D586–D594 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 17 October 2017*
doi: 10.1093/nar/gkx925

The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis

Karel Miettinen[1,2], Sabrina Iñigo[1,2], Lukasz Kreft[3], Jacob Pollier[1,2], Christof De Bo[3], Alexander Botzki[3], Frederik Coppens[1,2], Søren Bak[4] and Alain Goossens[1,2,*]

[1]Ghent University, Department of Plant Biotechnology and Bioinformatics, 9052 Ghent, Belgium, [2]VIB Center for Plant Systems Biology, 9052 Ghent, Belgium, [3]VIB Bioinformatics Core, 9052 Ghent, Belgium and [4]Plant Biochemistry Laboratory and Villum Research Center 'Plant Plasticity', Copenhagen Plant Science Center, Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark

Received August 14, 2017; Revised September 23, 2017; Editorial Decision September 26, 2017; Accepted October 03, 2017

ABSTRACT

Triterpenes constitute a large and important class of plant natural products with diverse structures and functions. Their biological roles range from membrane structural components over plant hormones to specialized plant defence compounds. Furthermore, triterpenes have great potential for a variety of commercial applications such as vaccine adjuvants, anti-cancer drugs, food supplements and agronomic agents. Their biosynthesis is carried out through complicated, branched pathways by multiple enzyme types that include oxidosqualene cyclases, cytochrome P450s, and UDP-glycosyltransferases. Given that the number of characterized triterpene biosynthesis enzymes has been growing fast recently, the need for a database specifically focusing on triterpene enzymology became eminent. Here, we present the TriForC database (http://bioinformatics.psb.ugent.be/triforc/), encompassing a comprehensive catalogue of triterpene biosynthesis enzymes. This highly interlinked database serves as a user-friendly access point to versatile data sets of enzyme and compound features, enabling the scanning of a complete catalogue of experimentally validated triterpene enzymes, their substrates and products, as well as the pathways they constitute in various plant species. The database can be accessed by direct browsing or through convenient search tools including keyword, BLAST, plant species and substructure options. This database will facilitate gene mining and creating genetic toolboxes for triterpene synthetic biology.

INTRODUCTION

Importance of triterpenes

Triterpenes compose a diverse class of plant natural products, both in structure and function. They comprise (i) primary metabolites such as the phytosterols, which are the indispensable structural components of cell membranes, and hormones, such as brassinosteroids, and (ii) specialized (also called secondary) metabolites with diverse biological functions (1,2). The latter include among others defence compounds with antimicrobial, antifungal, antiparasitic, insecticidal, anti-feedant and allelopathic activities (3) and leaf wax components (4). Primary metabolite triterpenes occur in all plant species. In contrast, specialized metabolite triterpenes are often restricted to a specific species or taxa although some of them can be quite widespread as a result of convergent evolutionary events. For example, compounds of the most common type of specialized metabolite triterpenes, i.e. the oleananes and their derivatives, are present in most flowering plant orders (5).

Specialized metabolite triterpenes show a striking diversity of structures from fairly simple unsubstituted triterpene backbones, such as the ones in leaf wax components, to complex molecules carrying multiple oxidative decorations, heterocycles, elaborate sugar chains and/or other chemical groups (2). Because these structures represent a wide range of biological activities, triterpenes have received considerable industrial interest as pharmaceuticals, cosmetics, agronomic agents, etc. Renowned examples are the quinine methide celastrol from *Tripterygium wilfordii* (thunder god vine), which has powerful antioxidant, anti-inflammatory, and anti-cancer activities (6) and QS-21, a soluble triterpene fraction from *Quillaja saponaria*, which is used as an adjuvant in vaccine formulations that are currently in clinical trials for anti-HIV, HPV, malaria, melanoma, mycobacterium tuberculosis and varicella zoster virus activities (7).

Triterpenes also have uses in the food industry, in which for instance glycyrrhizin from *Glycyrrhiza* species is used as

*To whom correspondence should be addressed. Tel: +32 9 331 38 51; Fax: +32 9 331 38 09; Email: alain.goossens@ugent.vib.be

a natural sweetener (8) and *Q. saponaria* extracts are being used as foaming agents and emulsifiers (9,10). Furthermore, plant sterols have been shown to reduce cholesterol serum levels. Plant sterols are frequently esterified with fatty acids to increase their lipid solubility when present in a food ingredient. Several studies have shown a significant reduction in the LDL-cholesterol levels when people consumed food enriched with plant sterol or stanol esters (11). This fact highly increased the interest of food manufacturers in using phytosterols as supplements and food additives and thus in phytosterol biosynthesis.

Biosynthesis of triterpenes

Plants produce triterpenes in long and branched biosynthetic pathways (Figure 1). Actual, known primary and specialized triterpenes are all built around a triterpene backbone based on six isopentenyl pyrophosphate (IPP) building blocks originating from the mevalonate (MVA) pathway (2). Part of the generated IPP pool is converted into the allylic isomer dimethylallyl pyrophosphate (DMAPP) by IPP isomerase. Two molecules of IPP and one molecule of DMAPP are assembled to farnesyl pyrophosphate (FPP) by prenyltransferase farnesyl pyrophosphate synthase. FPP is the building block for various MVA-dependent terpene pathways, including the tri- and sesquiterpenes (12). The first committed step in triterpene biosynthesis is the condensation of two molecules of FPP by squalene synthase (SQS) to produce the first triterpene, squalene. This molecule is then activated by epoxidation by a squalene epoxidase (SQE), resulting in 2,3-oxidosqualene (2,13,14).

The first branching point in triterpene biosynthesis is the cyclization of 2,3-oxidosqualene by oxidosqualene cyclase (OSC) enzymes into one of the more than hundred known triterpene backbones, such as cycloartenol for primary sterol synthesis or β-amyrin for specialized triterpenes (Figure 1). This cyclization is driven by an initial epoxy group protonation, followed by a cyclization and subsequent rearrangement of carbocation species and finally terminated by deprotonation or water addition (15). As an exception, compounds from the fern-specific class of triterpenes, hopanes, are cyclized directly from squalene by squalene cyclases (SCs) (16) (Figure 1).

