Université de Montréal

Generation and screening of natural product-like compounds for antibiotic discovery

par

Samuel Jacques

Programmes de Biologie Moléculaire,
Faculté de Médecine

Mémoire présenté à la Faculté de Médecine
en vue de l'obtention du grade de Maître es Science (M.Sc.)
en Biologie Moléculaire, option générale

Avril 2016

Résumé

Avec l'apparition de plus en plus de souches de bactérie résistante aux antibiotiques, le développement de nouveaux antibiotiques est devenu une important problématique pour les agences de santé. C'est pour cela que la création de nouvelles plateformes pour accélérer la découverte de médicaments est devenu un besoin urgent. Dans les dernières décennies, la recherche était principalement orientée sur la modification de molécules préexistantes, la méta-analyse d'organismes produisant des molécules activent et l'analyse de librairies moléculaires pour trouver des molécules synthétiques activent, ce qui s'est avéré relativement inefficace. Notre but était donc de développer de nouvelles molécules avec des effets thérapeutiques de façon plus efficace à une fraction du prix et du temps comparé à ce qui se fait actuellement. Comme structure de base, nous avons utilisé des métabolites secondaires qui pouvaient altérer le fonctionnement des protéines ou l'interaction entre deux protéines. Pour générer ces molécules, j'ai concentré mes efforts sur les terpènes, une classe de métabolites secondaires qui possède un large éventail d'activités biologiques incluant des activités antibactériennes. Nous avons développé un système de chromosome artificiel de levure (YAC) qui permet à la fois l'assemblage directionnel et combinatoire de gènes qui permet la création de voies de biosynthèse artificielles. Comme preuve de concept, j'ai développé des YACs qui contiennent les gènes pour l'expression des enzymes impliquées dans la biosynthèse de la β-carotène et de l'albaflavenone et produit ces molécules avec un haut rendement. Finalement, Des YACs produits à partir de librairies de gènes ont permis de créer une grande diversité de molécules.

Mots Clés : découverte de médicaments, métabolite secondaire, biologie synthétique, terpène, chromosome artificiel de levure, synthèse de gènes

Abstract

With the appearance of more and more antibiotic resistant strains of bacteria, the development of new antibiotics becomes an issue of utmost importance for society. It is for that reason that new platforms and methodologies to accelerate the discovery of novel antibiotics are urgently needed. For the last decades, research was mainly oriented on modifying existing antibiotics, mining natural producers or screening for synthetic molecules from giant chemical libraries but these approaches did not manage to keep the pipelines filled with a sufficient number of novel antibiotics. Therefore, our goal was to develop a way to create and screen new molecules more efficiently at a fraction of the cost when compared to traditional approaches and within a short time frame. As chemical scaffolds we use natural product-like compounds that modulate the function of individual proteins or of protein-protein interactions. To generate these compounds, I focused first on the terpene scaffold class, a class containing molecules with a wide range of biological activities and includes compounds with antibacterial activities. We developed a yeast artificial chromosome (YAC) platform that allows both directional and combinatorial assembly of biosynthetic genes that can be used to create artificial biosynthetic pathways. As a proof of principle, YACs were successfully assembled containing genes coding for enzymes involved in the biosynthesis of both B-carotene and albaflavenone, and that allowed high yield production of these compounds. Finally, YACs encoding terpene gene libraries were also created and which produced a diversity of terpenoid molecules.

Keywords: drug discovery, natural product, synthetic biology, yeast artificial chromosome, gene synthesis, terpene.

Table of contents

List of Figures

List of Tables

List of Acronyms

GPCR: G protein-coupled receptor

FDA: Food and Drug Administration

UV: Ultraviolet

RiPPs: Ribosomally synthesized and post-translationally modified peptides

NRP: non-ribosomal peptide

NRPS: non-ribosomal peptide synthase

PKS: polyketide synthase

CDS: coding DNA sequence

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

Cas9: CRISPR associated protein 9

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

sgRNA: single guide RNA

Da: Dalton

bp: base pair

kb: kilobase

Mb: million base

pmol: picomole

ml: milliliter

μl: microliter

g: gravitational

YAC: yeast artificial chromosome

BAC: bacterial artificial chromosome

BMV: Brome mosaic virus

ARS: autonomously replicating sequence

MS: mass spectrometry

NMR: nuclear magnetic resonance

PCR: polymerase chain reaction

GC-MS: gas chromatography mass spectrometry

SICLOPPS: split-intein circular ligation of peptides and proteins

HIF-1: Hypoxia inducible factor-1

MDR: multi drug resistant

AMR: anti-microbial resistant

HTP: high-throughput

À mes parents, qui m'ont toujours supporté dans mes projets.

Acknowledgement

I would like to thank first, Mike Tyers who greeted me in his laboratory, gave me the chance to work on this great project during the last two years and allowed me a great freedom in the progress of my experiments.

I would like to also thank you all members of the laboratory for their advices throughout my master. A special thanks to my two supervisors, Almer van der Sloot and Susan Moore, who answered all my questions, helped me plan my experiments and provided comments on my thesis. A special thanks also to Samir Bouayad-Gervais for his help with all the chemistry experiments.

1. Introduction

1.1 Drug Discovery

The use of natural or synthetic compounds to treat specific medical conditions, such as infection or cancer, is the foundation of modern drug discovery. The use of drugs can be traced back as far as antiquity where medicinal plants were already used by Egyptians as described in the Ebers Papyrus that has 877 prescriptions and recipes to cure many kind of diseases like skin problems.[1] Even before that, it was shown by the analysis of coprolites (fossilized feces) that humans in the prehistoric era were likely consuming natural products as drugs in an attempt to treat disease and injuries.[2] It is only in the 19th century that breakthroughs in chemistry appeared that allowed isolation, purification, chemical characterization and dosage of active molecules from medicinal plants. A famous example is the isolation of morphine from opium poppy by F.W. Sertürner in 1806 which is associated with the launch of the first generation of drugs.[1,3] Afterward, the method of Sertürner was used to extract many other compounds. In 1826, Pierre Joseph Pelletier and Francois Magendie processed 150 tons of cinchona bark and succeed to purify 1800kg of pure quinine sulfate, an antimalarial drug.[1,4] It is considered to be the first natural product to be commercialized and their enterprise became the first pharmaceutical industry in history.[5] Soon after in Germany, Emmanuelle Merck launched the second pharmaceutical company, but instead of selling his natural products only in bulk, drugs were packaged in more convenient forms such as powders, syrups, pills, etc.[5] The next generation of drugs introduced many revolutionary medicines such as vaccines, serums and analgesics. It was during this period that one of the

most important conceptual frameworks in drug research was proposed, namely the receptor concept. Postulated by Paul Ehrlich, this theory posited that specific molecules will selectively bind to cells that have the right receptors and mediate a drug-induced biological response. This idea was further elaborated by Ehrlich to the concept of chemotherapy.[3,6] The third generation of drugs, known as the golden age of drug discovery, began around 1935. This generation introduced the concept of hormones, vitamins, sulfonamides and antibiotics. The discovery of antibiotics is due to Alexander Fleming who observed in 1928 that a *Penicillium* mold was active against staphylococcus bacteria.[1] It is only in 1938 that Howard Florey and his collaborator Ernst Chain showed this property was due to a metabolite produced by *Penicillium* and named it Penicillin. This compound can lyse staphylococci with relative little toxicity in humans.[3] This discovery paved the way for the use of microorganisms as source of drugs, initially for antibiotics. Many laboratories started to screen other soil microorganisms and by 1943, Selman Wacksman and Albert Schatz discovered the Streptomycin, an anti-tuberculosis drug produce by *Streptomyces griseus*.[7] Streptomycin inhibits protein synthesis by binding the 30S ribosome subunit of *Mycobacterium tuberculosis*.[8] By killing bacteria, antibiotics allowed highly effective infection control, preventing the death of millions of people and making all forms of surgery much safer. The fourth generation of drugs developed from 1960-1980 yielded, central nervous system agents (antidepressants, antipsychotics, anxiolytics), cardiovascular agents and semisynthetic antibiotics. The latter class of new antibiotics entailed modification of an existing natural product-derived scaffold by chemical synthesis.[5] The first fully synthetic antibiotics were also developed during this generation with the quinolone class, such as Ciprofloxacin and related analogs.[9] It was also during this era that pharmaceutical companies started to screen extremely large libraries of natural product and

synthetic molecules against potential targets to find active compounds. Cyclosporin A was discovered in the context of one of these programs.[3] The fifth and current generation is focused on specific enzymes or other proteins, such as kinases or proteases, often revealed by genome-based screens for targets in specific disease indications. Due to the requirement for specific mechansim of action, these molecules are often more complicated to develop into efficacious drugs.[5,10] Monoclonal antibodies, which are used in immunotherapy and other application, are also part of this fifth generation of drugs. Monoclonal antibodies can be developed to bind specific targets, such as a receptor on a cancer cell or a soluble cytokine, for activation of the immune system or delivery of conjugated toxins.[11] Throughout the latest era of drug discovery, the pharmaceutical industry has tended to move away from natural products in favor of synthetic compounds for screening campaigns, despite the proven success of natural products from plants and microorganisms as a source of drugs. In part, this shift was caused by duplication and deconvolution issues of natural product hits, resupply issues, time consuming and costly isolation, and synthetic intractability of initial hits.

1.2 Lack of Novel Drugs

The last decade has shown that number and ratio of small molecules in the development of pharmaceutical companies is decreasing. This trend has occurred despite the growing number of available potential targets provided by the human genome and the power of new sequencing technologies to sequence patient and pathogen genomes, advances in molecular biology techniques and other –omics technologies and sophisticated bioinformatics analysis. This lack of new drugs is of particular concern for the control of the spread of

antibiotic resistant strains. Examples include highly drug-resistant gram negative bacteria such as *Klebsiella pneumoniae*, *Escherichia coli* and *Acinetobacter baumanii,* multi drug resistant (MDR) *Mycobacterium tuberculosis* and Methicillin-resistant *Staphylococcus aureus* (MRSA). The development of new antibiotics is thus an issue of utmost importance for health agencies.[12,13] The emergence of anti-microbial resistant (AMR) strains and the current low rate of discovery of new antibiotics is now widely recognized in the media as a disaster in the making. As a consequence, in the near future simple operations may become life-threatening due to the risk of incurable infection. Part of the problem with many current combinatorial chemistry approaches and high-throughput (HTP) screens is the low chemical diversity, which is often shaped by convenience in chemical synthesis and/or the history of previous screening campaigns. New genome-derived targets can have many forms, for example myriad protein-protein interactions involved in signalling or molecular machineries (e.g. BCL2, MDM2, ribosome subunits), transcription factors (e.g. BCL6, RUNX) or receptors (e.g. Toll like receptors, GPCRs). High affinity and specificity of small molecule action is thus needed, as well as novel modes of action, such as the ability to efficiently target allosteric sites or shallow protein-protein interaction surfaces. A greater variety of scaffold classes is required to overcome these challenges. Given the diversity of potential targets, cost and efficiency of early stage drug discovery is a further major challenge. Many years of investment are required before a novel molecule is approved for clinical application and most molecules in fact fail at a late stage due to lack of efficacy or side effects. The cost of development for a new drug is now estimated at over 2 billion dollars.[14] New platforms and methodologies that mine novel large sources of chemical diversity inherent to natural products would help mitigate this enormous cost.

