out to investigate the pharmacological and biological activities of this class of metabolites and their role in human medicine. Withanolide extracts from W. somnifera showed to possess anti-inflammatory, cytotoxic and antitumor activities [222]; there are also indications that the administration of Withania extracts improved memory retention in rats [223] and cognitive functions in humans [224, 225]. Withanolide A, withanolide B, withaferin A, and withanone, in particular, showed protective effect on the neuronal tissues of frontal cortex and corpus striatum in rats and prevented increase of lipid peroxidation [226, 227]. These early investigations on the effects of Withania extracts in attenuating cerebral functional deficits led to more targeted studies on the potential beneficial effects of withanolides in neurodegenerative diseases. Recent studies showed, for example, that a root extract of W. somnifera was effective in decreasing the accumulation of β-amyloid peptides in the brains of rats affected by Alzheimer’s disease [228]. Also, a crude Withania extract relieved significantly the symptoms of drug-induced parkinsonism (tremor, rigidity) in model rats [229].

Withanolides have also shown promising antitumor activities. Withanolide A and Withaferin A are two of the best studied withanolides for their capacity to significantly reduce the survival of various cancer cell lines and decrease the size of breast tumors implanted in rats [230–232]. The effect of Withaferin A, in particular, seems related to its capacity to interfere in the pathways of protein degradation and recycling (which are highly active in cancer cells), through inhibition of tubulin polymerization: this inhibition would prevent the formation of autophagy-related structures, which are essential for protein recycling [233].

Also, other withanolides (e.g., withanolide D, 17α-hydroxy-withanolide D, physagulines) were extracted from stems, roots, and leaves of Tubocapsicum anomalous (Franch. & Sav.) Makino (Solanaceae) and Physalis angulata L. (Solanaceae), and all exhibited high and significant cytotoxicity against several human cancer cell lines [234–236].

Despite the increasing evidence concerning the beneficial effects of these compounds, there are still many areas that remain to be investigated, especially regarding the biosynthesis and reg-
ulation of the withanolide pathway. *W. somnifera* is an important and highly valued plant in traditional medicine and showed promising effects in small-scale clinical trials [237, 238]. In the future, the full elucidation of withanolide biosynthesis will help to transfer the pathway to heterologous hosts for cost-effective biosynthesis of the active components; on the other hand, the development of biotechnology protocols for *Withania* spp. will guide future efforts for functional studies in this important genus and will provide the genetic materials for targeted breeding and commercial exploitation.

### Artemisinin

Artemisinin is a sesquiterpene lactone isolated from the Chinese herb *Artemisia annua* L. (Asteraceae), known as qinghaosu (sweet worm-wood) in traditional medicine, and mainly used for its antimalarial effect. In addition to that, recent studies showed promising anticancer, antiviral, and anti-inflammatory activities [239].

The first report on the healing properties of *A. annua* extracts dates back to 340 AD by Ge Hong in his book *Zhou Hou Bei Ji Fang* (*A Handbook of Prescriptions for Emergencies*). It was only in 1971, however, that the active compound was isolated and characterized, due to the work of the Chinese chemist Youyou Tu [240, 241], who was later awarded the Nobel prize for medicine in 2015 for her discovery of artemisinin.

Artemisinin became essential in the treatment of uncomplicated malaria caused by the parasite *Plasmodium falciparum* and has established itself as the most potent of all antimalarial drugs [242]. Although the mechanism of action is still not completely understood, the use of artemisinin and its derivatives in combined therapies contributed significantly to the reduction in malaria mortality [243]. Artemisinin is currently the first-line treatment against malaria [244, 245], despite the emergence in recent years of cases of resistance in Southeast Asia. Recent studies showed that the resistance is mainly due to the K13 mutation in *P. falciparum* parasites [246, 247].

Given the complex structure of natural artemisinin, the main commercial source for this compound so far is the natural plant. Artemisinin is produced by the glandular trichomes of *A. annua*, but its accumulation in planta is low (0.01–1.4% dry weight) and highly dependent on the plant variety [248]. Based on this, the extraction of artemisinin is relatively expensive and its production cannot meet the global demand.