The generated backbones are further functionalized by oxidations catalyzed by cytochrome P450 (P450) enzymes, which further enhances their diversity in structures and bio-activities (2,13,14,17,18). These reactions include simple hydroxylation reactions but also more complicated oxidation reactions such as additional cyclization, epoxidation, ring-opening or dealkylation (Figure 1). Finally, triterpenes can be modified by addition of sugars, sugar chains or other chemical groups to the activated carbons. Typical substitutions include glycosylation by family-1 UDP-glycosyltransferases (UGTs), but also other groups can be added and chemical modifications can be performed by e.g. acyltransferases, acetyltransferases, methyltransferases and different oxidoreductases (2,13,14,17,18) (Figure 1). Glycosylated triterpenes are usually referred to as triterpene saponins.

Phytosterols encompass some of the best-studied triterpenes. They are essential components of eukaryotic cell membranes and act as precursors for the biosynthesis of the plant brassinosteroid hormones and of specialized taxa-specific steroidal triterpenes that include the steroidal (glyco)alkaloids. Phytosterols are synthesized from 2,3-oxidosqualene via cycloartenol, generated by the OSC cycloartenol synthase (CAS), followed by a 14–17 step enzymatic pathway (19,20). This pathway contains many different types of enzymes and results in the production of sterols such as stigmasterol, β-sitosterol and campesterol, which are the most commonly consumed plant sterols provided by vegetable oils (19,21). Using partly the same enzymes, but in a parallel pathway, plants also synthesize cholesterol (22), which in turn can serve as the starting point for the steroidal glycoalkaloids in Solanaceae species (22,23). Whereas stigmasterol and β-sitosterol phytosterols are involved in the structure and function of cell membranes, campesterol is the starting point of brassinosteroid biosynthesis, plant hormones with main roles in plant growth, development and stress responses (24–26).

Enzymes from the phytosterol and brassinosteroid biosynthesis pathways have been fairly well characterized in the model plant *Arabidopsis thaliana* and to a lesser extent in rice, pea and tomato (27,28). In contrast, only a small fraction of the specialized metabolism pathway enzymes has been characterized to date, mainly because of their immense variability. Indeed, most plant families have their own unique subset of triterpene biosynthesis enzymes that contribute to a species-specific compendium of often unique structures with unique bio-activities. The diversity of specialized metabolism triterpenes arises from their modular biosynthesis where a variety of species-, genus-, or family-specific OSCs, P450s, UGTs and other decoration enzymes act in rather web-like than linear pathways, producing arrays of bio-active compounds rather than single effective compounds (1,2,13,14,17,18).

In recent years, the number of characterized triterpene biosynthesis enzymes has increased extensively, mostly because of the tremendous increase in genome and transcriptome sequencing data available. This has enabled large-scale discovery programmes based on, for example, genetics and genome cluster organization (29–31), co-expression (32) or molecular evolution of enzymes (33), and aimed at the elucidation of entire biosynthetic pathways or the characterization of entire enzyme families. Likewise, there is an increased interest in triterpene biosynthesis, which is driven by the potential commercial value of some of them and efforts towards enhanced production of triterpenes of interest in plants and in various heterologous hosts (34–39). The TriForC database presented here (i) is a user-friendly portal for the utilization of all available and validated genetic and biochemical triterpene biosynthesis resources and (ii) provides a broad view reference on plant triterpene biosynthesis.

DATABASE CONTENT

Database concept

The TriForC database (http://bioinformatics.psb.ugent.be/triforc/) is a fully manually curated resource with a fairly simple setup and comprising all plant triterpene biosynthesis enzymes biochemically characterized to date. Presently, this database includes 271 enzymes from 70 different plant