1.3 Importance of Natural products

Natural products can be differently defined according to the field of research, but in the context of this thesis I will define natural products to encompass any secondary metabolite produced by any organism.[15] These secondary metabolites are produced by both prokaryotes and eukaryotes species, but those used in the human pharmaceutical arsenal are usually provided by bacteria, fungi and plants. These secondary metabolites are not required for normal metabolism, growth, development and reproduction but instead provide a survival advantage to the host. Typically natural products are involved in defense mechanisms and have often cytotoxic properties, which can be exploited for therapeutic purposes.[16] Moreover, as natural products are the result of million years of natural selection for optimized binding to specific biological targets, natural products are intrinsically enriched in scaffolds with drug-like properties, such as solubility and membrane permeability. Finally, natural products represent many structural and biosynthetic classes, each of which can contain many thousands of unique compounds. This diversity and bioactivity ideally positions natural products as productive lead scaffolds for the development of novels drugs. Indeed, even today, more than 50% of current drugs on the market are derived from or inspired by natural products.[17]

1.4 Natural product classes

As alluded to above, natural products represent an impressive number of molecules, which can be classified according to the precursors and biosynthetic pathways they originate from. Many classes exist, but only those that our laboratory is currently focused on will be

described here. Special attention will be given to the terpenes in the next section as this class is the focus of my thesis (Fig. 1).
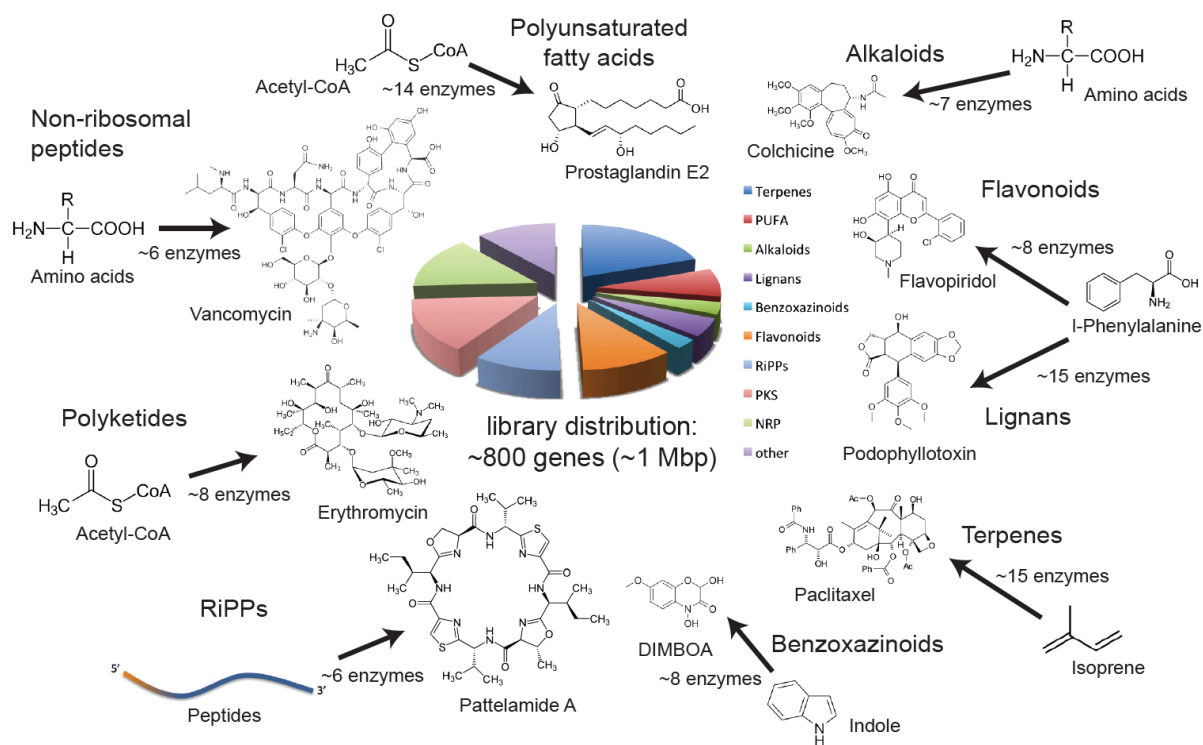
Figure 1. Synthetic genes for natural product biosynthesis

Alkaloids are small natural products with over 12,000 molecules identified to date, mostly produced by plants. Recent studies also indicate that marine invertebrates might be a potential important source of alkaloids.[18] Alkaloids are derived from amino acids and typically contain one or more nitrogen atoms.[19] Historically, the beneficial effects of traditional medicines were due to the presence of alkaloids.[20] Some alkaloids, such as ergot alkaloids are

6

toxic at high doses, but when used at a proper dose can have strong medicinal effects[21] Alkaloids have a wide range of pharmacological properties that include analgesics (codeine), central nervous depressants (morphine), antihypotensives (ephedrine), antitumor activities (vinblastine) and antimalarial activities (quinine).[19,22] Many alkaloids are also widely consumed as recreational drugs like caffeine or nicotine. Despite their great diversity, alkaloids share many points in common: they are basic because of their nitrogen with an unshared pair of electrons, when purified they are often solids (except those that lack oxygen, such as nicotine, which are liquids) and they are able to bind to a wide array of proteins, such as enzymes, receptors or structural scaffold proteins.[18]

Flavonoids are secondary metabolites produced by plants and fungi and currently number more than 6000 documented compounds.[23] Flavonoids have many biological roles in plants such as structural integrity, UV photoprotection, reproduction and internal regulation of plant cell physiology and signaling.[24] In addition, flavonoids are often the pigments that give the color to many flowers and fruits. The precursor to all flavonoids, 4-coumaroyl-CoA, is supplied by phenylalanine through the phenylpropanoid pathway.[23] Afterward, 4-coumaroyl-CoA enters the flavonoid pathway, which is divided into many sub-groups. Flavonoids in plants are always found as α or β glycoside forms because glycosylation increases their solubility and stability.[24] Despite their biological activities, only few flavonoids, are currently known to have therapeutic effects. Epidemiological and medical data, however, suggest potential roles in the prevention of diseases such as cancer, diabetes and cardiovascular conditions.[24] Plus, it was shown that flavonoids can act as adjuvants towards antibiotic

resistant strains, whether through direct inhibition of resistance determinants or through other mechanisms.[25]

Nonribosomal peptides (NRPs) represent a wide range of biological activities, for example immunosuppressants (cyclosporine A), antibiotics (daptomycin) and anticancer agents (bleomycin A2).[26] NRPs are special type of small proteinogenic and non-proteinogenic peptides that are assembled by large multifunctional enzymes called nonribosomal peptide synthases (NRPS), and are thus independent from ribosome-mediated translation of genetically encoded peptides.[27] The NRPs are therefore advantageous in terms of chemical diversity as they are not limited to the 20 proteinogenic amino acids. Over 500 monomers are known to be used by NRPSs which include non-proteinogenic amino acids, fatty acids, and α-hydroxy acids.[28] This is possible because NRPSs use modules to assemble peptidic scaffolds. Each incorporation of a new amino acid or other building block requires a dedicated module, such that NRPSs are typically extremely large enzymes. For example cyclosporine A is assembled by a NRPS of 1.6 MDa. Each NRPS module is composed of 2 or more domains.[27] The first module, known as the initiation module, has an adenylation domain (A) which is responsible for activation of the amino acid and a thiolation domain (T), also known as peptidyl carrier protein domain (PCP), which transfers the growing peptide to the next module. Every other module also has an upstream condensation domain (C) that incorporates a new amino acid to the peptide. The last module in the assembly line has an additional domain called the thioesterase domain (TE), which can either hydrolyse the linear peptide or catalyze cyclisation. Some modules also have specialized domains, such as epimerisation (E),

methylation (MT) or cyclization (Cy) domains, which incorporate specific modifications in to the amino acid recruited by the C domain.[26,27,28]

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important new class of natural products. In contrast to NRPs, the RiPP scaffold is genetically encoded and thus limited to the canonical 20 proteinogenic amino acids. One advantage of the RiPPs is that their chemical structure can be easily manipulated by changing the codon sequence in the corresponding precursor gene. RiPP length can vary widely, from ~12 to 110 amino acids.[29] Many organisms produce RPs, including bacteria, plants, fungi and some metazoans.[30] Known RiPPs show a diverse array of potential therapeutic activities, such as antibacterial, antitumor, anti-HIV, and antimalarial action. Almost all ribosomal peptides are synthetized in the same manner. The initial structure is a genetically encoded precursor peptide that contains a leader peptide, a core peptide and a recognition sequence.[29] Subsequently, this precursor peptide is processed and heavily post-translationally modified. The leader peptide, normally at the N-terminal of the peptide, serves for the recognition of the peptide by specific post-transcriptional modification enzymes and, in some cases, serves for the export of the peptide. The sequence that will become the natural product is the core peptide and is cleaved from precursor peptide. Finally, the recognition sequence, found at the C-terminal, is involved in the excision and, in some classes, cyclisation of the mature peptide.

Polyketides are reduced natural products with structurally intriguing carbon skeletons which can have numerous stereocenters and exceeding 500 Da in mass.[31,32] They are made of repeated acyl and malonyl building blocks supplied by acetyl-CoA and malonyl-CoA.[33] Like NRPs, polyketides are assembled by huge multimodular enzymes called polyketides synthases (PKSs) which assemble monomers by Claisen condensation reaction.[34] PKSs modules are also made of different domains to allow the extension of the growing polyketides.[33,34] Three mains domains are needed: an acyltransferase (AT), which recruits the appropriate building block, an acyl carrier protein (ACP), which holds the fragment recruited by AT and transfers the growing polyketide to the next module and a ketosynthase (KS) which joins the growing polyketide to the new fragment. Furthermore, this condensation can be followed by other optional modifications carried out by ketoreductase (KR), dehydratase (DH) and enoylreductase (ER) domains. At the end, a thioesterase (TE) domain liberates the polyketide by the cleavage of the thioester link between the polyketide and the ACP and (often) followed by a macrocylisation.[33,34] Polyketides have a large array of therapeutic properties including antibiotic (erythromycin), antifungal (amphotericin), immunosuppressant (rapamycin), antiparasitic (avermectin) and antitumoral (epothilone).[33]

## 1.4.1 Terpenes

Terpenes, also known as isoprenoids, are the most diverse class of natural products with over than 50 000 compounds found in almost every forms of life.[35,36] Terpenes can be classified depending on their number of isoprene units or their number of carbons, as follows: hemiterpene (C_5, 1 unit), monoterpene (C_{10}, 2 units), sesquiterpenes (C_{15}, 3 units), diterpene

(C_{20}, 4 units), sesterterpene (C_{25}, 5 units), triterpene (C_{30}, 6 units), sesquarterpene (C_{35}, 7 units) and tetraterpene (C_{40}, 8 units) (Fig. 2).[35,37] Isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) are the two precursors which all terpenes are derived from and they can originate from two different pathways: the mevalonate or the methylerythritol phosphate (MEP) pathway.[36,38] The mevalonate pathway is prevalent in eukaryotes and archea and consists of six enzymes that convert Acetyl-CoA to IPP. Afterward, an IPP isomerase is needed for two functions: producing DMAPP from IPP and keeping a balance between IPP and DMAPP. While the MEP pathway is privileged by prokaryotes and transforms the molecules glyceraldehyde-3-phosphate and pyruvic acid to IPP and DMAPP in a ratio 5:1. Afterward, IPP and DMAPP can be condensed by a prenyltransferases geranyl diphosphate synthase (GPPS) into geranyl diphosphate (GPP) which is the precursor of monoterpenes or further transformed into farnesyl diphosphate (FPP) by a farnesyl diphosphate synthase (FPPS), the precursor to sesquiterpenes and triterpenes. Finally, FPP can also be converted into geranylgeranyl diphosphate (GGPP) by geranylgeranyl diphosphate synthase (GGPPS), the precursor of the diterpenes. For the other higher order terpene classes, it is a mix of these precursors. For example, the production of sesquarterpenes requires one molecule of FPP and 4 molecules of IPP to create its backbone. The specific terpene backbones, regardless the terpene class, are created by terpene synthases, the first committed step in the terpene pathway, and subsequently modified and diversified by cytochrome P450s, to give rise to the huge number of terpenes family that are presently known. Terpene synthases are converting the precursors (ex. GPP) into their scaffolds (ex. monoterpene backbone), while the cytochromes P450 (P450s), the most important terpene-modifying enzyme class, can catalyse reactions such as insertion of oxygen on the backbone.[36,39] The development of this

