Review

Synthetic biology advanced nature product discovery

Junyang Wang[1], Jens Nielsen[1,2,3,*] and Zihe Liu[1,*]

[1] Beijing Advanced Innovation Center for Soft Matter Science and Engineering, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China; junyang199036@163.com
[2] Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[3] BioInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen N, Denmark
* Correspondence: nielsenj@chalmers.se (J.N.); zihe@mail.buct.edu.cn (Z.L.)

Abstract: A wide variety of bacteria, fungi and plants can produce bioactive secondary metabolites, which are often referred to as natural products. With the rapid development of DNA sequencing technology and bioinformatics, a large number of putative biosynthetic gene clusters have been reported. However, only a few natural products can be detected when isolated species are grown under conventional laboratory conditions, as most biosynthetic gene clusters are not expressed or are expressed at extremely low levels at these conditions. With the rapid development of synthetic biology, advanced genome mining and modification strategies have been reported, and they provide new opportunities for discovery of natural products. This review discusses advances in recent years that can accelerate the design, build, test, and learn (DBTL) cycle of natural product discovery, and prospects trends and key challenges for future research directions.

Keywords: natural products; biosynthetic gene clusters; synthetic biology; genome mining strategies; modification strategies; design-build-test-learn (DBTL) cycle

1. Introduction

Natural products (NPs) derived from secondary metabolites of bacteria, fungi and plants have played an important role in traditional drug development. Since the discovery of penicillin and its widespread use as an anti-infective drug, the research and development of NP-derived drugs has opened a rich chapter in the history of human health. NPs and their semi-synthetic derivatives have been playing a crucial role in clinical medicine as antibacterials, antifungals, antivirals, immunosuppressants and enzyme inhibitors [1]. In addition, they are also widely used in agriculture as herbicides, insecticides and fungicides [2]. However, since the 21st century more and more pathogenic bacteria have become drug-resistant. In order to deal with the increasing resistance to antibiotics, there is an urgent need to discover NPs with new structures and new biological activities [3].

Traditional NP discovery strategies, either through chemical synthesis or direct extraction from native hosts, have been successful and discovered many compounds. However, since the 1960s, after a short 10-year golden age, the NP discovery has faced many challenges. A large number of known compounds have been repeatedly discovered. Moreover, traditional methods are inefficient and mostly low-throughput, and could not reach the discovery speed of putative biosynthetic gene clusters (BGCs). Another related challenge is that most BGCs are silent or weakly expressed under laboratory conditions [4], and the production level of NPs in the native host is very low, which will increase the discovery cost, mainly on product extraction and detection [5]. Thus, it is urgent to develop new design, build, test strategies that can allow efficient detection of novel NPs.

With the advances of sequencing technology and bioinformatics analysis, a large amount of genome sequence data and putative BGCs have accumulated in public databases. For example, each fungal genome contains 50-90 NP BGCs, which means that these microorganisms have the ability to synthesize 50-90 kinds of NPs [6]. However, the actual

number of identified NPs in each genome is far from being exploited. Meanwhile, the gap between the number of predicted BGCs and the identified NPs is still increasing. The rapid development of synthetic biology, including advances in Design-Build-Test-Learn (DBTL) technologies, have greatly enabled mining of novel NPs. The DBTL cycle includes the design of the initial strains or the establishment of a preliminary model system to achieve the determined engineering goals, the construction of the strains, and the testing of their functions on the specified indicators to evaluate the outcomes and understand which engineering strategy is effective and which are invalid (and why), and then incorporate the learned knowledge into the decision of the subsequent DBTL cycle (Figure 1). The latest developments in synthetic biology technologies, including data mining and pathway design [7,8], rapid DNA assembly [9], genome editing [10], comprehensive pathway reconstruction [11], single-cell technology, high-throughput screening and compound monitoring, are accelerating the production process of the target compound [12]. Based on the iterative application of the DBTL cycle, academic and industrial biofoundries have been developed to boost the next wave of NP discovery [13]. Here, we discuss recent developments of synthetic biology methods in the design, build, test, and learning steps of NP discovery.