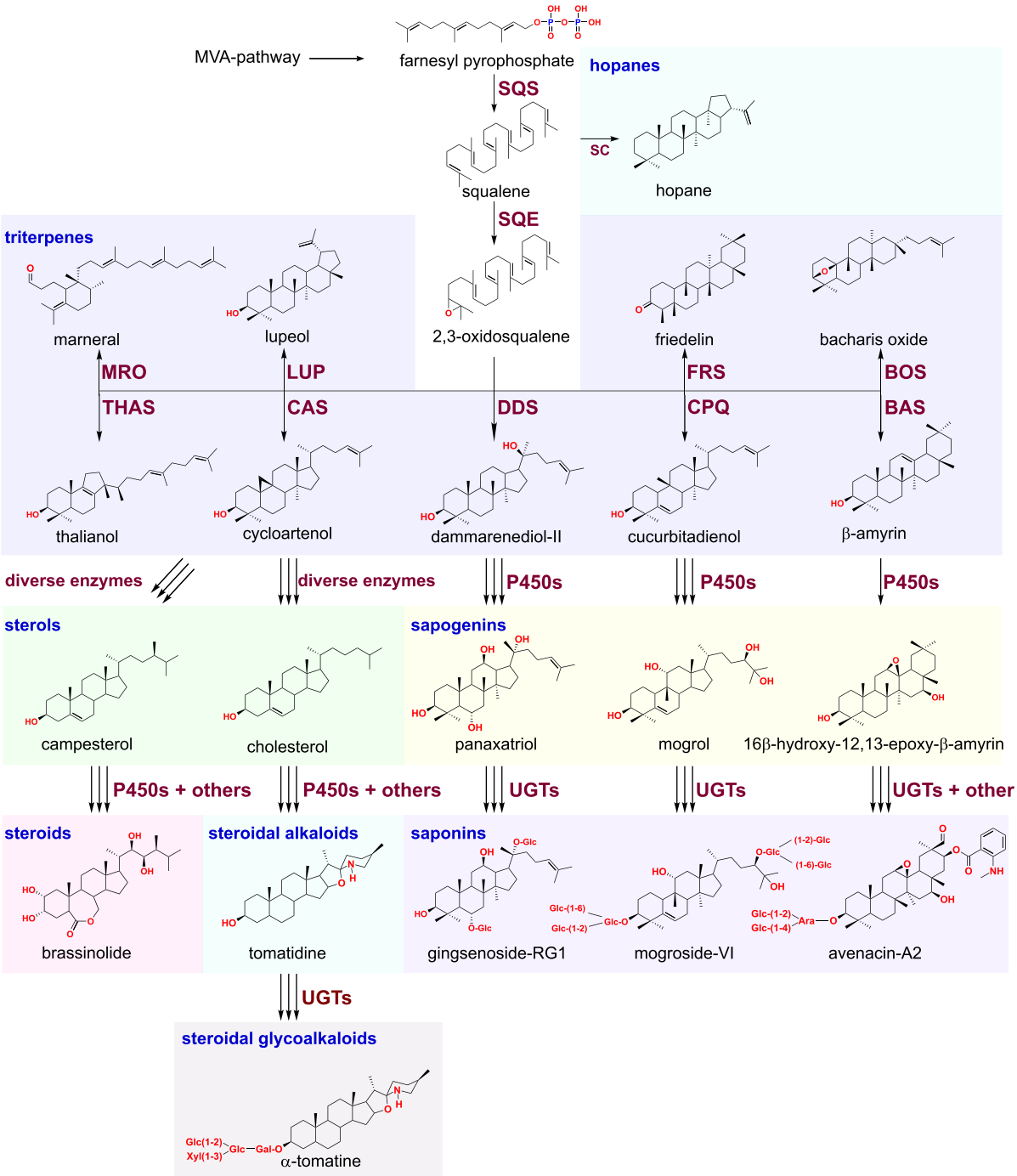

Figure 1. Overview of plant triterpene biosynthesis. Plants produce diverse types of triterpenes, some of which are depicted here. BAS, β-amyrin synthase; BOS, baccharis oxide synthase; CAS, cycloartenol synthase; CPQ, cucurbitadienol synthase; DDS, dammarenediol synthase; FRS, friedelin synthase; LUP, lupeol synthase; MRO, marneral synthase; MVA, mevalonate; P450s, cytochrome P450s; SC, squalene cyclase; SQE, squalene epoxidase; SQS, squalene synthase; THAS, thalianol synthase; UGTs, UDP-glycosyltransferases.

species that are able to catalyse >700 reactions (Figure 2 and Table 1). This information has been compiled from data published in 172 peer-reviewed scientific papers. Given their importance and general occurrence in triterpene biosynthetic pathways, we have collected data (e.g. substrate, product, pathway and species) for 14 different types of enzymes, namely OSCs, P450s, UGTs squalene synthases, squalene epoxidases, acyltransferases, methyltransferases, methyloxidases, isomerases, reductases, transaminases, desaturases, dehydrogenases/decarboxylases, and epoxide hydrolases. Accordingly, we have included 256 different compounds in the database, all linked to the corresponding enzymatic reactions, either as the substrate or the product. All of the included enzymatic reactions have been experi-

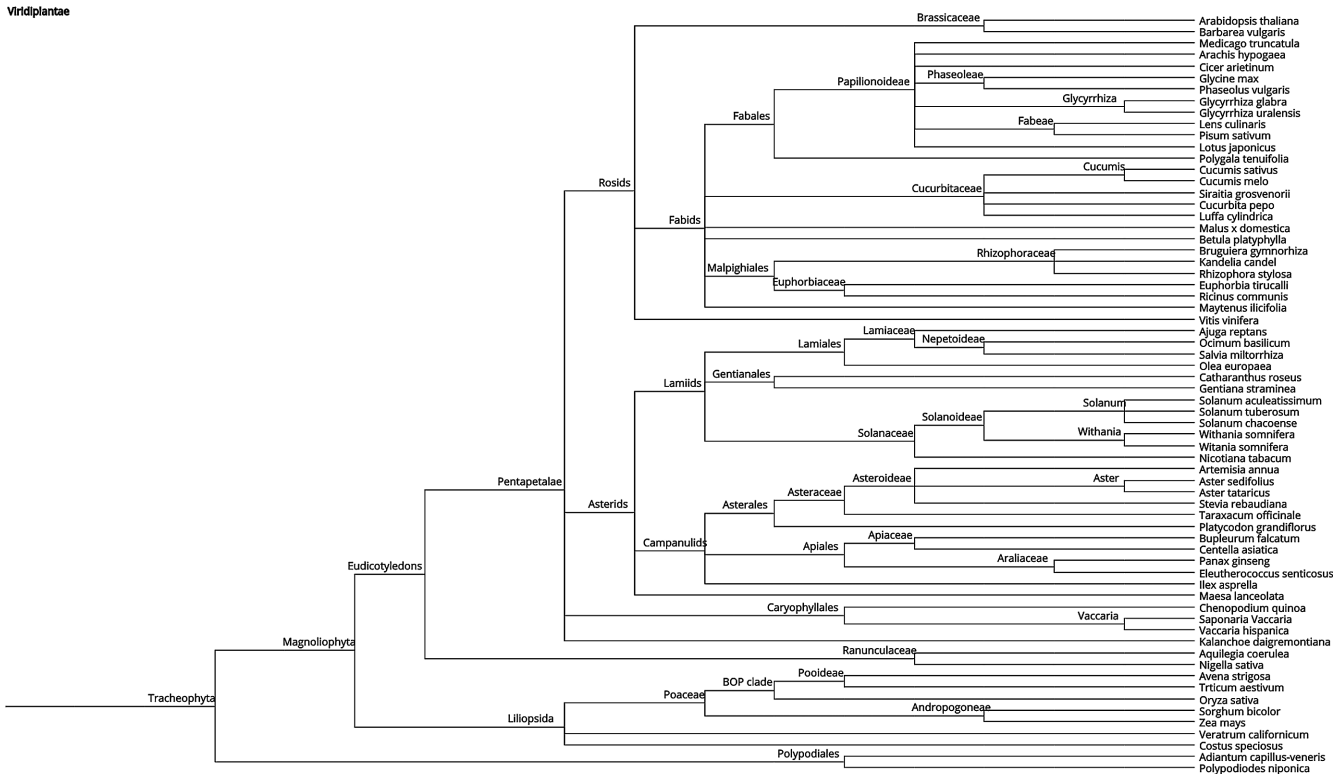

Figure 2. Screenshot of the interactive phylogenetic tree containing all plant species included in the TriForC database. All triterpene enzymes produced by each of these plants and included in the TriForC database are linked to the respective plant name.