11

diversity over the evolution allowed their producers to use these metabolites for many functions. In plants, terpenes can be used as primary metabolites (sterols and hormones) or can have a more specific role like acting as a defense mechanisms or attracting insects for pollination.[40] Humans took advantage of all these properties to create compounds used in the daily life like fragrances (ex. citronellol), cosmetics (ex. Astaxanthin), dietary supplements (ex. Lupeol), agrichemicals (ex. Gibberellins), pesticides (ex. Germacrene D1) flavors (ex. Limonene) and, of course, as pharmaceuticals (ex. paclitaxel) molecules.[36,41] Since many terpenes have pharmacological activity or other valuable properties, companies show an interest not only in purifying terpenes from their natural sources, but engineering new heterologous organisms to produce a higher yield or provide a more reliable source than the natural producer or to modify existing molecules to produce novel effects. With the new sequencing technologies available and the creation of powerful algorithms, many terpene pathways were discovered in the last decade and gave the opportunity to recreate synthetic pathways in a heterologous hosts like *E. coli* or *S. cerevisiae*. A successful example is the production of artimisinic acid in yeast. Arimisinic acid is a precursor of the antimalarial drug artemisin, which is normally extracted from *Artemisia annua L* (Sweet wormwood), but now has been engineered to be produced by yeast with a yield of 25g/L.[42,43] Another example is taxadiene, a precursor to taxol also known under the commercial name paclitaxel, a successful anti-cancer drug produced by *Taxus brevifolia* (Pacific yew).[44] Due to its complexity to synthesize, its elevated cost to isolate and the scarcity of the source species, taxol is become a subject of interest and many laboratories are trying to elucidate its complete pathway.[45] But until now, only taxadiene has been produced in *E. coli and yeast* and there are still many steps to elucidate the complete pathway.[46,47]
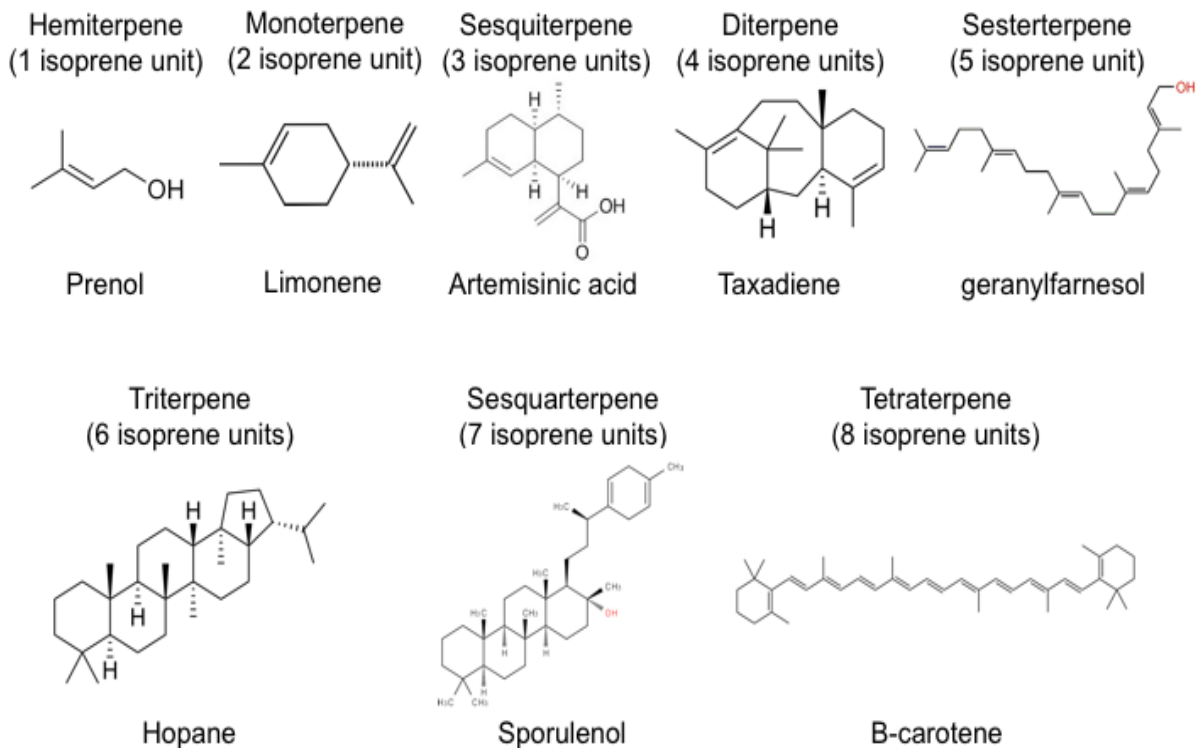
Figure 2. Examples of terpene from different sub-classes.

1.5 Synthetic biology

Synthetic biology is a recent interdisciplinary branch of science in which the principles of biology and engineering are drawn together. This emerging field is focused in designing and engineering artificial biological systems and organisms to process information and generate new outputs or biosynthesized products. In the 1990s, before synthetic biology was defined, biologists were already performing metabolic engineering, which is the optimized production of a desired compound by the integration of a specific pathway in the original host or in a heterologous host. A successful example, is the overproduction of lysine in *C.*

glutamicum, where the flux of glucose has been deviated from its original path for maximal production of lysine.[48] With the recent decrease in gene synthesis cost and the advent of new gene assembly techniques, it has become much simpler to assemble many DNA parts (promoter, terminator and CDS) into synthetic modules and to introduce these into heterologous organisms. Synthetic biology also allows the decoupling of production from normal host regulation, for example by removing negative feedback or increasing metabolic fluxes for higher yield. A host organism exploited for millennia by humans is the budding yeast *S. cerevisiae*, which is traditionally used to bake bread and brew beer. *S. cerevisiae* has since become an important "chassis" for bioproduction because it is a unicellular eukaryote that can be grown rapidly in a controlled manner on minimal feedstocks at massive scale.[49] *S. cerevisiae* has been used for the production of highly complex natural products, perhaps most notably the anti-malarial agent artemisinin.

The Gibson assembly method and the CRISPR-Cas9 system have recently accelerated the ability to produce synthetic circuits and genetically engineered cells and organisms. Gibson assembly, so named for its inventor Daniel Gibson, allows the sequence independent assemble of many DNA fragments at the same time, unlike conventional restriction enzyme-based methods.[50] The clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated protein 9 (Cas9) system was originally discovered as an RNA-guided DNA endonuclease enzyme system used by the adaptive immunity system of certain bacteria, such as *Streptococcus pyogenes*, to cleave foreign DNA. The CRISPR/Cas9 system has been recently re-engineered into a powerful tool that can target and cut almost any DNA fragment

by only changing the sequence of its target specific guide RNA (sgRNA).[51,52] The CRISPR/Cas9 system thus allows precise genome-level alterations to be made, including deletions, insertions, and point mutations.[53] Despite these impressive technological breakthroughs, the modularization and standardization of biological parts envisioned by engineers remains a major challenge. The biological parts may be promoters, terminators, CDS or other regulatory elements, and can be conceptualized building blocks analogous to transistors, resistors or capacitors in electrical engineering.[54] Parts from different categories and organisms can be readily combined in the hope of building novel circuits or devices with specific utility. However, because the individual properties of proteins and nucleic acids are far more complex than the reiterated components used in electronics, introducing new parts in a cell usually produces unpredictable interactions that result for example in expression failure, cell death or other unwanted characteristics. To fully unleash the promise of synthetic biology, a better understanding of interplay between cellular components, such as interactions between proteins or metabolites with proteins is required. For instance, novel methods need to be developed to eliminate or compartmentalize toxic compounds that can be produced or to prevent non-desired interactions that can disturb normal metabolism of the cell. Despite these challenges, some notable achievements have already been made. One of the premiere examples of synthetic biology is the assembly of the first synthetic genome by the J. Craig Venter Institute in 2010.[55] This project entailed assembling cassettes of 1,080bp into successively larger and larger fragments to recreate the 1,077,947bp genome of *Mycoplasma mycoides*. Critically the function of the synthetic genome was proven by replacing the genome of *M. capricolum* cells with the synthetic genome of *M. mycoides,* which effectively 'rebooted' the host cells with a new genome. These new synthetically encoded cells had the

phenotypes of *M. mycoides* and were able to self-replicate. Aside from technical results of academic interest, synthetic biology has strong potential to revolutionize biotechnology. One such example is the conversion of carbon dioxide to biofuel (2,3-butanediol) by an engineered strain of *Synechococcus elongatus,* a cyanobacteria, where exogenous enzymes were integrated in the genome to produce compound that was not produced naturally by this organism.[56] Other successful examples include the commercial production of artemisinic acid (see above) and vanillin in yeast.[57]

1.5.1 Development of novel chemical matter using synthetic biology

The synthetic engineering of natural products, endowed by million years of natural selection to bind specific targets with physiological importance, hold tremendous potential for the creation of new therapeutic molecules. Natural product biosynthesis pathways from specific organisms can be transplanted into new host such as *E. coli* or *S. cerevisiae* to produce valuable compounds on an industrial scale. Moreover, genes from different natural product pathway and different organisms may be recombined to either create novel natural product-like compounds.[58,59] Genetic selection is a powerful method to identify molecules that confer a survival advantage to the host cell when interacting with a target reporter of interest. For example, selection can be applied by disrupting or inhibiting a target protein function or interaction that is toxic to the host cell. This principle has been shown with a genetically encoded cyclic peptide system based on split-intein circular ligation of peptides and proteins (SICLOPPS) technology.[60] The SICLOPPS system enabled discovery of cyclic peptides that reduced toxicity of alpha-synuclein aggregation in yeast or inhibited inhibit the Hypoxia

inducible factor-1 (HIF-1) heterodimerization.[61] Selection strategies have also been used in combination with novel natural product like compounds derived from a small library of natural product pathway genes expressed in yeast. In one instance, genes from a specific class of natural products were randomly integrated into a yeast artificial chromosome (YAC). A Brome Mosaic Virus (BMV) replication assays was used to screen the library, such that only the cells that produced molecules that inhibited BMV replication were able to survive. A number of novel compounds produced by these selected cells were isolated and analysed by mass spectrometry (MS) and nuclear magnetic resonance (NMR). This approach yielded novel scaffolds with potential activities, but medicinal chemistry optimisation was still needed.[58,59]

1.6 Yeast artificial chromosome technology

YACs are genetically engineered chromosomes that can be stably propagated in *S. cerevisiae*. The basic structural elements of YACs required to mimic a natural chromosome are: (i) a centromere for segregation during mitosis, (ii) telomeres for chromosome end maintenance and (iii) an autonomously replicating sequence (ARS) for initiation of DNA replication.[62] Selectable markers, such as antibiotic resistance or auxotrophic markers, allow YACs to be stably maintained over many generations. Circularized YACs that contain a bacterial origin of replication and a bacterial selection marker can be propagated in *E. coli* as well as in *S. cerevisiae*. Cleavage of the circular YAC into a linear form allows the liberation of both of its telomeres and function as a chromosome in yeast. The telomere containing YAC arms can also be built as separate fragments on different YAC arms. When introduced into cells, YAC arms can be ligated with other DNA modules (genes, promoter, terminators) that bear

complementary ends. The homologous recombination system of *S. cerevisiae* recombines all introduced DNA fragments in the proper order, as dictated by sequence overlaps. Large fragments of DNA, on the order of 1Mb or more, can be inserted into YACs to construct complex genetic circuits. The large genetic payload capacity of the YAC avoids the need for many different plasmid constructions and selection markers. Although YACs can have potential drawbacks, such as the potential for deletions, rearrangements and chimerism.[63] These issues can be mitigated by careful design of the YAC, for example by reducing homologous stretches of DNA between modules. YACs also provide more flexibility when compared to other methods of incorporating recombinant DNA, such as chromosomal integration. In directed integration strategies, a neutral integration site must be found for every fragment incorporated, which is not practical for dozens of fragments. Random integration strategies, such as insertion into yeast Ty elements are less optimal because of inefficient expression and/or interference with the expression of endogenous genes.[64] In summary, precisely engineered YAC systems can be readily optimized for the combination and expression of many diverse genes, as required for the research described in this thesis.

2. Research project

The aim of this thesis was to implement a YAC platform for combinatorial assembly of biosynthetic genes under the control of inducible promoters, with the goal of creating artificial biosynthetic pathways that synthesize a wide variety of novel natural product-like compounds against specific targets. For this purpose, our laboratory custom synthesized a large library of genes involved in the production of a variety of natural product classes (see above). All sequences were optimised for GC content and codon bias to maximize expression in yeast. As a proof of principle, I assembled a YAC containing genes coding for enzymes involved in the biosynthesis of β-carotene with the goal of high yield production. β-carotene was chosen because it only requires three enzymes and its biosynthesis is easy to monitor due to the orange color development in producer cells. I also assembled a second YAC with the genes that code for the enzymes involved in the biosynthesis of the sesquiterpene albaflavenone.[65] This compound has known antibiotic properties and serves as as a positive control for pathogen co-culture experiments to screen for novel antibiotics. Finally, I created a library of terpenes to create entirely novel natural product-like scaffolds. These initial steps have set the stage for application of the synthetic natural product platform in selections screens and co-culture screens for the identification of new anti-mycobacterial compounds active against *Mycobacterium tuberculosis* and other bacterial pathogens.

3. Materials and Methods

3.1 Yeast strain

The haploid *Saccharomyces cerevisiae* yeast strain used for the construction pYAC7 and derivative YACs was BY4742 (ATCC® 201389™) with the following genotype: *MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*. The haploid yeast strain used for the construction pYAC8 was VL6-48 (ATCC® MYA-3666™) with the following genotype: *MATα his3Δ200 trplΔ1 ura3-52 ade2-101 lys2 psi+cir°*.