In order to face these fundamental problems, many efforts to increase artemisinin production have been attempted. Significant results in this direction were obtained in the field of molecular biology, synthetic biology, and genetic and metabolic engineer-
ing. All these achievements would have not been possible without the characterization of the genes and enzymes related to artemisinin biosynthesis. In the early studies, radioactive-isotope labeling has been used to show that artemisinin derives from IPP and DMAPP, which are synthesized both from the cytosolic mevalonate (MVA) and from the plastidial 2-c-methyl-d-erythritol 4-phosphate (MEP) pathway [249–252]. The condensation of two molecules of IPP with one molecule of DMAPP forms FPP, which is then converted to amorpha-4,11-diene by amorphaadiene synthase (ADS) [253]. Amorphaadiene is subsequently oxidized, first, to amorphaalcohol and then to amorphaaldehyde by a CYP71AV1 and its redox partner cytochrome P450 reductase (CPR) [254, 255]. Artemisinic aldehyde is then converted to dihydroartemisiminic aldehyde by the enzyme D8R2 (artemisinic aldehyde Δ11(13) reductase) and oxidized to dihydroartemisinic acid (DHAA) by aldehyde dehydrogenase (ALDH1) [256, 257]. The export of DHAA to the trichome and its photoxidation then yields artemisinin (Fig. 8).

The elucidation of the artemisinin biosynthetic pathway has been a fundamental step in exploring and developing the bioengineering tools used to enhance its production. Different directions have been undertaken in order to improve the artemisinin biosynthesis in the same A. annua species or in different host organisms.

Germplasm selection and breeding have been used for creating superior cultivars [258]. The studies reported so far describe a number of cultivars with increased artemisinin content from 1 to 2.4% (DW), but due to instable artemisinin production, these lines have not been considered as a valuable commercial source [259, 260].

Transgenic A. annua plants have also been produced with the aim of increasing the amount of artemisinin. In general, two main strategies have been used: the first one based on the overexpression of structural or regulatory genes [261–263], and the second one based on the inhibition of competing pathways, such as, for example, the squalene pathway [264].

Overexpression of several genes responsible for key steps of artemisinin biosynthesis, such as farnesyl pyrophosphate synthase (FPS), ADS, CYP71AV1, CPR, and D8R2 led to approximately a double increase of artemisinin production [265–267].

Based on these conclusions, many research groups focused their interest in co-overexpressing two or more genes in A. annua to further increase the amount of artemisinin [262, 263]. For example, co-overexpression of FPS, CYP71AV1, and CPR genes increased the artemisinin content by 3.6 fold (2.9 mg/g fw) in comparison with control plants [267], and the simultaneous overexpression of ADS, CYP71AV1, and CPR resulted in 2.4-fold increase of artemisinin (15.1 mg/g DW) compared to control plants [268].

Recently, several transcription factors of different families, including WRKY, bHLH, NAC, and MYC have been isolated and characterized in A. annua. The overexpression of these genes also increased the final amount of total artemisinin [261, 266, 269–271].

The other approach used to enhance the artemisinin amount is to block the key enzymatic steps in competitive pathways to divert the flow predominantly into artemisinin biosynthesis [262]. Inhibition of the expression of the SS gene, which uses farnesyl pyrophosphate as a substrate and catalyzes the first step of the sterol pathway, increasing the artemisinin content up to 31.4 mg/g (a three-fold increase with respect to control plants) [264].

In order to explore the metabolic engineering approaches for alternative artemisinin production, several heterologous hosts have been tested. The steps leading to the synthesis of amorphadiene have been engineered in E. coli by introducing the MVA pathway from yeast (S. cerevisiae) and a synthetic ADS gene [272]. The results obtained reached a titer of 300 mg/L amorphadiene [273].

Another attempt to enhance artemisinin production has been made in plant hosts. Nicotiana species have been selected as potentially the most suitable ones because of their favorable characteristics (rapid growth and high biomass) [263]. An innovative approach consisted in the insertion of biosynthetic genes in both the nucleus and chloroplast genomes, leading to a final yield of 120 µg/g artemisinic acid [274]. Despite these efforts, however, the production levels in Nicotiana remained low and therefore not suitable for commercial production.

To date, the most prominent achievement in the field of metabolic engineering is the production of artemisinic acid in yeast. In this case, the MVA pathway has been introduced into S. cerevisiae along with ADS and CYP71AV1, allowing the conversion of amorphadiene to artemisinic acid in three oxidation steps. As a result, around 100 mg/L of artemisinic acid have been obtained [254]. The system was further improved by the introduction of two additional enzymes, a plant dehydrogenase (ADH1) and a second cytochrome (CYB5), which were both positive regulators of artemisinin biosynthesis. The process reached titer up to 25 mg/L of artemisinic acid, which is the maximum amount achieved so far [275]; this improved yeast system has, however, found modest market impact due to the lower costs associated to the direct extraction of artemisinin from plants [276].