Table 1. Current status of the data curated to construct the TriForC database

Type of information	Total number
Triterpene enzymes	271
Enzymatic reactions	716
Substrates	164
Products	266
Plant species	70

mentally demonstrated to be functional, at least in heterologous systems, such as yeast (*Saccharomyces cerevisiae*) or tobacco (*Nicotiana benthamiana*), and, in several cases, also by gain- or loss-of-gene-function analysis in the host plant.

The main characteristics of the TriForC database are that it is manually curated, complete in terms of triterpene enzymology, and user-friendly (i.e. easily accessible, easily maintainable, and easily queriable). As such, this database distinguishes itself from the well-known and excellent MetaCyc and KEGG metabolic databases that cover many more biological aspects than triterpene biosynthesis alone and, therefore, require prior knowledge of metabolic pathways for accurate usage. Accordingly, to benefit fully from these more elaborate metabolic databases, prior information about what it is expected to be found is required. In contrast, the TriForC database is a database exclusively devoted to triterpenes, directly showing the complete picture, where possible and relevant, also in a species-specific manner. The easy-to-use search tools (see the section below) allow the user to query directly for enzymes and obtain all the corresponding enzymatic reactions together with the respective substrates and products on the same web page and displayed in an interactive way. In addition, the TriForC database contains actualized information and therefore a higher coverage of triterpene biosynthesis enzymes than the previously mentioned databases. In other words, whereas the KEGG and MetaCyc databases are more comprehensive in terms of different metabolic pathways, the TriForC database is more comprehensive and up-to-date than the aforementioned databases in terms of triterpene biosynthesis and aspires to maintain that status.

Furthermore, the KEGG and MetaCyc databases compile both manually curated information and automatically generated data. Consequently, in some cases, doing a search in these databases will result in a joint listing of characterized (validated) enzymes as well as numerous 'predicted' enzymes, in some cases with erroneous annotations, thus demanding more mining time and effort from the user. In contrast, the TriForC database exclusively contains manually curated data, which will ensure high quality data and thus opens it up for researchers from different scientific fields, not only plant biology, that do not necessarily have in-depth expertise in triterpene biosynthesis but a common interest in these valuable and ubiquitous compounds.

Database structure and usage

The content of the database is organized in three different sections (or levels) that contain enzymes, compounds (substrates and products of the enzymes) and pathways. The

information present in each of these sections is fully interconnected via specific links. Furthermore, filters for enzyme type and plant species are included in the enzymes and compounds sections to narrow the search.

The enzymes section contains a list of all triterpene-metabolizing enzymes that have been validated up to date (Figure 3A). Each enzyme page presents the enzyme name, GenBank and UniProt accessions, description, protein and cDNA sequences, type of enzyme, reaction and regiospecificity, as well as links to each compound page (substrates and products), pathways, plants and scientific references (Figure 3B). Each external reference (e.g. GenBank, references) is provided as a link-out to the relevant resource.

The compounds section contains a list of all triterpene compounds included in the database. Each compound page includes the compound name, alias, molecular formula, molecular weight, Chemical Abstracts Service (CAS) registry number and links to PubChem database (40), involvement in pathways, plant association and scientific references. In addition, in each compound page, it is shown whether this compound is a 'substrate of' a determinate enzyme or the 'product of' another enzyme.

The pathways section shows reconstructed biosynthetic pathways from several plant species (Figure 4A). In this section, the user can view detailed pathway maps, for example, depicting brassinosteroid, phytosterol or a species-specific specialized triterpene saponin or steroidal glycoalkaloid biosynthesis pathway. Each pathway page contains the structure of the substrates and products together with all the enzymes that are catalysing each reaction step.

The highly-linked structure of our database ensures flexibility during the navigation of the website, which permits all information to be reachable regardless of the starting point.

Search tools

All data in the database can be browsed using the search tool that presents keyword, sequence, BLAST, substructure and plant species options.

The keyword search tool allows the user to search by compound and enzyme name, alias, CAS number, short description or GenBank ID.

Identical, homologous or similar genes or proteins included in the TriForC database can be found by submitting a FASTA query sequence (nucleotide or amino acid) in the 'BLAST search' tool.

In addition, an innovative substructure search tool has been included in our database. This tool is based on the SMILES representation (a textual representation of chemical structures) (41) and utilizes the openchemlib-js library developed for 'Wikipedia Chemical Structure Explorer' (http://www.cheminfo.org/wikipedia/) (42). This tool permits to search for certain substructures within a triterpene structure. For example, it can be used to find all structures present in the database with a given carbon backbone or with a specific functional group at a certain position of the backbone (Figure 4B). The search results that list the compounds with a similar structure are displayed in real time as a substructure is being drawn. Furthermore, this search tool is interactive, meaning that when the user modifies the query by removing or adding atoms or functional groups, a new search is automatically carried out and the new results are displayed.