3.2 Construction of pYAC9

The initial vector that supplies the long and short YAC arms, which contain the telomeres, the centromere, the ARS and the selectable markers, was called pYAC5 and is a modified version of pYAC4 (ATCC® 67379™) with two additional restriction sites, AscI and NotI, introduced in the multiple cloning site (MCS). BY4742 was transformed with pYAC5 and cells were grown on a SC, -His, -Ura plate at 30°C for 48 hours to select the cells with the plasmid. The URA3 gene was replaced by homologous recombination by the LEU2 gene amplified from the plasmid pRS425 (ATCC® 77106™) with the primers YAC0077_URAtoLEU_FW and YAC0078_URAtoLEU_RV (see Table S1 for sequences), which contain homologous sequences with the regions flanked on each side of the URA3 gene in pYAC5, to give pYAC6. Cells were grown on a SC, -Leu, -His plate at 30°C for 48 hours to select the cells with the marker swap. Afterward, the HIS3 gene of pYAC6 was replaced by

URA3 gene amplified from the plasmid pYAC5 with the primers YAC0083_HIStoURA_2_FW and YAC0084_HIStoURA_2_RV then inserted in the cells to give pYAC7. Cells were grown on a SC, -Leu, -Ura plate at 30°C for 48 hours to select the cells with the modification. A culture of BY4742 with pYAC7 was prepared by transferring a colony into 5ml of SC, -Leu, -Ura and incubating the culture at 30°C and 250rpm, overnight. pYAC7 was isolated by performing genomic DNA extraction from the culture, and *E. coli* cells transformed with this genomic DNA.[66] Bacterial transformants were grown on a LB plate with ampicillin (100µg/mL) at 37°C. A culture of *E. coli* with pYAC7 was prepared by transferring a colony into 5ml of LB with ampicillin and incubating the culture at 37°C and 250 rpm, overnight. A plasmid miniprep preparation was performed from 5 mL *E. coli* culture to purify pYAC7, using a miniprep DNA column (EZ-10 Spin Column plasmid DNA Miniprep, Biobasic) according to the manufacturer's instructions. Finally, VL6-48 cells were transformed with pYAC7 and cells were grown on a SC, -Leu, -Ura plate at 30°C for 48 hours to select the cells with the plasmid. The HIS3 gene was amplified from the plasmid pYAC5 with the primers YAC0142_TrptoHIS_2_FW and YAC0143_TrptoHis_2_RV then inserted into the cells to replace the TRP1 gene which gave pYAC8. Cells were grown on a SC, -His, -Ura plate at 30°C for 48 hours to select the cells with the modification. A culture of Vl6-48 with pYAC8 was prepared by transferring a colony into 5ml of SC, -Leu, -Ura and incubating the culture at 30°C and 250 rpm, overnight. Subsequently, pYAC8 was isolated by performing an extraction of genomic DNA from the culture and *E. coli* cells were transformed with this genomic DNA. Cells were grown on a LB plate with ampicillin (100µg/mL) at 37°C. A plasmid preparation was performed from the *E. coli* culture to purify pYAC8. Two restriction sites (XbaI and PvuII) were subsequently added at the gene insertion site to give pYAC9.

3.3 Creation of entry modules

The DNA sequences of all genes in the natural product biosynthesis gene library were codon optimised by an in-house algorithm (J. Coulombe-Huntington, unpublished) and synthesized by Gen9 using BioFab® technology (Gen9 Inc, Boston, MA, USA). Adapter sequences to aid isothermal assembly flanked the 5' and 3' end of each gene and allowed both PCR amplification with the same primer pair (pri0561 and pri0562) and directional assembly between a promoter and terminator in any of the 24 pASS vectors for modular assembly into the YAC. Every gene amplification was performed with the same conditions (Table S2). pASS vectors needed to be opened prior the insertion of a gene, which was achieved by PCR amplification. The 24 vectors had the same forward primer (pri0643) and according to the specific promoter in each vector the reverse primer differed as follows: pASS 1 to 6 (pri0642), pASS 7 to 12 (YAC0213_pASS7-12_RV), pASS 13 to 18 (YAC0214_pASS13-18_RV) and pASS 19 to 24 (YAC0215_pASS19-24_RV) (see Table S1 for sequences). The amplification of the backbones was performed under the same conditions as for the genes above (Table S2). For each position in the YAC, 0.2 pmol of every selected gene was mixed with 0.4 pmol of open backbone. Gibson assembly was used to create entry modules for every position in the YAC. For every entry module, an amplification by PCR was performed on the Gibson assembly reaction sm577_panASS_FW and sm580_panASS_RV primer pair to release the entry modules. These entry module PCR products were used in subsequent YAC assemblies described below.

3.4 YAC assembly

pYAC9 was digested by BamHI-HF, XBaI and PvuII (Table1) for 1 hour at 37°C in CutSmart buffer (NEB B7204S) to linearize and liberate the YAC arms from the the bacterial resistance marker and origin and URA3 marker. Long and short YAC arms were isolated by separating the digested pYAC9 sample on a 1% agarose gel and purifying the arms (~6500bp and ~4150bp, respectively) using a DNA spin column gel extraction kit (EZ-10 Spin Column DNA Gel Extraction Kit, Biobasic) according to the manufacturer instructions. For every pmol of arms, 2 to 4 pmol of entry modules for each of the 24 positions of the YAC were mixed together. For positions that did not include a gene, empty entry modules were used to bridge the YAC slot. In control assembly mixes, entry modules at random position were left out to monitor unspecific homologous recombination. The pooled YAC fragments were transformed into BY4742 with a modified version of the large-scale high-efficiency yeast transformation method using the LiAc/SS carrier DNA/PEG method as described by Gietz and al.[67] The transformation mix for each reaction was 240µl of PEG 3500 50% (w/v), 18µl LiAC 2.0M, 50µl salmon sperm DNA (2mg/ml in TE), 36µl DTT 1.0M. After mixing cells with DNA, heat shock was performed for 40 minutes at 42°C. Cells were then re-suspended into 1ml of SD, -His, -Leu and incubated for 45 min at 30°C. Serial 10-fold dilutions of the cells were prepared (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}) in a volume of 100µl and then plated on SD, -His, -Leu and incubated at 30°C for two days.

Table 1: pYAC digestion conditions

Components	Volume (μl)
pYAC9	sufficient volume to reach 10μg
5x Cutsmart Buffer (NEB)	5
BamHI-HF (20U/μl) (NEB)	2
XBaI (20U/μl) (NEB)	2
PvuII-HF (20U/μl) (NEB)	2
H_2O	Add to 50

3.6 Isolation and identification of terpenes

Ten of the terpene YAC library colonies that grew on the on SD, -His, -Leu plate were picked and grown in 5ml of SD, -His, -Leu, -Met at 30^oC for 24 hours to induce transcription of modules encoded in the YAC. Yeast clones that had a fresh earth-like smell were selected for culture in 500ml of SC, -His, -Leu, -Met at 30C for 48 hours at 250 rpm. Yeast clones that contained a YAC that did not contain biosynthetic genes served as a negative control. Cultures were pelleted by centrifuging at 10 000g for 10min (4°C) and the pellets were washed with 30ml of PBS and pelleted at 10 000g for 10min (4°C). This wash step was repeated 3 times. The pellets were resuspended in 20ml of PBS and the cells were lysed by passing three times through a high pressure homogenizer (EmulsiFlex-C3) at ~21 000kPa. Two extractions were performed by adding 10ml of dichloromethane (DCM) to the lysed cells, followed by 3 min of vortexing. Lysed cells were pelleted at 21 000g for 3min. Afterward, the organic phase (top

layer) was transferred to a round-bottom flask. The combined organic layers were dried using Na_2SO_4, filtered by passing through Celite, and concentrated by a rotary evaporator (<5 Torr, 40-50°C). The residue was washed twice with 500µl of DCM and transferred to a 2ml vial and concentrated by evaporation under a stream of N2. The final residue was re-dissolved in 50ul of DCM and 0.5µl was analyzed by GC-MS.

3.7 GC-MS analysis

Electron impact (EI) mass spectra were obtained using a 7000C Triple Quadrupole GC/MS System from Agilent Technologies in the EI positive ion full-scan mode with ion source at 250 °C and with energy of ionization of 70 electron volts. GC separations were performed on a fused silica HP5 (25-m, 0.32-mm inside diameter) capillary column with a linear temperature program (70 °C hold for 4 min, to 300 °C at 10 °C/min, 300 °C hold for 2 min).[65] 0.5µl was injected.

3.8 mCherry expression Analysis

Thirty of the colonies from the albaflavenone YAC assembly that grew on the on SD, -His, -Leu selection plate were picked and grown in a 5ml of SD, -His, -Leu, -Met, at 30°C for 24 hours, which induced the transcription of genes contain on the YAC, including an mCherry reporter gene. 100µl of each culture were transferred to a 96 wells plate and analysed for mCherry expression with a fluorescent plate reader (Infinite® M1000 PRO plate reader, Tecan) (excitation wavelength: 587 nm, Emission wavelength: 610 nm).

4. Results

4.1 YAC platform, assembly and transformation

The YAC assembly method was adapted from a previously published method by Naesby et al., but instead of a random assembly of constituent parts, the platform I used allows directional and combinatorial assembly of the constituent parts in a more streamlined fashion.[59] The platform includes 24 different parts (termed entry modules) that are flanked by 100 bp homology arms that allow directional homologous recombination-mediated assembly (Fig. 3). Each module has a unique combination of homology arms that allow the modules to occupy only one specific position in the YAC. The homologous recombination system of yeast will recognize and recombine modules that share homologous sequences. For each position, a mix of entry modules that carry different genes can be created to allow position-specific combinatorial assemblies. Genes coding for enzymes involve in different natural product pathways are inserted by Gibson assembly in one or more of the 24 entry modules using a standardized assembly patches. The gene(s) in each module are placed under the control of promoter and terminator elements for regulated expression in yeast. Every entry module has a different combination of a promoter and terminator. For the system I used four different methionine-dependent promoters (Met2 and Met25) from different *S. cerevisiae* species and six different terminators are available.[59] In addition to these entry modules, the two YAC arms are required to complete the assembly. These YAC arms contain the telomeres, the centromere, the ARS for stable replication and yeast auxotrophic markers for YAC selection and maintenance. The YAC arms were derived from the pYAC9 plasmid, a modified version

26

of pYAC4, where restriction sites were added to the MCS and auxotrophic markers changed to meet requirement for future screening experiments (Fig. S1). A mix of entry modules and YAC arms DNA was introduced into *S. cerevisiae* (BY4742) by use of the lithium acetate protocol and following transformation, the homologous recombination system assembles the modules into a complete YAC. The auxotrophic markers on each arm were used to select the cells that assembled a functional YAC. The fluorescent protein (FP) mCherry was incorporated into one of the entry modules to monitor false positive assemblies. The presence of genes in each module was confirmed by colony PCR (data not shown). Since many terpenes are colored , in some cases visual analysis of the yeast clone was sufficient for confirmation of successful YAC assembly.
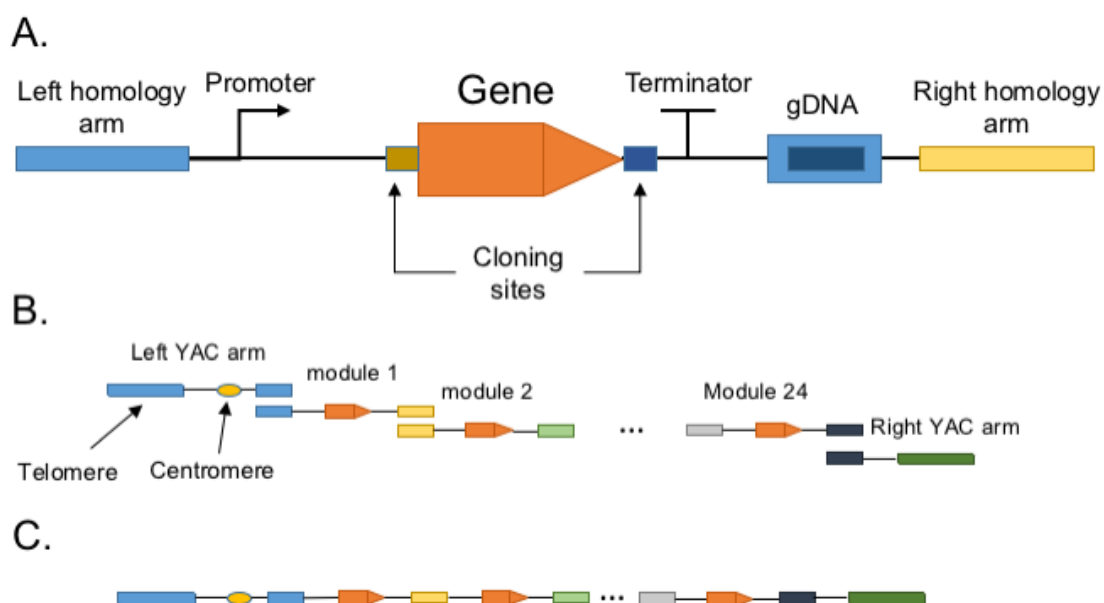
Figure 3. YAC platform for combinatorial assembly of natural product biosynthesis enzymes. (A) Representation of an entry module with all its elements. (B) Directional assembly of the modules and YAC arms by the homologous recombination system of *S. cerevisiae*. (C) Representation of a fully assembled YAC.