Taxol

Taxol (paclitaxel) is a complex diterpenoid extracted from the bark of the Pacific yew (Taxus brevifolia Nutt., family Taxaceae), a tree native to the west coastal region of North America. In 1960, taxol was discovered during a large phytochemical screening aimed at the identification of cytotoxic natural products from plants. This effort was jointly conducted by the National Cancer Institute and the U.S. Department of Agriculture [277, 278]. Taxol belongs to a large family of taxoids (taxane diterpenoids) that accumulate in Taxus species, where they play an important role in plant defense. Taxoids deter the feeding activities of mammals and insects and protect the plants from fungi colonization [279].

Taxol is formed by a tetracyclic oxahaptenedane skeleton decorated with eight functional oxygen groups, two acyl groups, and a benzyl group [280]. After the elucidation of its structure in 1971 [277], several clinical trials led to its approval by the FDA as an anticancer drug for the treatment of a wide range of cancers (ovarian, breast, lung, Kaposi’s sarcoma, cervical, and pancreatic) [281]. Since then, taxol has become a leading anticancer drug, whose total sales exceed several billion U.S. dollars per year [282]. The mechanism of action of taxol is based on its capacity...
to interfere with the function of microtubules during cell division, causing their polymerization even at low temperatures. This property renders taxol highly cytotoxic to cancer cells [281].

The amount of taxol that can be extracted from the bark of the adult trees of *T. brevifolia* is however, extremely low. Around 12 kg of bark material yield only 0.5 g of purified taxol [278]; therefore, alternatives sources or methods for taxol production must be developed to avoid the need to rely on destructive bark harvesting [283].

In addition to that, the knowledge of the pathway of taxol biosynthesis remains incomplete. Of the 20 hypothesized enzymatic steps, only 14 have been well characterized [280, 284, 285] (Fig. 9). The current understanding of the taxol biosynthetic pathway includes at least eight oxidation steps, five acetyl/aroyl transferase steps, a C4β,C20-epoxidation reaction, a phenylalanine aminomutase step, N-benzoylation, and two CoA esterifications [282]. The presence of several putative enzymes in the pathway was recently suggested by analyzing the transcripts of *Taxus baccata* L. cells elicited with methyl jasmonate [285].

The precursors of taxol are IPP and DMAPP from the plastidial MEP pathway. Geranylgeranyl pyrophosphate synthase catalyzes the condensation of three molecules of IPP and one of DMAPP into geranylgeranyl pyrophosphate (GGPP), which is then cyclized by taxadiene synthase into taxa-4(5),11(12)-diene (taxadiene). Taxadiene is then the central precursor from which all taxane diterpenoids originate. In the branch leading to taxol biosynthesis, taxadiene is hydroxylated by different P450 hydroxylases. The order of the reactions and some of the genes responsible for these subsequent catalytic steps are, however, not clear yet: from the isolation of the putative intermediates, several hydroxylations should occur at positions C1, C2, C4, C7, and C9, as well as a further oxidation at C9 and a C4β,C20 epoxidation. The product of this series of poorly characterized steps is baccatin III, a key intermediate that can be also extracted from the needles of *T. brevifolia* and constitutes the starting substrate for semisynthesis of taxol and other taxane diterpenoids [280]. Baccatin III is then esterified on C13 with a β-phenylalanol moiety yielding 3′-N-debenzoyl-2′-deoxy-taxol, in a reaction catalyzed by baccatin III: 3-amino,13-phenylpropanoyltransferase (Fig. 9).

From 3′-N-debenzoyl-2′-deoxy-taxol, the last two steps of the biosynthesis leading to taxol require the hydroxylation and terminal N-benzoylation of the β-phenylalanine side chain by a yet uncharacterized taxane-2′α-hydroxylase and a N-benzoyl transferase (DBTNBT) [21, 285].