Finally, the plant species search page displays a phylogenetic tree displaying all the plant species included in our database with links to the list of enzymes that can be found in each plant species (Figure 2). The plant species coverage of this database spans all vascular plants, from terrestrial or epiphytic ferns (*Polypodium*, *Adiantum*) to angiosperms. Inside the dicotyledon group, herbaceous perennial plants (*Aquilegia coerulea*, *Centella asiatica*, etc.), crops (*Solanum tuberosum*, *Cucumis sativus*, etc.), wormwoods (*Artemisia annua*), trees (*Malus domestica*) and legumes (*Lotus japonica*, *Medicago truncatula*) can be found. The selection of plant species is driven by a sole criterion, which is the existence of a peer-reviewed scientific paper describing a triterpene biosynthesis enzyme in that species.

Together, these four search tools constitute a unique search engine compendium that, to the best of our knowledge, does not exist in any other metabolic database.

CONCLUSIONS AND PERSPECTIVES

The TriForC database is a comprehensive, easily accessible and user-friendly database containing a complete and state-of-the-art overview of characterized triterpene biosynthesis enzymes. Our database permits to retrieve all available and curated information from each included enzyme and offers interactive metabolic pathway views and an unprecedented compendium of 'classic' and innovative search engines. Overall, the concept of our database, as well as the way the type of information and tools are compiled in a single database, are, to the best of our knowledge, unique among its kind. As such, our database aspires to facilitate gene mining in general and also creates a genetic toolbox for synthetic biology programmes for the production of high-value triterpene compounds. Furthermore, this database may prove useful for non-plant researchers as well, given that triterpenes are not restricted to plants, but can be encountered in any eukaryote, minimally as sterols, which are the vital components for eukaryote cell membrane structures and precursors of fat-soluble vitamins and steroid hormones.

To safeguard longevity, sustainability and robust maintenance of the TriForC database, two main measures have been taken. First, the database is hosted by our centre (VIB-UGent Center for Plant Systems Biology), which has a long-standing reputation in the generation and maintenance of databases and web tools. Second, an updating system was developed that does not require extensive professional IT skills but merely relies on a regular updating of a limited number of database spreadsheet and structure vector files with novel manually curated data. The updated files are then easily uploaded in the database webpage, ensuring the maintenance of a state-of-the-art database. We aim to actualize these files at least twice a year, when new publications become available in the triterpene field. Furthermore, in order to guarantee the correctness of the information, some data such as nucleotide and protein sequences, compound information and reference details are automatically retrieved from other databases such as GenBank, UniProt, PubChem and PubMed.

A

HOME SEARCH ENZYMES COMPOUNDS PATHWAYS CONTACT

Browse through the list of individual enzymes. You can narrow your search by using filters for enzyme type or plant species.

Type - any - Plant - any -

Name	Type	Plant	Description	Genbank	References
AaBAS	OSC	Artemisia annua	Artemisia annua beta-amyrin synthase (BAS) mRNA, complete cds.	EU330197	Moses et al., Plant Cell, 2015, Kirby et al., FEBS J., 2008,
AaCAS	OSC	Artemisia annua	Artemisia annua cycloartenol synthase mRNA, complete cds.	KM670093	Moses et al., Plant Cell, 2015,
AaLUS	OSC	Artemisia annua	Artemisia annua lupeol synthase mRNA, complete cds.	KM670094	Moses et al., Plant Cell, 2015,
AaOSC2	OSC	Artemisia annua	Multi-functional OSC	KF309252	Moses et al., Plant Cell, 2015,
AaOSC3	OSC	Artemisia annua	unknown	KM670095	Moses et al., Plant Cell, 2015,
AcACX	OSC	Adiantum capillus-veneris	Adiantum capillus-veneris ACX mRNA for cycloartenol synthase, complete cds.	AB368375	Shinozaki et al., FEBS Lett., 2008,
AsbAS1	OSC	Avena strigosa	Avena strigosa mRNA for beta-amyrin synthase (bAS1 gene)	AJ311789	Haralampidis et al., Proc. Natl. Acad. Sci. U.S.A., 2001, Papadopoulou et al., Proc. Natl. Acad. Sci. U.S.A., 1999,
AsCS1	OSC	Avena strigosa	Avena strigosa mRNA for cycloartenol synthase (cs1 gene)	AJ311790	Haralampidis et al., Proc. Natl. Acad. Sci. U.S.A., 2001,
AsOXA1	OSC	Aster sedifolius	Aster sedifolius beta-amyrin synthase (OXA1) mRNA, complete cds.	AY836006	Cammareri et al., 2008 Plant Science,
AsSAD1 (S728F)	OSC	Avena strigosa	dammarenediol-II synthase		Salmon et al., Proc. Natl. Acad. Sci. U.S.A., 2016,
AsSAD7/AsSCPL1	acyltransferase	Avena strigosa	Serine Carboxypeptidase-Like Acyltransferase	FJ475130	Mugford et al., Plant Cell, 2009,
At3βHSD/D1	dehydrogenase/decarboxylase	Arabidopsis thaliana	3β-hydroxysterol-dehydrogenase/C4-decarboxylase	AY957470	Rahier et al., J. Biol. Chem., 2006,
At3βHSD/D2	dehydrogenase/decarboxylase	Arabidopsis thaliana	3β-hydroxysterol-dehydrogenase/C4-decarboxylase	DQ302749	Rahier et al., J. Biol. Chem., 2006,
AtaSHS	OSC	Aster tataricus	Aster tataricus SHS1 mRNA for shionone synthase, complete cds.	AB609123	Sawai et al., FEBS Lett., 2011,
AtBARS1	OSC	Arabidopsis thaliana	Arabidopsis thaliana baruol synthase 1 (BARS1), partial mRNA.	NM_117625	Lodeiro et al., J. Am. Chem. Soc., 2007,
AtBAS	OSC	Arabidopsis thaliana	Arabidopsis thaliana AtBAS mRNA for beta-amyrin synthase, complete cds.	AB374428	Shibuya et al., Plant Physiol. Biochem., 2009,
AtCAMS1	OSC	Arabidopsis thaliana	Arabidopsis thaliana camelliol C synthase 1 (CAMS1), partial mRNA.	NM_148667	Kolesnikova et al., Org. Lett., 2007,
AtCAS	OSC	Arabidopsis thaliana	Arabidopsis thaliana cycloartenol synthase 1 (CAS1), mRNA.	NM_126681	Corey et al., Proc. Natl. Acad. Sci. U.S.A., 1993,
AtCYP51	P450	Arabidopsis thaliana	Obtusifoliol 14a-Demethylase	AY091203	Kushiro et al., Biochem. Biophys. Res. Commun., 2001,