4.2 Reconstitution of the β-carotene pathway

Genes coding for the enzymes in the β-carotene pathway were used for an initial construction of a test YAC (fig. S2). Eight copies of each gene were used to construct the YACs to maximise the production of β-carotene: the first eight modules encoded CrtYB, the next eight modules CrtI and last eight modules CrtE. Cells transformed with the YAC arms only served as a negative control to verify the absence of ligation between the two arms, which would otherwise give a false positive assembly and severely dilute the library pool. A plasmid called pHV004 (kindly provided by Patrick Cai, University of Edinburgh), that contained the genes for β-carotene production under control of the *TDH1* promoter was used as positive control to monitor the colony color. After 48 hours, no colonies were observed on the plate with cells that were transformed with only the YAC arms, indicating that arms did not recombine together to give false positives. When observing cells that contained pHV004, an orange color was obvious, showing that β-carotene was produced. When observing the cells from the full YAC assembly, half of the colonies appeared orange. When compared with wild type BY4742 yeast, it was easy to distinguish between cells that produced β-carotene and that do not due to the orange color of the former (Fig. 4). Cells transformed with 1μg of pHV004 led to ~$3x10^6$ transformants, while cells transformed with 200ng of YAC arms (and 100ng of each entry module) resulted $7x10^5$ transformants per 1μg of YAC arm DNA. This result demonstrates that YAC assembly *in vivo* from 26 different parts proceeded with high efficiency, as required. GC-MS analysis was attempted on an extract of β-carotene to confirm analytically the presence of the compound, but the boiling temperature of β-carotene (654.7

°C) was too high for the instrument used, which was only able to reach 300°C (data not shown).
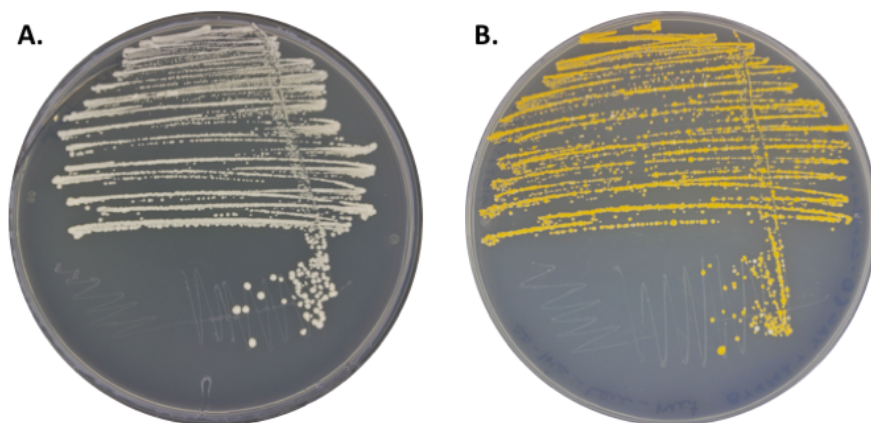
Figure 4. Comparison between wild type yeast and yeast expressing the β-carotene biosynthesis pathway. (A) *S. cerevisia* **BY4742 wild type cells. (B)** *S. cerevisiae* **BY4742 cells transformed with a YAC that contain the genes for the biosynthesis of β-carotene.**

4.3 Reconstitution of the albaflavenone pathway

The two genes, *sco5223* and *sco5222*, coding for the enzymes in the albaflavenone pathway were used in the construction of a YAC (Fig. S3). Again many copies of each gene, whereby the first six modules carried a copy of *sco5222* and the next six carried a copy of *sco223*, were used to maximise the production of albaflavenone. For this experiment, a gene module that codes for fluorescent protein mCherry was used to monitor complete assembly of the YAC. Analysis of 30 colonies showed that 12 of the clones had mCherry expression. To verify the presence of albaflavenone in these colonies, the same extraction method as described for β-carotene was used for albaflavenone and extracts were analysed by GC-MS.

29

These results were compared with those from literature, where albaflavenone production was analysed from the original bacterial producer strain *Streptomyces coelicolor.*[65] The fragmentation spectrum for the molecule with a retention time of 8.35 minutes (43.1, 57.1, 71.1, 85.1 and 99.1 pattern) was in agreement with fragmentation spectrum of albaflavenone from literature, thereby confirming the production of albaflavenone by the YAC-encoded enzymes (Fig. 5).[65] The presence of three biosynthetic intermediates (epi-isozizaene and 4S,4R-albaflavenol; retention times of 6.8min, 7.9min and 8.0min, respectively) were also confirmed by comparing their fragmentation spectrum patterns. Extracts produced from a culture with a YAC without biosynthetic genes did not show peaks with such retention times, indicating that peaks were due to presence of albaflavenone genes. However, because no standard was available, quantification of the amount of albaflavenone produced by the cells was not possible. One of the genes involved in albaflavenone production in natural producer species is the cytochrome P450 (CYP170A1) encoded by *sco5223*. Cytochrome P450s require a reductase for regeneration and I speculated that the endogenous yeast reductase was not able to efficiently regenerate the Streptomyces P450. Therefore, to improve yield, I incorporated a bacterial cytochrome P450 system. A copy of the *E. coli* flavodoxin (Fld) and *E. coli* flavodoxin reductase (Fpr) was added to position 13 and 14 of the original albaflavenone YAC, with the premise that this reductase system would efficiently regenerate active CYP170A1.[65] Comparison of GC/MS results from the first YAC version, without flavodoxin and flavodoxin reductase, and the second YAC version, with flavodoxin and flavodoxin reductase, showed that cells bearing the the flavodoxin and the flavodoxin reductase produced substantially more albaflavenone (Fig. 6; note the log scale). I performed all experiments described in this section except Figure 5, which was performed by Louiza Marouche.
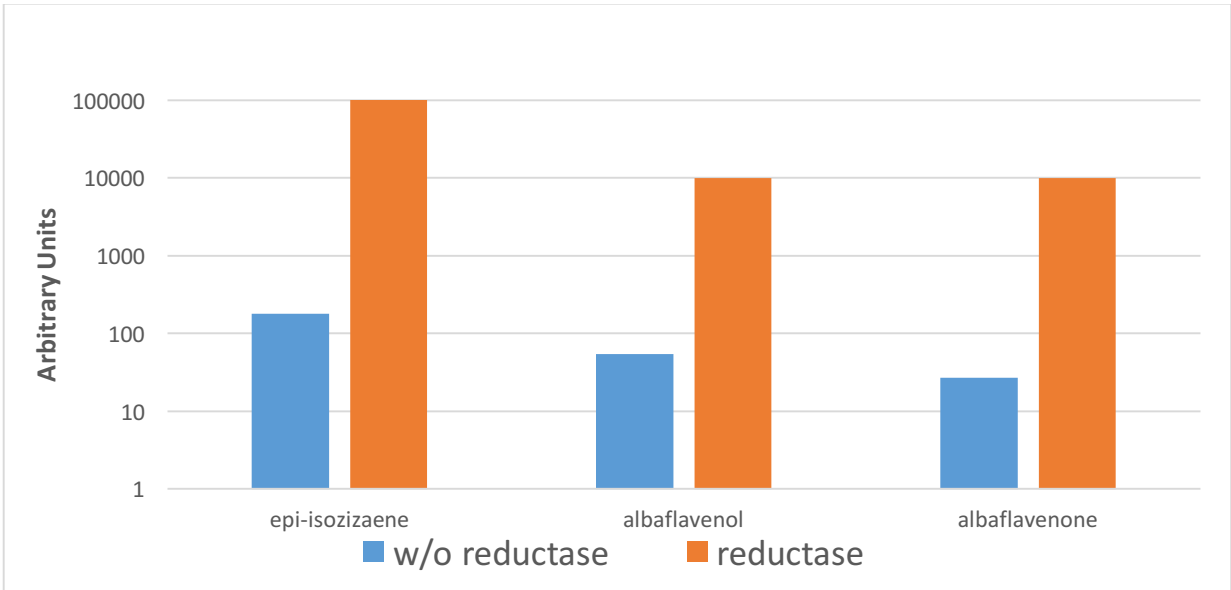
Figure 6. Normalized signal intensity for the production of albaflavenone in the absence or presence of n exogenously encoded reductase.
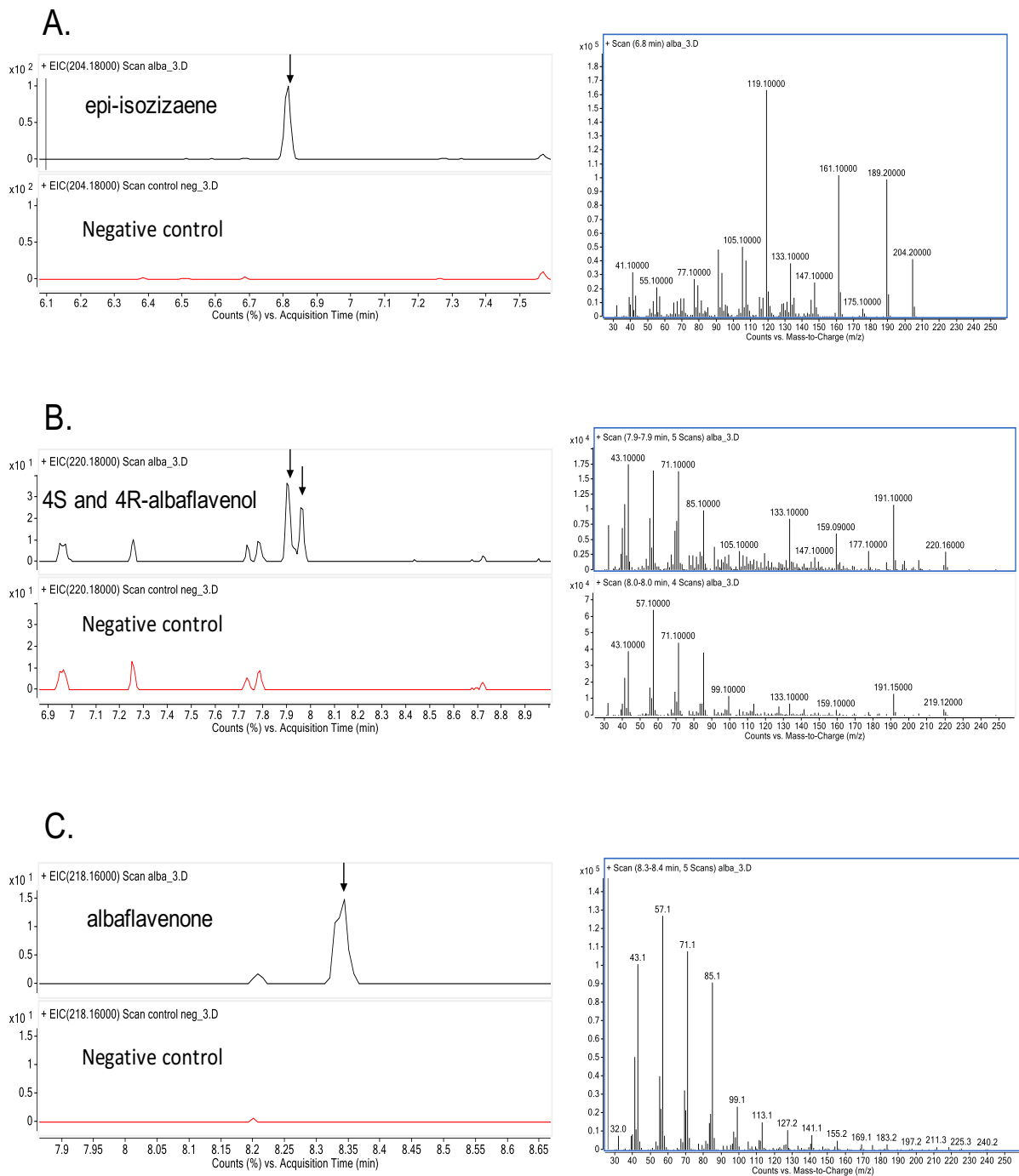
Figure 5. GC/MS analysis of in vivo products from a 2day culture of a YAC clone. Left panels represent the GC trace of (A) epi-isozizaene, (B) 4S and 4R-albaflavenol, and (C) albaflavenone. Cells lack the YAC served as a negative control. Right panels show the mass spectrum of each compound. The analysis were performed by Louiza Marouche.