Today, the supply of taxol for medical use cannot be achieved from natural sources. As a consequence of the initial overharvesting of the bark for taxol extraction, *T. brevifolia* is now in a near threatened state [286]. On the other hand, total chemical synthesis of taxol, which was achieved in 1994 [287], has never been considered as an economically feasible alternative, due to the high complexity of the process. The current standard for taxol production is now semisynthesis, starting from the isolation of the intermediates baccatin III or 10-deacetylbaccatin III from *Taxus* cell cultures. Taxol can also be produced entirely from *Taxus* cell suspension cultures. The whole process, after decades of optimization based on the use of chemical elicitors (e.g., methyl jasmonate) and improvement of growth conditions, has now reached yields in the range of several hundred mg per liter of culture [282, 288].

A partial alternative to *Taxus* cell culture was represented by the transfer of the known part of the pathway—up to taxadiene—to *E. coli*. Bacteria (and yeast) offer in fact a higher growth rate with respect to plant cell cultures and are generally easier to manipulate. The insertion of two pathway modules into *E. coli* (the
 MEP pathway and the GGPP synthase/taxadiene synthase pathway) resulted in final yield of around 1 g/L of taxadiene. Although taxadiene is a distant precursor of baccatin III (and thus several steps—some of which still unknown—separate taxadiene from taxol), the metabolic engineering of _E. coli_ was an important achievement for the future full transfer of this important pathway to a microbial host [289].

Bioinformatic Resources for Medicinal Plants

In recent years, the decreasing costs associated with sequencing and assembly of genomic data led to the release of a high number of whole-plant genome sequences, including several from medicinal plants [290]. In some cases, as we detail below, this was accompanied by the development of several communal bioinformatics resources that integrated various types of omics datasets. Clearly, given the complexity of secondary metabolism of medicinal plants with respect to crops and model plants species, these resources offer the opportunity to mine specifically the metabolic pathways of medicinal plants and correlate, for example, the number of specific metabolites with the genomic data (e.g., gene expression, sequence polymorphisms). We provide below a survey of the main genomic databases that have been recently developed for some of the most studied medicinal plants.

Medicinal Plant Genomics Resource [291] is an example of a large, collaborative effort between several research institutions containing genome and metabolome data of 14 taxonomically diverse medicinal species, including _Atropa belladonna_ L. (family Solanaceae), _C. sativa_, _C. roseus_, _Panax quinquefolius_ L. The website offers an easy-to-use interface for a BLAST (basic local alignment search tool) search against the sequenced species and provides access to the various genome browsers of medicinal plants. The files related to the genome and transcript assemblies are also available for download. _C. acuminate_ (the “happy tree” of Chinese traditional medicine, [292]), _Calotropis gigantea_ (L.) W. T. Aiton (a shrub of the Apocynaceae family growing in Southeast Asia, which is known for producing cardiac glycosides [293]), and a new variety of _C. roseus_ are the latest medicinal plants whose genomic and transcriptomic data have been added to the database. The database also contains metabolic profiling data (mainly acquired through LC-MS), collected from several tissues of medicinal plants.

Another example of a resource offering a range of tools for visualization and analysis of metabolic networks and ‘omics’ data is CathaCyc, a metabolic pathway database built from metabolic and RNA-seq data of the plant _C. roseus_ [82]. CathaCyc is a repository for genes, enzymes, reactions, and pathways of primary and...
secondary metabolism; it contains 390 pathways with more than 1300 enzymes. The database also integrates the draft genome data of C. roseus [74]. The enzymes in CathaCyc have also been linked to ORCAE [294], a genome annotation resource, allowing the users to validate and edit gene annotations [295].

In 2011, a consortium of U.S. research organizations, funded by NIH, launched the project Transcriptome Characterization, Sequencing, and Assembly of Medicinal Plants Relevant to Human Health [296]. Currently, the database contains transcriptome data related to 31 species of medicinal importance, including, among others, Cinchona pubescens Vahl (the quinine tree, family Rubiaceae), Colchicum autumnale L. (family Colchicaceae, the source of colchicine), Datura stramonium L. (family Solanaceae), and Podophyllum peltatum L. (family Berberidaceae) (mayapple; the roots of Podophyllum accumulate podophyllotoxin, the precursor of the chemotherapeutic etoposide [297]).

Recently, another database has been established within the Phytometasyn project (www.phytometasyn.ca). It contains de novo transcript assemblies of around 20 medicinal plants including the plant Eschsholzia californica Cham. (California poppy, a member of Papaveraceae accumulating several active BIAs, mainly those of the pavine-type, e.g., eschscholtzidine [83]).