B

HOME SEARCH ENZYMES COMPOUNDS PATHWAYS CONTACT

Name	AaBAS
Genbank	EU330197
UniProt	B1P7H3
Type	OSC
Description	Artemisia annua beta-amyrin synthase (BAS) mRNA, complete cds.
Plant	Artemisia annua
Reaction	cyclization
Substrates	2,3-oxidosqualene
Co-substrates	
Regiospecificity	
Products	β-amyrin
Pathways	
References	Moses et al., Plant Cell, 2015, Kirby et al., FEBS J., 2008
Genome link	
Protein sequence	MWRLKIAEGRNDPYLYSTNNFVGRQIWEFDPNYGTPEERAEVEQARVDFWNHRHEVKPSSDVLWRMQFLRE KGFEQTIPQVKIEDGEEISYEKATTTLRRSVNFFAALQADDGHWPAENAGPLYFMQPLVICLYITGHLNTV FPAEYRKEILRYIYCHQNEDGGWGFHIEGHSTMFCTTLSYICMRLLGEGRDGGLDGACTKARKWILDHGSV TTIPSWGKTWLSILGVCEWAGTNPMPPEFWILPSFLPMYPAKMWCYCRLVYMPMSYLYGKRFVGPITPLIL QLRDELYAQPYDEIKWRSIRHLCAKEDLYYPHPLLQDLMWDSLYVFTEPVLNHWPFNKLREKALQTTMKHI HYEDENSRYITIGSVEKALCMLACWVEDPNGVCFKKHIARIPDYLMVAEDGMKMQSFGSQEWDAGFAIQAL MATDLTDEIGSTLMKGHEFIKASQVKDNPSGDFKSMHRHISKGSWTFSDQDHGWQVSDCTAEALKCCLLFA TMPPEIVGEKMKPEQLNDAVNVILSLQSKNGGLAAWEPAGSSEWLEILNPTEFFADIVIEHEYVECTSSAI QALVMFKKKYPGHRKKEIENFLLGSSGYLEKIQMEDGSWYGNWGVCFTYGTWFALGGLSAVGKTYDNCPAI RKAVKFLLETQLEDGGWGESYKSCPEKKYIPLEGGRSNLVHTAWAMMGLIHSRQAERDATPLHRAAKLLIN SQLETGDFPQQEIAGVFMKNCMLHYALYRNIYPMWALADYRKQVLPQLKGT

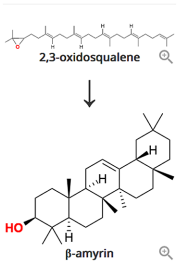

Figure 3. Screenshot of the enzyme pages of the TriForC database. (**A**) The enzyme catalogue page shows for each enzyme the name, type, plant of origin, functional description and links to the GenBank accession number and the scientific paper describing it. The enzymes listed can also be restricted per enzyme type or plant species by clicking on one of the pop-up choices in the 'Type' or 'Plant' boxes on top. Every enzyme 'line' provides a link to a page devoted to one enzyme (**B**). The AaBAS enzyme page shows all available information for that enzyme as well as links to each compound page (substrates and products), pathways, host plant species, GenBank and UniProt accession numbers and scientific references (PubMed).