4.4 Assembly of a novel YAC-based terpene biosynthetic libraries

I then tested the capacity of the YAC platform to assemble a combinatorial library of terpene biosynthesis genes. Two types of terpene library were created: the first was made by regrouping enzymes involved in the production of terpenes belonging to the same sub-class (sesquiterpene, diterpene, triterpene and tetraterpene) and the second was made by combinatorial assembly of enzymes from different sub-classes (Fig. 7).
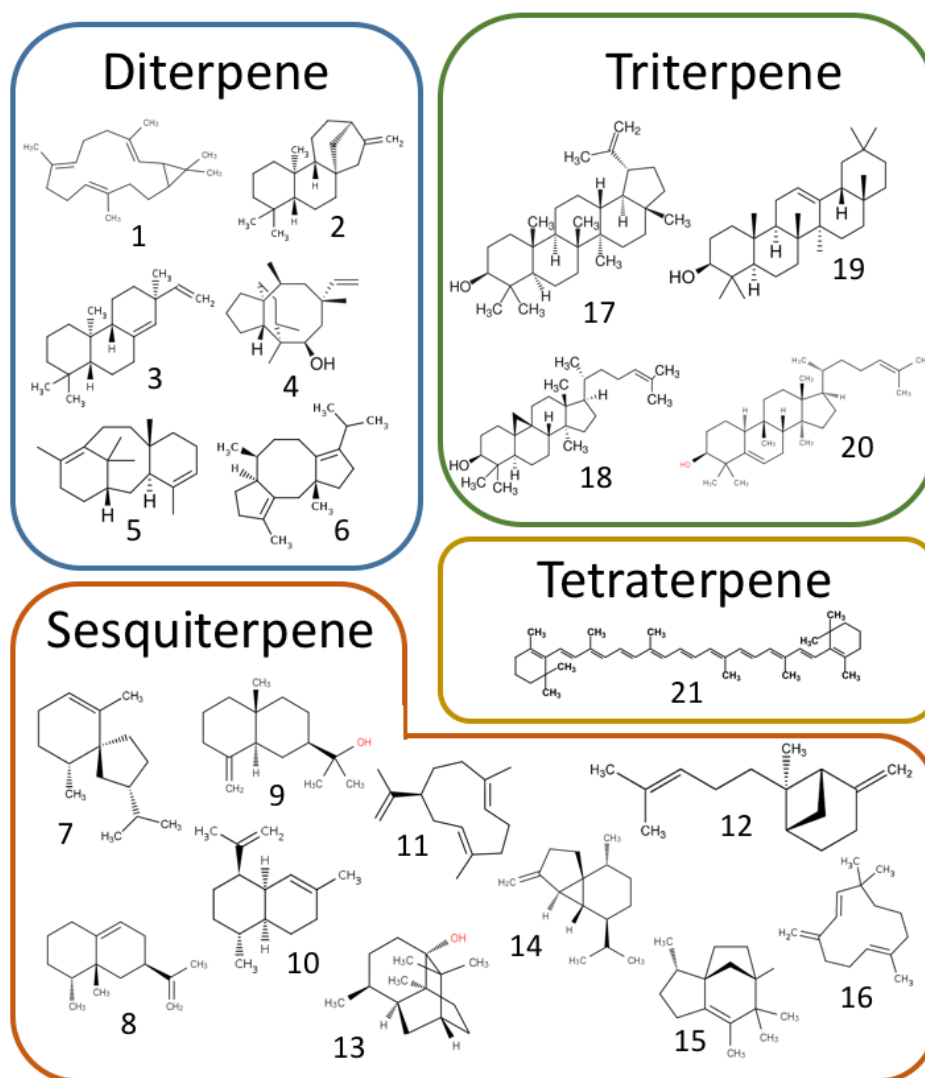
Figure 7. Terpene scaffolds that can be produced with the terpene synthases available in the library and classified according to their sub-classes. Diterpene: casbene (1), ent-kaurene (2), ent-sandaracopimaradiene (3), deoxymutilin (4), taxadiene (5), (+)-fusicocca-2,10(14)-diene (6). Sesquiterpene: vetispiradiene (7), 5-epi-aristolochene (8), β-eudesmol (9), amorpha-4,11-diene (10), germacrene B (11), trans-α-bergamotene (12), patchoulol (13), β-cubebene (14), albaflavenol (15), γ-humulene (16). Triterpene: lupeol (17), cycloartenol (18), β-amyrin (19), cucurbitadienol (20). Tetraterpene: β-carotene. Scaffolds produced by monoterpene terpene synthases are not shown.

Some enzymes were redundant for all terpene classes, such as cytochrome P450 enzymes (all accompanied by a reductase for enzyme regeneration), an important group of enzymes for modification of scaffolds, and a HMG-CoA reductase to increase the quantity of compounds produced from mevalonate pathway precursors.[39,42] In total, six libraries of compounds were created; diterpenes, triterpenes, tetraterpenes, sesquiterpenes, diterpenes/tetraterpenes and triterpenes/sesquiterpenes (Table S1A to S1F). Library clones were not screened for active compounds because the necessary target and cell-based assay readouts were not completely established. However, the diterpene/tetraterpene library confirmed that our platform was able to produce a diversity of compounds. When 20 library clones were picked at random and plated in the absence of methionine on SC, -His, -Leu, -Met to induce the expression of the genes under control of the MET promoters, half of the clones started to produce different colored products (Fig. 8). This result is consistent with the fact that many tetraterpenes are known to be colored.
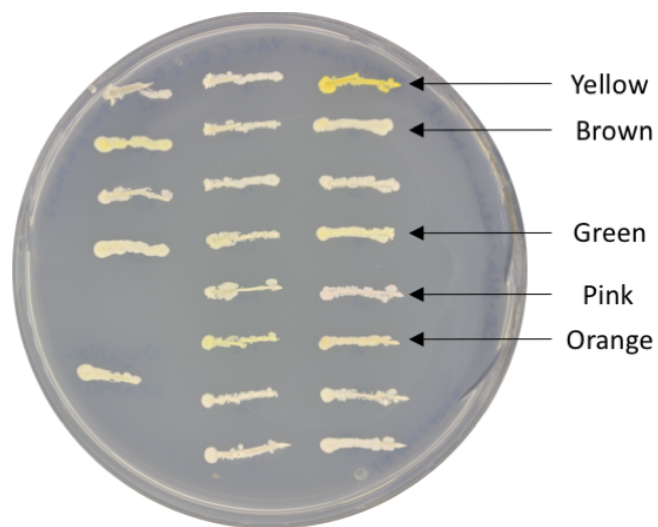
Figure 8. Diversity of pigments created by the terpene libraries. YAC cloness with genes from diterpene and tetraterpene biosynthesis pathways produce colored colonies.

5. Discussion

The advantage of our YAC assembly platform over other YAC assembly methods, and other methods to introduce heterologous genes in *S. cerevisiae* host chassis, is that it allows directional assembly of a large number of genes in a facile manner. Specific genes can be assembled for a given position in the YAC, while at other positions mixes of genes can be introduced. The use of this kind of system gives the possibility to have an initial set of genes present in every assembly to which a mix of random genes can be added to create variant molecules of a common scaffold. In addition, the number of cassettes is not limited to the 24 existing modules, but can be increased by the addition of additional entry modules with new homology arms, which allows flexibility for pathways that require the introduction of a larger number of enzymes genes. Universal standardized adaptors allow us to amplify any gene of the library with the same primer pair for cloning in any of the 24 pASS vectors by Gibson assembly. This strategy allows us to work with a large number of genes with minimal assembly errors. To ensure both YAC stability and efficient translation, all biosynthetic genes were refactored to reduce GC content, minimize potential secondary structures, and optimize codons for expression in yeast. At the moment, only four methionine-dependent promoters and six terminators are used for the gene expression, but the ultimate goal is to have 24 different promoters and terminators which will reduce the possibility of unwanted homologous recombination among the cassettes during their assembly. As an alternative system for NP pathway expression, I also tested an integrative platform , which may be useful for future yeast chassis modification (data not shown).[64]

As a proof of principle to show that the YAC platform allows efficient assembly and expression, genes that code for enzymes of the β-carotene pathway were introduced into a YAC. β-carotene was chosen because it only requires three enzymes and its biosynthesis is easy to monitor as a successful assembly and reconstitution of the biosynthetic pathway gives an orange color to the cells. In addition, β-carotene was successfully expressed in yeast by other groups using different systems, such as a simple plasmid-based system.[68] The orange colour of YAC library clones demonstrated that β-carotene was produced, and confirmed that the fragments were properly recombined into a YAC. Extraction of the β-carotene was performed, but its analysis by GC-MS was not possible as the instrument was unable to reach the β-carotene boiling point. This technical limitation prevented precise quantification of the yield of β-carotene produced by the cells. However, the production of the diagnostic orange color was sufficient to conclude that β-carotene production was successfully reconstituted in yeast using our system.

The next step in benchmarking the system was to produce a compound that had not been produced in yeast before, and ideally to use such a biosynthetic pathway as a positive control for future screens for novel antibiotics. I decided to reconstitute biosynthesis of albaflavenone, a sesquiterpene produced by the soil bacteria *S. coelicolor*. Albaflavenone was previously shown to have antibiotic properties and has a strong earth like smell, which enables simple detection of compound production. Cultures that emitted an earth-like smell were selected for extraction to test for the presence of albaflavenone. The same extraction protocol as used for β-carotene was used because both molecules are terpenes and have somewhat

similar physiochemical properties. GC-MS confirmed the presence of albaflavenone in a culture of the YAC clone, along with the 3 known intermediates of the pathway. The produced compound will next be tested in a bacterial liquid growth inhibition assay, with purified albaflavenone as a control. Subsequently, a co-culture screen will be performed with the full library, in which yeast expressing the terpene YAC library and a bacterial pathogen will be co-plated to detect compounds that are active against the pathogen by virtue of a halo of growth inhibition. Yeast that produce albaflavenone and *Micrococcus luteus* will be used to optimize the co-culture assay format. For specific protein targets, a yeast two hybrid assay has been developed to monitor protein-protein interactions (data not shown). The reporter in this system consists of a URA3 gene under the control of a modified version of the SPO13 promoter which contains eight repeated sequences of the LexA binding site.[69] A combination vector was developed where the sequences of both bait (fused with the LexA binding domain) and prey (fused with the B42 transcriptional activation domain) proteins were combined on the same plasmid, which has the advantage to need only a single auxotrophic marker for selection of both parts. My preliminary characterization of this system suggests it is effective as a selection reporter for the interaction between mycobacterial ribosomal proteins L10 and L12 and the interaction between the human transcription factor HIF1-alpha and beta (data not shown).

Based on the positive results obtained with YAC-based reconstitution β-carotene and albaflavenone biosynthesis in yeast, I then tested the generation of combinatorial libraries designed to produce terpene-like molecules. In total, six different terpene libraries were designed and constructed. Four libraries were composed of enzymes from the same terpene sub-class, namely sesquiterpenes, diterpenes, triterpenes and tetraterpenes. With this design, I

hoped to increase the chance of producing active compounds because scaffolding enzymes can only recognize specific precursors within a given sub-classes; for example monoterpene synthase can only convert geranyldiphosphate into a scaffold. Moreover, I speculated that tailoring enzymes would be more likely to recognise and interact with a given scaffold from the same sub-class.[36] The two other libraries were a mix of enzymes from two different sub-classes of terpenes, namely diterpene/tetraterpene and sesquiterpene/triterpene. These two libraries were made to assess whether novel molecules can be created by the action of enzymes that normally interact with substrates that are from different subclasses. In addition, the same enzyme from different organisms, for example a lupeol synthase from different species of plants, were used for many of the libraries.[70] It is thought that a same enzyme can perform different modifications depending on its evolutionary origin and thereby result in a higher level of diversity. Unfortunately, the assays that will serve to detect active compounds were not quite completed by the end of my MSc term and so it was not possible to screen the libraries for novel antibiotic activities. Nevertheless I was able to confirm that our approach could produce an array of diverse compounds based on a simple colony colour assay. Colonies from the diterpene/tetraterpene library showed an array of different colors caused by the presence of different combinations of genes in the YAC of each clone. Analytical chemistry assays are currently ongoing to determine the identity of these coloured molecules. Taken together, all of the above results confirmed that the gene optimization algorithm that was used to guide gene synthesis allowed adequate expression levels and protein translation to express sufficient amounts of enzymes for the YAC-mediated generation of natural product-like molecules.

In the future, once anti-mycobacterial activity is detected, multiplex PCR, qPCR and sequencing will be used to identify the YAC-encoded genes involved in the production of bioactive compounds. Subsequently, additional metabolic engineering may be performed to improve the yield of promising screen hits, for example by increasing the availability of relevant biosynthetic precursors. Over-expression of the enzymes in the mevalonate pathway, which is the source of all terpene precursors, will be required for this purpose. Previous results from other groups shown impressive results for the yield of other terpenes, such as artemisinic acid, when the mevalonate pathway is upregulated.[42,71]

6. Conclusion

The results obtained in the course of my MSc research project have provided strong proof of principle that the YAC-based natural product biosynthesis platform functions well. Achieving expression of specific pathways for the production of the natural products β-carotene and albaflavenone in yeast proved that our YAC platform enables the integration of complex pathways from a variety of source organisms into a heterologous yeast host. Moreover, the modular design of our YAC systems will allow almost infinite versatility in the future design of complex combinatorial enzyme libraries. I provided the first proof of concept for this approach with the successful generation of combinatorial terpene biosynthesis libraries. In addition to combinatorial assembly of novel biosynthetic pathways, our system will allow valuable compounds normally produced in small amounts, under specific conditions in the original producer host or limited due to scarce sources to be produced by an engineered strain of yeast with far higher yield and lower cost. Now that proof of concept has been established, the research priority is to develop a high affinity small molecule inhibitor that targets the interaction between mycobacterial ribosomal proteins L10 and L12 by selection screening of our various YAC libraries. My hope is that this screening effort will form the basis for discovering a novel first-in-class antibiotic, and as such be important in the treatment of multidrug resistant *Mycobacterium tuberculosis*. This re-emerging pathogen is a massive health threat for developing countries, where millions of new cases arise every year with high mortality rates.[72] More generally, our YAC-based biosynthetic platform and integrated selection screen assays have the potential to revolutionize discovery of natural product-like lead compounds for drug development in many if not all disease areas.