Future Prospects

For centuries, plants have always been used as remedies to treat a great number of symptoms. Even today, a large part of the world population relies on herbal medicines as a major source of health care, especially in Asia, Africa, and Latin America. In some rural areas, traditional medicines based on herbal drugs are the only source of health care. Almost 30% of the modern drugs we use today are actually derived from natural products; an ever-increasing number of these, coming from plants, are now in the process of being approved for market either as main active ingredients or as supplements. Several clinical trials of herbal medicines are now underway in the United States for the treatment of food allergies, asthma, and gastric inflammation [298].

We are now at the beginning of a new phase in which integrative approaches of genomics and metabolomics are applied to the study of the metabolism of medicinal plants. These approaches have begun to revolutionize our understanding of at least two main aspects of herbal medicines: (i) the biosynthesis, and pathway regulation, of many plant secondary metabolites of medicinal importance [299]; (ii) the mechanism of action of many of these plant herbal components on human metabolism and health [299, 300]. We see in this avalanche of knowledge both challenges and avenues for further research. We think there is a urgent need to develop faster, more informative and comprehensive analytical approaches for profiling and characterizing a larger number of metabolites; these challenges can be overcome also with the development of computational metabolomics strategies for metabolite annotation [301,302], de novo pathway reconstruction [303], and analysis of natural variation [304]. We clearly recognize the long history and the potential of traditional medicines as a source of well-being, but we also reason that a more intense scrutiny should be conducted on herbal drugs—including rigorous studies on their chemical composition and clinical trials—before claims could be made in relation to their therapeutic efficacy. This new knowledge could then be used—as we have seen in the case studies presented here (especially in the case of artemisinin)—to set up platforms for metabolic engineering and enable sustainable production of medicinal phytochemicals. Finding alternative ways for production of these compounds—outside of their respective native plant hosts— is also relevant to preserve natural resources in their native habitats, as the case of taxol has shown during the initial overharvesting of T. brevifolia. Scientists and policy makers need to find a better balance to promote a sustainable use of genetic resources, especially from the hot spots of world biodiversity (e.g., the Amazonian forest). A new equilibrium need to be established between ecological conservation and bioprospecting for novel drug discoveries from plants [11].

Acknowledgements

The authors thank the German Federal Ministry of Education and Research, project Plant-INNO, and the European Union’s Horizon 2020 research and innovation program, project PlantaSYST (SCA-CSA No. 739582 under FPA No. 664620).

Conflict of Interest

The authors declare no conflicts of interest.

References


Scossa F et al. The Integration of... Planta Med 2018; 84: 855–873


[37] Sung B, Ahn KS, Aggarwal BB. Noscapine, a benzylisoquinoline alkaloid, sensitizes leukemic cells to chemotherapeutic agents and cytokines by modulating the NF-kappa B signaling pathway. Cancer Res 2010; 70: 3259–3268


Frischknecht PM, Baumann TW. Stress-induced formation of purine.


Chang KL, Ho PC. Gas chromatography time-of-flight mass spectrom-

Friedman J, Waller GR. Caffeine hazards and their prevention in germi-


Heemery TE. Ginseng farm in Lawrence-County, Tennessee. Econ Bot 1977; 31: 160–162


Osbourne AE. Preformed antimicrobial compounds and plant defense against fungal attack. Plant Cell 1996; 8: 1821–1831


Han JY, Kim HJ, Kwon YS, Choi YE. The Cyt P450 enzyme CYP71A14A7 catalyzes the formation of protopanaxadiol from dammaranediol-II during ginsenoside biosynthesis in Panax ginseng. Plant Cell Physiol 2011; 52: 2062–2073

Han JY, Kwon YS, Yang DC, Jung YR, Choi YE. Expression and RNA interference-induced silencing of the dammaranediol synthase gene in Panax ginseng. Plant Cell Physiol 2006; 47: 1653–1662


Dhar N, Rana S, Bhat WW, Dhar RS, Vishwakarma R. Withaferin A induces impaired autophagy and unfolded protein response in hu...


[238] Tu Y. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. Nat Med 2011; 17: 1217


[244] Tang K, Shen Q, Yan T, Fu X. Transgenic approach to increase artemisinin content in Artemisia annua L. Plant Cell Rep 2014; 33: 605–615


This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.