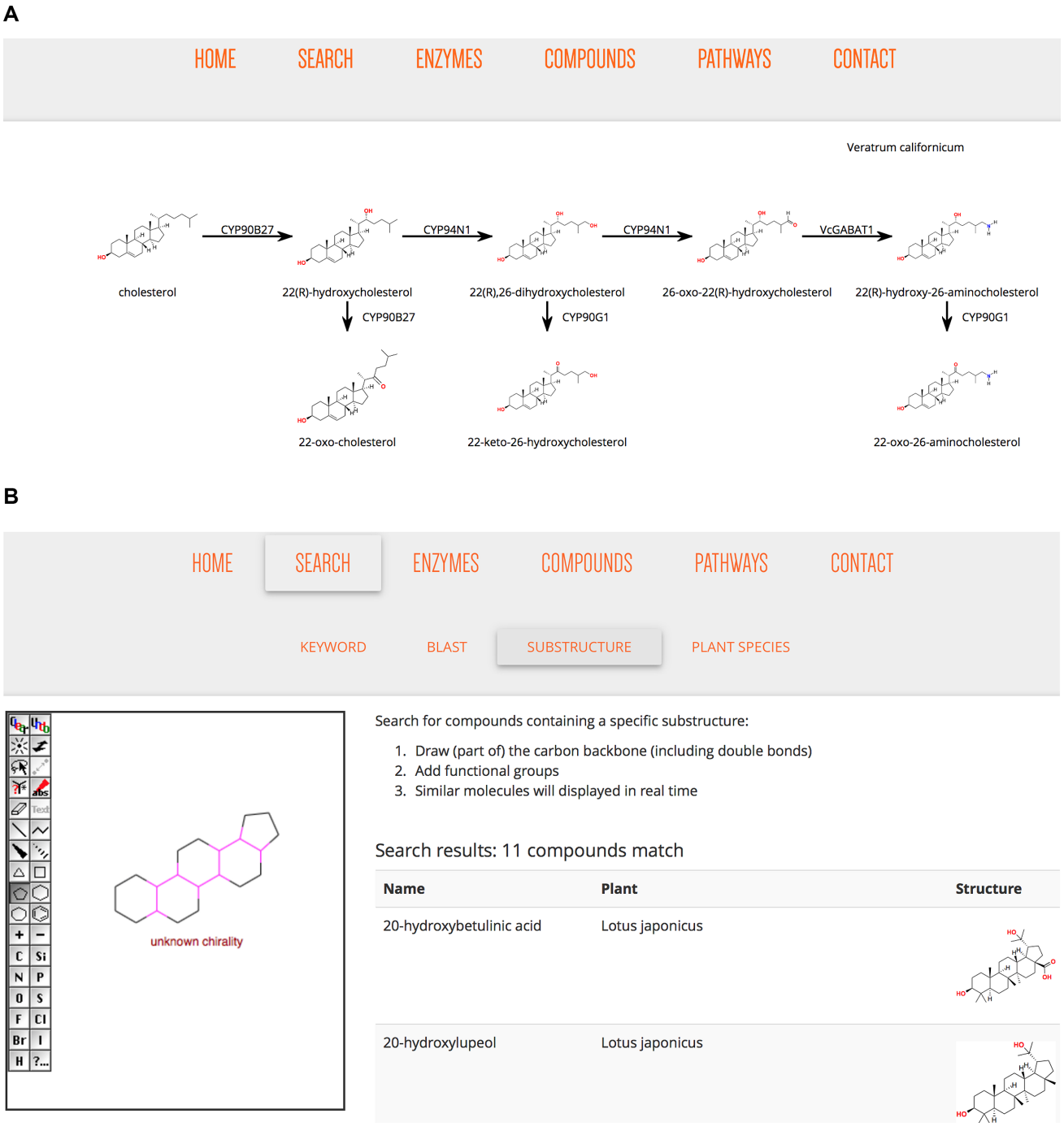

Figure 4. Screenshots of some of the TriForC database tools. (**A**) Pathway section for *Veratrum californicum*. This scheme shows the complete triterpene biosynthesis pathway that is experimentally validated from this plant species. Full arrows correspond to established enzymatic reactions catalysed by identified and experimentally validated enzymes. Dashed arrows indicate that there is a putative intermediate compound between the substrate and product that still has not been experimentally isolated. (**B**) Substructure search tool. A carbon backbone with or without functional groups can be drawn in the designated area (left box). The search results display, simultaneously while drawing, the compounds present in the TriForC database that match the drawn structure.

In the future, expansion or translation to other terpene classes could be envisaged, such as the mono-, di-, or sesquiterpene classes that are as important and diverse as the triterpenes. The way the TriForC database has been designed, this will only require the commitment and a straightforward curation effort of only a few leading researchers for each of the metabolite classes.

ACKNOWLEDGEMENTS

We thank Annick Bleys for help with preparing the article and all TriForC consortium partners for contributing to and curating the database content.

FUNDING

European Community's Seventh Framework Programme [FP7/2007–2013] under grant agreement [613692-TriForC]. Funding for open access charge: VIB-UGent Center for Plant Systems Biology.
Conflict of interest statement. None declared.