References

1. Ng, R. History of Drug Discovery and Development. *Drugs From Discov. to Approv.* 391–397 (2008). doi:10.1002/9780470403587

2. Sobolik, K. D. & Gerick, D. J. PREHISTORIC MEDICINAL PLANT USAGE : A CASE STUDY FROM COPROLITES. *J. Ethnobiol.* **12,** 203–211 (1992).

3. Drews, J. Drug discovery: a historical perspective. *Science (80-.).* **287,** 1960–64 (2000).

4. Roersch van der Hoogte, A. & Pieters, T. Science in the service of colonial agro-industrialism: The case of cinchona cultivation in the Dutch and British East Indies, 1852-1900. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **47,** 12–22 (2014).

5. Ravina, E. in *The Evolution of Drug Discovery* 1–22 (Wiley-VCH, 2011).

6. Limbird, L. E. The receptor concept: a continuing evolution. *Mol. Interv.* **4,** 326–336 (2004).

7. Comroe, J. H. Pay dirt: the story of streptomycin. Part I. From Waksman to Waksman. *Am. Rev. Respir. Dis.* **117,** 773–81 (1978).

8. Sharma, D., Cukras, A. R., Rogers, E. J., Southworth, D. R. & Green, R. Mutational Analysis of S12 Protein and Implications for the Accuracy of Decoding by the Ribosome. *J. Mol. Biol.* **374,** 1065–1076 (2007).

9. Ball, P. Quinolone generations: natural history or natural selection? *J. Antimicrob. Chemother.* **46,** 17–24 (2000).

10. Regulska, K., Stanisz, B., Regulski, M. & Murias, M. How to design a potent, specific, and stable angiotensin-converting enzyme inhibitor. *Drug Discov. Today* **19,** 1731–

1743 (2014).

11. Carter, P. Improving the efficacy of antibody-based cancer therapies. *Nat. Rev. Cancer* **1,** 118–129 (2001).

12. Antonanzas, F., Lozano, C. & Torres, C. Economic Features of Antibiotic Resistance: The Case of Methicillin-Resistant Staphylococcus aureus. *Pharmacoeconomics* 285– 325 (2014). doi:10.1007/s40273-014-0242-y

13. Lim, J. Y., Yoon, J. & Hovde, C. J. A brief overview of Escherichia coli O157:H7 and its plasmid O157. *J. Microbiol. Biotechnol.* **20,** 5–14 (2010).

14. Matthews, M. & Ph, D. *The High Cost Of Inventing New Drugs—And Of Not Inventing Them*. (2015).

15. Hanson, J. R. in *Natural Products* (ed. Hanson, J. R.) **17,** 1–34 (Royal Society of Chemistry, 2003).

16. Hunter, P. Harnessing Nature ' s wisdom. *Embo Rep* **9,** 838–840 (2008).

17. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75,** 311–335 (2012).

18. Ziegler, J. & Facchini, P. J. Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* **59,** 735–769 (2008).

19. Evans, W. C. *Alkaloids. Trease and Evans' Pharmacognosy* (2009). at <http://www.us.elsevierhealth.com/media/us/samplechapters/9780702029332/9780702 029332_2.pdf>

20. Schardl, C. L., Panaccione, D. G. & Tudzynski, P. Ergot alkaloids--biology and molecular biology. *Alkaloids. Chem. Biol.* **63,** 45–86 (2006).

21. Jakubczyk, D., Cheng, J. Z. & O'Connor, S. E. Biosynthesis of the ergot alkaloids. *Nat.*

Prod. Rep. **31,** 1328–38 (2014).

22. Cushnie, T. P. T., Cushnie, B. & Lamb, A. J. Alkaloids: An overview of their antibacterial, antibiotic-enhancing and antivirulence activities. *Int. J. Antimicrob. Agents* **44,** 377–386 (2014).

23. Falcone Ferreyra, M. L., Rius, S. P. & Casati, P. Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Front. Plant Sci.* **3,** 222 (2012).

24. Xiao, J., Muzashvili, T. S. & Georgiev, M. I. Advances in the biotechnological glycosylation of valuable flavonoids. *Biotechnol. Adv.* **32,** 1145–1156 (2014).

25. Shakya, T. *et al.* A Small Molecule Discrimination Map of the Antibiotic Resistance Kinome. *Chem. Biol.* **18,** 1591–1601 (2011).

26. Strieker, M., Tanović, A. & Marahiel, M. A. Nonribosomal peptide synthetases: Structures and dynamics. *Curr. Opin. Struct. Biol.* **20,** 234–240 (2010).

27. Winn, M., Fyans, J. K., Zhuo, Y. & Micklefield, J. Recent advances in engineering nonribosomal peptide assembly lines. *Nat. Prod. Rep.* (2016). doi:10.1039/C5NP00099H

28. Calcott, M. J. & Ackerley, D. F. Genetic manipulation of non-ribosomal peptide synthetases to generate novel bioactive peptide products. *Biotechnol. Lett.* **36,** 2407–2416 (2014).

29. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30,** 108–60 (2013).

30. McIntosh, J. a., Donia, M. S. & Schmidt, E. W. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. *Nat. Prod. Rep.* **26,** 537 (2009).

31. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48,** 4688–716 (2009).

32. Hagen, A. *et al.* Engineering a polyketide synthase for in vitro production of adipic acid. *ACS Synth. Biol.* acssynbio.5b00153 (2015). doi:10.1021/acssynbio.5b00153

33. Hertweck, C. Decoding and reprogramming complex polyketide assembly lines: Prospects for synthetic biology. *Trends Biochem. Sci.* **40,** 189–199 (2015).

34. Weissman, K. J. Uncovering the structures of modular polyketide synthases. *Nat. Prod. Rep.* **32,** 436–53 (2015).

35. Sato, T. Unique Biosynthesis of Sesquarterpenes (C35 Terpenes). *Biosci. Biotechnol. Biochem.* **77,** 1155–1159 (2013).

36. Misawa, N. Pathway engineering for functional isoprenoids. *Curr. Opin. Biotechnol.* **22,** 627–633 (2011).

37. Davies, F. K., Jinkerson, R. E. & Posewitz, M. C. Toward a photosynthetic microbial platform for terpenoid engineering. *Photosynth. Res.* 265–284 (2014). doi:10.1007/s11120-014-9979-6

38. Kirby, J. & Keasling, J. D. Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu. Rev. Plant Biol.* **60,** 335–355 (2009).

39. Pateraki, I., Heskes, A. M. & Hamberger, B. in *Advances in biochemical engineering/biotechnology* **123,** 107–139 (2015).

40. Singh, B. & Sharma, R. A. Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech* **5,** 129–151 (2015).

41. Tarapore, R. S., Siddiqui, I. A., Adhami, V. M., Spiegelman, V. S. & Mukhtar, H. The dietary terpene lupeol targets colorectal cancer cells with constitutively active Wnt/β-

catenin signaling. *Mol. Nutr. Food Res.* **57,** 1950–1958 (2013).

42. Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 940–943 (2006).

43. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496,** 528–32 (2013).

44. Croteau, R., Ketchum, R. E. B., Long, R. M., Kaspera, R. & Wildung, M. R. Taxol Biosynthesis and Molecular Genetics. *Phytochem. Rev.* **5,** 75–97 (2006).

45. Heinig, U. & Jennewein, S. Taxol : A complex diterpenoid natural product with an evolutionarily obscure origin. *African J. Biotechnol.* **8,** 1370–1385 (2009).

46. Engels, B., Dahm, P. & Jennewein, S. Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production. *Metab. Eng.* **10,** 201–206 (2008).

47. Haas, M. J. Paclitaxel plants routes in bacteria. *Sci. Exch.* **3,** 1–2 (2010).

48. Woolston, B. M., Edgar, S. & Stephanopoulos, G. Metabolic Engineering: Past and Future. *Annu. Rev. Chem. Biomol. Eng.* **4,** 259–288 (2013).

49. Mattanovich, D., Sauer, M. & Gasser, B. Yeast biotechnology: teaching the old dog new tricks. *Microb. Cell Fact.* **13,** 34 (2014).

50. Gibson, D., Young, L. & Chuang, R. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. ...* **6,** 12–16 (2009).

51. Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31,** 230–2 (2013).

52. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2647

53. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32,** 347–55 (2014).

54. Luo, Y., Lee, J.-K. & Zhao, H. Challenges and opportunities in synthetic biology for chemical engineers. *Chem. Eng. Sci.* **103,** 115–119 (2013).

55. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).

56. Oliver, J. W. K., Machado, I. M. P., Yoneda, H. & Atsumi, S. Cyanobacterial conversion of carbon dioxide to 2,3-butanediol. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 1249–54 (2013).

57. Brochado, A. R. *et al.* Improved vanillin production in baker's yeast through in silico design. *Microb. Cell Fact.* **9,** 84 (2010).

58. Klein, J. *et al.* Yeast synthetic biology platform generates novel chemical structures as scaffolds for drug discovery. *ACS Synth. Biol.* **3,** 1–29 (2014).

59. Naesby, M. *et al.* Yeast artificial chromosomes employed for random assembly of biosynthetic pathways and production of diverse compounds in Saccharomyces cerevisiae. *Microb. Cell Fact.* **8,** 45 (2009).

60. Tavassoli, A. & Benkovic, S. J. Split-intein mediated circular ligation used in the synthesis of cyclic peptide libraries in E. coli. *Nat. Protoc.* **2,** 1126–33 (2007).

61. Miranda, E. *et al.* A cyclic peptide inhibitor of HIF-1 heterodimerization that inhibits hypoxia signaling in cancer cells. *J. Am. Chem. Soc.* **135,** 10418–25 (2013).

62. Burke, D., Carle, G. & Olson, M. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science (80-.).* **236,** 806–812 (1987).

63. Copeland, N. G., Jenkins, N. a & Court, D. L. Recombineering: a powerful new tool for

mouse functional genomics. *Nat. Rev. Genet.* **2,** 769–779 (2001).

64.   Yuan, J. & Ching, C. B. Combinatorial Assembly of Large Biochemical Pathways into Yeast Chromosomes for Improved Production of Value-added Compounds. *ACS Synth. Biol.* (2014). doi:10.1021/sb500079f

65.   Zhao, B. *et al.* Biosynthesis of the Sesquiterpene Antibiotic Albaflavenone in Streptomyces coelicolor A3(2). *J. Biol. Chem.* **283,** 8183–8189 (2008).

66.   Harju, S., Fedosyuk, H. & Peterson, K. R. Rapid isolation of yeast genomic DNA: Bust n' Grab. *BMC Biotechnol.* **4,** 8 (2004).

67.   Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 38–41 (2007).

68.   Mata-Gómez, L. C., Montañez, J. C., Méndez-Zavala, A. & Aguilar, C. N. Biotechnological production of carotenoids by yeasts: an overview. *Microb. Cell Fact.* **13,** 12 (2014).

69.   Huang, J. & Schreiber, S. L. A Yeast Genetic System for Selecting Small Molecule Inhibitors of Protein--Protein Interactions in Nanodroplets. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 13396–13401 (1997).

70.   Segura, M. J., Meyer, M. M. & Matsuda, S. P. Arabidopsis thaliana LUP1 converts oxidosqualene to multiple triterpene alcohols and a triterpene diol. *Org. Lett.* **2,** 2257–2259 (2000).

71.   Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* **12,** 355–367 (2014).

72.   Manjelievskaia, J., Erck, D., Piracha, S. & Schrager, L. Drug-resistant TB: deadly,

costly and in need of a vaccine. *Trans. R. Soc. Trop. Med. Hyg.* **110,** 186–191 (2016).

Supplementary data
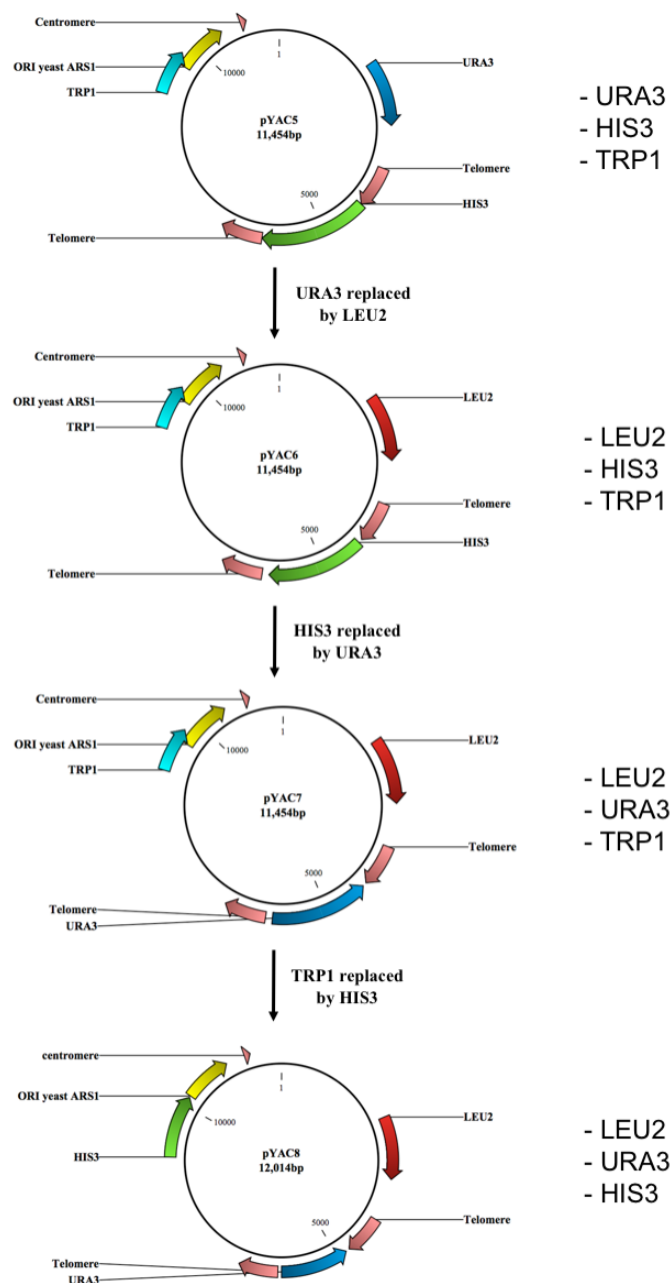
Figure S1. Creation of pYAC8. The URA3 gene of pYAC5 was first replaced by an LEU2 gene which gave pYAC6. Afterward, the HIS3 gene of pYAC6 was replaced by an URA3 gene which gave pYAC7. Finally, the TRP1 gene of pYAC7 was replaced by an HIS3 gene which gave pYAC8. Two restriction sites (XbaI and PvuII) were subsequently added by Susan Moore to yield pYAC9 (Data not shown).

Figure S2. The biosynthetic pathway of β-carotene reconstituted into a YAC. The precursor, farnesyl diphosphate (FPP), was provided by the mevalonate pathway, endogenous to *S. cerevisiae*. Afterward, it is converted into geranylgeranyl diphosphate (GGPP) by a geranylgeranyl diphosphate synthase (CrtE) and into phytoene by a phytoene synthase (CrtYB), a bifunctional enzyme that has also a lycopene cyclase activity. A phytoene dehydrogenase catalyzes two reactions where phytoene is converted into neurosporene, which is converted into lycopene by the same enzyme. Finally, lycopene is converted into β-carotene by CrtYB. (All molecule structures were taken from Wikipedia.)

Figure S3. The biosynthetic pathway of albaflavenone reconstituted in yeast. The precursor (2) farnesyl disphosphate is converted into (1) epi-sozizaene by an epi-sozizaene synthase by *sco5222*. Afterward, CYP170A1 coded by *sco5223* is responsible for the two sequential allylic oxidations which lead (4a) (4S)-albaflavenol and (4b) (4R)-albaflavenol, then to (3) albaflavenone. (Figure adapted from Bin Zhao et al. J. Biol. Chem. 2008;283:8183-8189)

Table S1: List of oligonucleotides used in the present study.

Name	Sequence
YAC0077	CAACCCTTGGCAGAACATATCCATCGCGTCCGCCATCTCCAGCAGCCGCA CGCGGCGCATAACTGTGGGAATACTCAGGTATCG
YAC0078	ACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGC ACCCGTCTACCCTATGAACATATTCCATTTTGTAATTTCG
YAC0083	GATCCTCGGGGACACCAAATATGGCGATCTCGGCCTTTTCGTTTCTTGGA GCTGGGACATCATTACGCCCGAGTAATAACTG
YAC0084	GATCCGCTGCACGGTCCTGTTCCCTAGCATGTACGTGAGCGTATTTCCTTT TAAACCACGCCAACCCTTGGCAGAACATATC
YAC0142	ATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTT CGTCTTCAACTTGGCCTCCTCTAGTACACTC
YAC0143	CAGTCCACCAGCTAACATAAAATGTAAGCTTTCGGGGCTCTCTTGCCTTCC AACCCAGTCCCTCCATCTCTTTTATATTTTTTTTCTCG
YAC0213	TTTGTATATAGGGGTTGCGAGTTCC
YAC0214	TTTGTATATAGGGGTTGCGAGTGTGC
YAC0215	TTTGTATATAGGGGTTGCGAGTTGG
sm577	GGCAGTTCAGGCTCATCAGC
sm580	CGGGCAGTTCAGGCGC
pri0561	ACTCGCAACCCCTATATACAAAGCTTA
pri0562	CTGACATCAGCAGAGGACATCCG
pri0642	TTGTATATAGGGGTTGCGAGTATATCT
pri0643	GGATGTCCTCTGCTGATGTCAGAG

Table S2: Amplification conditions for gene inserted into entry vectors.

Amplification mix:

Components	Volume (µl)
5X KAPA Buffer	5,00
10mM dNTPs	0,75
10µM pri561	0,75
10µM pri562	0,75
DNA template	1,00
HiFi KAPA hotstart	0,50
H20	16,25
total volume	25,00

PCR parameters:

Steps	Temperature(°C)	Time
initial denaturation	95	5min
denaturation	98	20s
annealing 25x	60	15s
elongation	72	4min
final elongation	72	7min

Table S3A to S3F: Gene libraries for terpene biosynthesis. Enzymes used for the six different YAC pathway libraries with their positions into the YAC. The cDNA sequence of each gene was optimized for expression in yeast. The MT identifiers was used for archival of biosynthetic gene plasmids.

Table S3A: Diterpenes pathway genes		
Module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
5	cytochrome P450 monooxygenase	MTL00210
		MTL00211
		MTL00212
		MTL00213
6	Retinol dehydrogenase	MTL00220
		MTL00221
7	casbene synthase	MTL00400
8	taxadiene synthase	MTL00184
9	ent-sandaracopimaradiene synthase	MTL00185
10	ent-cassa-12,15-diene synthase	MTL00186
11	syn-copalyl diphosphate synthase	MTL00187
12	stemer-13-ene synthase	MTL00188
13	9b-pimara-7,15-diene synthase	MTL00189
14	(+)-fusicocca-2,10(14)-diene synthase (FPPS)	MTL00377
15	geranylgeranyl diphosphate synthase	MTL00190
		MTL00191
		MTL00192
		MTL00193
		MTL00194
		MTL00195
		MTL00196
		MTL00412

		MTL00197
16	copalyl diphosphate synthase	MTL00198
		MTL00407
		MTL00397
17	kaurene synthase	MTL00199
		MTL00200
		MTL00201
		MTL00202
		MTL00203
		MTL00371
18	kaurene synthase (bi-functional)	MTL00204
		MTL00205
		MTL00206
		MTL00207
19	kaurene oxidase (require cytochrome P450 to be active)	MTL00208
		MTL00209
20	empty	
21	empty	
22	mCherry	
23	emtpy	
24	empty	

	Table S3B: Triterpenes pathway genes	
module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
5	cytochrome P450 monooxygenase	MTL00210
		MTL00211
		MTL00212
		MTL00213
6	empty	
7	lupeol synthase	MTL00243
		MTL00244
		MTL00245

module position	enzymes	MT identifiers
		MTL00246
8	putative beta-amyrin synthase	MTL00247
9	putative gamma-glutamyl hydrolase	MTL00378
10	beta-amyrin synthase	MTL00248
		MTL00249
		MTL00250
		MTL00251
11	cycloartenol synthase	MTL00252
		MTL00253
		MTL00254
		MTL00255
		MTL00256
12	mixed-amyrin synthase	MTL00257
13	cucurbitadienol synthase	MTL00258
14	empty	
15	empty	
16	empty	
17	empty	
18	empty	
19	empty	
20	empty	
21	empty	
22	mCherry	
23	empty	
24	empty	

Table S3C: Tetraterpenes pathway genes

module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
5	cytochrome P450 monooxygenase	MTL00210
		MTL00211
		MTL00212
		MTL00213

		MTL00220
6	Retinol dehydrogenase	MTL00221
7	phytoene-beta carotene synthase	MTL00215
8	phytoene dehydrogenase	MTL00216
		MTL00217
9	beta-carotene 15,15'-monooxygenase	MTL00218
		MTL00219
		MTL00392
		MTL00222
10	beta-carotene-9',10'-dioxygenase	MTL00223
		MTL00224
		MTL00225
11	empty	
12	empty	
13	empty	
14	empty	
15	empty	
16	empty	
17	empty	
18	empty	
19	empty	
20	empty	
21	empty	
22	mCherry	
23	empty	
24	empty	

Table S3D: Sesquiterpenes pathway genes

module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
		MTL00210
5	cytochrome P450 monooxygenase	MTL00211
		MTL00212

		MTL00213
6	(+)-fusicocca-2,10(14)-diene synthase (FPPS)	MTL00377
7	5-epi-aristolochene synthase	MTL00228
		MTL00231
8	aristolochene synthase	MTL00229
		MTL00230
9	germacrene B synthase	MTL00233
10	gamma-humulene synthase	MTL00234
11	beta-eudesmol synthase	MTL00235
12	beta-cubebene synthase	MTL00236
13	trans-alpha-bergamotene synthase	MTL00237
14	patchoulol synthase	MTL00238
15	amorpha-4,11-diene synthase	MTL00239
16	vetispiradiene synthase	MTL00240
		MTL00242
17	terpene synthase/cyclase family protein	MTL00241
18	empty	
19	empty	
20	empty	
21	empty	
22	mCherry	
23	empty	
24	empty	

Table S3E: Diterpenes/Tetraterpenes pathway genes		
module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
5	cytochrome P450 monooxygenase	MTL00210
		MTL00211
		MTL00212
		MTL00213
6	Retinol dehydrogenase	MTL00220

		MTL00221
7	casbene synthase	MTL00400
8	taxadiene synthase	MTL00184
9	ent-sandaracopimaradiene synthase	MTL00185
10	ent-cassa-12,15-diene synthase	MTL00186
11	syn-copalyl diphosphate synthase	MTL00187
12	stemer-13-ene synthase	MTL00188
13	9b-pimara-7,15-diene synthase	MTL00189
14	(+)-fusicocca-2,10(14)-diene synthase (FPPS)	MTL00377
15	geranylgeranyl diphosphate synthase	MTL00190
		MTL00191
		MTL00192
		MTL00193
		MTL00194
		MTL00195
		MTL00196
		MTL00412
16	copalyl diphosphate synthase	MTL00197
		MTL00198
		MTL00407
		MTL00397
17	kaurene synthase	MTL00199
		MTL00200
		MTL00201
		MTL00202
		MTL00203
		MTL00371
18	kaurene synthase (bi-functional)	MTL00204
		MTL00205
		MTL00206
		MTL00207
19	kaurene oxidase (require cytochrome P450 to be active)	MTL00208
		MTL00209
20	phytoene-beta carotene synthase	MTL00215
21	phytoene dehydrogenase	MTL00216
22	Cherry	
23	beta-carotene-9',10'-dioxygenase	MTL00222
		MTL00223

		MTL00224
		MTL00225
24	beta-carotene 15,15'-monooxygenase	MTL00217
		MTL00218
		MTL00219
		MTL00392

Table S3F: Triterpenes/Sesquiterpenes pathway genes

module position	enzymes	MT identifiers
1	HMG-CoA reductase	MTL00214
2	NADPH-cytochrome P450 reductase CPR1	MTL00226
3	P450 reductase ATR1	MTL00227
4	elicitor-inducible cytochrome P450	MTL00232
5	cytochrome P450 monooxygenase	MTL00210
		MTL00211
		MTL00212
		MTL00213
6	terpene synthase/cyclase family protein	MTL00241
7	lupeol synthase	MTL00243
		MTL00244
		MTL00245
		MTL00246
8	putative beta-amyrin synthase	MTL00247
9	putative gamma-glutamyl hydrolase	MTL00378
10	beta-amyrin synthase	MTL00248
		MTL00249
		MTL00250
		MTL00251
11	cycloartenol synthase	MTL00252
		MTL00253
		MTL00254
		MTL00255
		MTL00256
12	mixed-amyrin synthase	MTL00257
13	cucurbitadienol synthase	MTL00258
14	(+)-fusicocca-2,10(14)-diene synthase (FPPS)	MTL00377

15	5-epi-aristolochene synthase	MTL00228
		MTL00231
16	aristolochene synthase	MTL00229
		MTL00230
17	germacrene B synthase	MTL00233
18	gamma-humulene synthase	MTL00234
19	beta-eudesmol synthase	MTL00235
20	beta-cubebene synthase	MTL00236
21	trans-alpha-bergamotene synthase	MTL00237
22	patchoulol synthase	MTL00238
23	amorpha-4,11-diene synthase	MTL00239
24	vetispiradiene synthase	MTL00240
		MTL00242