REFERENCES

1. Moses,T., Papadopoulou,K.K. and Osbourn,A. (2014) Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 439–462.
2. Thimmappa,R., Geisler,K., Louveau,T., O'Maille,P. and Osbourn,A. (2014) Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.*, **65**, 225–257.
3. Field,B., Jordán,F. and Osbourn,A. (2006) First encounters – deployment of defence-related natural products by plants. *New Phytol.*, **172**, 193–207.
4. Buschhaus,C. and Jetter,R. (2011) Composition differences between epicuticular and intracuticular wax substructures: how do plants seal their epidermal surfaces? *J. Exp. Bot.*, **62**, 841–853.
5. Vincken,J.-P., Heng,L., de Groot,A. and Gruppen,H. (2007) Saponins, classification and occurrence in the plant kingdom. *Phytochemistry*, **68**, 275–297.
6. Cascão,R., Fonseca,J.E. and Moita,L.F. (2017) Celastrol: a spectrum of treatment opportunities in chronic diseases. *Front. Med.*, **4**, 69.
7. Zhu,D. and Tuo,W. (2016) QS-21: a potent vaccine adjuvant. *Nat Prod. Chem. Res.*, **3**, e113.
8. Kitagawa,I. (2002) Licorice root. A natural sweetener and an important ingredient in Chinese medicine. *Pure Appl. Chem.*, **74**, 1189–1198.
9. Güçlü-Üstündağ,Ö. and Mazza,G. (2007) Saponins: properties, applications and processing. *Crit. Rev. Food Sci. Nutr.*, **47**, 231–258.
10. San Martín,R. and Briones,R. (1999) Industrial uses and sustainable supply of *Quillaja saponaria* (Rosaceae) saponins. *Econ. Bot.*, **53**, 302–311.
11. De Smet,E., Mensink,R.P. and Plat,J. (2012) Effects of plant sterols and stanols on intestinal cholesterol metabolism: suggested mechanisms from past to present. *Mol. Nutr. Food Res.*, **56**, 1058–1072.
12. Vranová,E., Coman,D. and Gruissem,W. (2013) Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.*, **64**, 665–700.
13. Augustin,J.M., Kuzina,V., Andersen,S.B. and Bak,S. (2011) Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry*, **72**, 435–457.
14. Gholami,A., De Geyter,N., Pollier,J., Goormachtig,S. and Goossens,A. (2014) Natural product biosynthesis in *Medicago* species. *Nat. Prod. Rep.*, **31**, 356–380.
15. Abe,I. (2007) Enzymatic synthesis of cyclic triterpenes. *Nat. Prod. Rep.*, **24**, 1311–1331.
16. Shinozaki,J., Shibuya,M., Takahata,Y., Masuda,K. and Ebizuka,Y. (2010) Molecular evolution of fern squalene cyclases. *ChemBioChem*, **11**, 426–433.
17. Moses,T., Pollier,J., Almagro,L., Buyst,D., Van Montagu,M., Pedreno,M.A., Martins,J.C., Thevelein,J.M. and Goossens,A. (2014) Combinatorial biosynthesis of sapogenins and saponins in *Saccharomyces cerevisiae* using a C-16α hydroxylase from *Bupleurum falcatum*. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 1634–1639.
18. Seki,H., Tamura,K. and Muranaka,T. (2015) P450s and UGTs: key players in the structural diversity of triterpenoid saponins. *Plant Cell Physiol.*, **56**, 1463–1471.
19. Nes,W.D. (2011) Biosynthesis of cholesterol and other sterols. *Chem. Rev.*, **111**, 6423–6451.
20. Schaller,H. (2003) The role of sterols in plant growth and development. *Prog. Lipid Res.*, **42**, 163–175.
21. Piironen,V., Lindsay,D.G., Miettinen,T.A., Toivo,J. and Lampi,A.-M. (2000) Plant sterols: biosynthesis, biological function and their importance to human nutrition. *J. Sci. Food Agric.*, **80**, 939–966.
22. Sonawane,P.D., Pollier,J., Panda,S., Szymanski,J., Massalha,H., Yona,M., Unger,T., Malitsky,S., Arendt,P., Pauwels,L. *et al.* (2016) Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism. *Nat. Plants*, **3**, 16205.
23. Itkin,M., Heinig,U., Tzfadia,O., Bhide,A.J., Shinde,B., Cardenas,P.D., Bocobza,S.E., Unger,T., Malitsky,S., Finkers,R. *et al.* (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, **341**, 175–179.
24. Belkhadir,Y. and Jaillais,Y. (2015) The molecular circuitry of brassinosteroid signaling. *New Phytol.*, **206**, 522–540.
25. De Bruyne,L., Höfte,M. and De Vleesschauwer,D. (2014) Connecting growth and defense: the emerging roles of brassinosteroids and gibberellins in plant innate immunity. *Mol. Plant*, **7**, 943–959.
26. Lozano-Durán,R. and Zipfel,C. (2015) Trade-off between growth and immunity: role of brassinosteroids. *Trends Plant Sci.*, **20**, 12–19.
27. Fujioka,S. and Yokota,T. (2003) Biosynthesis and metabolism of brassinosteroids. *Annu. Rev. Plant Biol.*, **54**, 137–164.
28. Vriet,C., Russinova,E. and Reuzeau,C. (2013) From squalene to brassinolide: the steroid metabolic and signaling pathways across the plant kingdom. *Mol. Plant Breed.*, **6**, 1738–1757.
29. Boutanaev,A.M., Moses,T., Zi,J., Nelson,D.R., Mugford,S.T., Peters,R.J. and Osbourn,A. (2015) Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E81–E88.
30. Itkin,M., Davidovich-Rikanati,R., Cohen,S., Portnoy,V., Doron-Faigenboim,A., Oren,E., Freilich,S., Tzuri,G., Baranes,N., Shen,S. *et al.* (2016) The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii*. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E7619–E7628.
31. Zhou,Y., Ma,Y., Zeng,J., Duan,L., Xue,X., Wang,H., Lin,T., Liu,Z., Zeng,K., Zhong,Y. *et al.* (2016) Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat. Plants*, **2**, 16183.
32. Moses,T., Pollier,J., Shen,Q., Soetaert,S., Reed,J., Erffelinck,M.-L., Van Nieuwerburgh,F.C.W., Vanden Bossche,R., Osbourn,A., Thevelein,J.M. *et al.* (2015) OSC2 and CYP716A14v2 catalyze the biosynthesis of triterpenoids for the cuticle of aerial organs of *Artemisia annua*. *Plant Cell*, **27**, 286–301.
33. Miettinen,K., Pollier,J., Buyst,D., Arendt,P., Csuk,R., Sommerwerk,S., Moses,T., Mertens,J., Sonawane,P.D., Pauwels,L. *et al.* (2017) The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nat. Commun.*, **8**, 14153.
34. Dai,Z., Wang,B., Liu,Y., Shi,M., Wang,D., Zhang,X., Liu,T., Huang,L. and Zhang,X. (2014) Producing aglycons of ginsenosides in bakers' yeast. *Sci. Rep.*, **4**, 3698.
35. Jiang,Z., Kempinski,C., Bush,C.J., Nybo,S.E. and Chappell,J. (2016) Engineering triterpene and methylated triterpene production in plants provides biochemical and physiological insights into terpene metabolism. *Plant Physiol.*, **170**, 702–716.
36. Khan,N.E., Nybo,S.E., Chappell,J. and Curtis,W.R. (2015) Triterpene hydrocarbon production engineered into a metabolically versatile host—*Rhodobacter capsulatus*. *Biotechnol. Bioeng.*, **112**, 1523–1532.
37. Kirby,J., Romanini,D.W., Paradise,E.M. and Keasling,J.D. (2008) Engineering triterpene production in *Saccharomyces cerevisiae* – β-amyrin synthase from *Artemisia annua*. *FEBS J.*, **275**, 1852–1859.
38. Reed,J., Stephenson,M.J., Miettinen,K., Brouwer,B., Leveau,A., Brett,P., Goss,R.J.M., Goossens,A., O'Connell,M.A. and Osbourn,A. (2017) A translational synthetic biology platform for

rapid access to gram-scale quantities of novel drug-like molecules. *Metab. Eng.*, **42**, 185–193.
39. Zhuang,X. and Chappell,J. (2015) Building terpene production platforms in yeast. *Biotechnol. Bioeng.*, **112**, 1854–1864.
40. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241.
41. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem Inf. Model.*, **28**, 31–36.
42. Ertl,P., Patiny,L., Sander,T., Rufener,C. and Zasso,M. (2015) Wikipedia Chemical Structure Explorer: substructure and similarity searching of molecules from Wikipedia. *J. Cheminformatics*, **7**, 10.