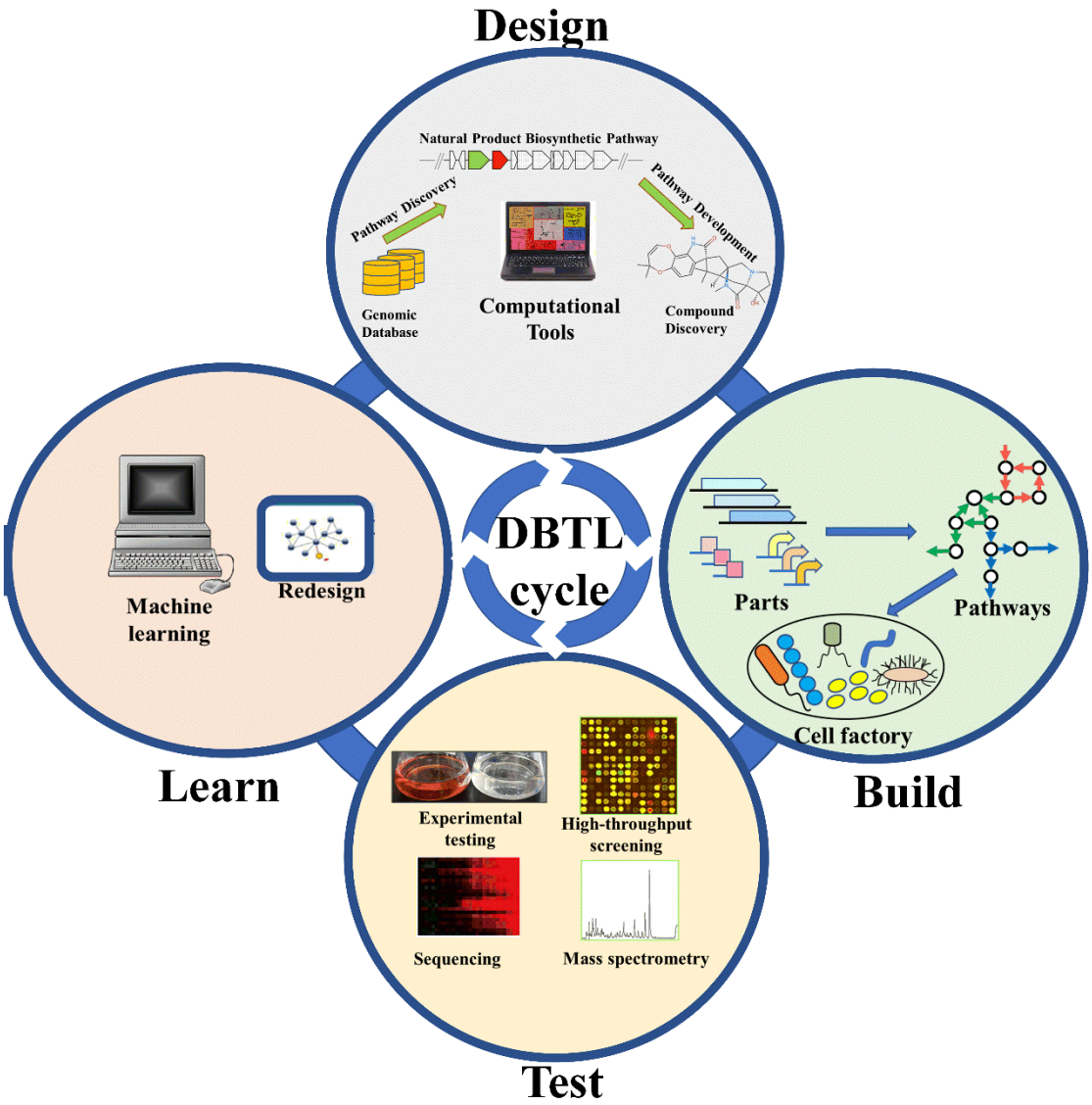

Figure 1. The design–build–test–learn cycle for strain engineering. The key aspects of each phase of the design–build–test–learn cycle are presented. The cycle starts with computational strategies for natural product biosynthetic pathway prediction and produces an engineered microbial cell factory that performs the desired function(s).

2. The design stage for natural product discovery

In the design stage, it is particularly important to identify genes involved in the synthesis of NPs. A number of databases have been established, including virtualized Registry of Standard Biological Parts that provides a platform for storage, exchange and retrieval of "component" information, KEGG, MetaCyc and BRENDA that store information on known metabolic reactions [14-16]. However, so far many genes involved in the synthesis of NPs are still unknown. Recently developed omics methods and computational biotechnologies provide abundant data and advanced tools for identifying target genes [17] (Table 1).

Table 1. Tools for natural product pathway discovery, prediction, and analysis.

Name	Description	Web Address
NP.searcher	Website for the prediction of gene clusters of PKs and NRPs in genome sequences	http://dna.sherman.lsi.umich.edu
CoExpNetViz	Website allows researchers to submit their transcriptomics data for cross-species coexpression analysis and to find whether the gene may have similar functions and regulation	http://bioinformatics.psb.ugent.be/webtools/coexpr/
ATTED-II	Database of coexpressed genes, can be used to predict coregulated gene groups	http://www.atted.bio.titech.ac.jp
PRISM 3	Web server for the prediction of genetically encoded NRPs, type I and II PKs, aminocoumarins, antimetabolites, bisindoles, and phosphonates based on chemical graphs	http://magarveylab.ca/prism/
antiSMASH50	Web server for the identification and analysis of 52 types of biosynthetic gene clusters in bacteria, fungi, and plants	https://antismash.secondarymetabolites.org
DeepBGC	Deep learning strategy to reduce false-positive rates of biosynthetic gene cluster (BGC) identification and improve the ability to discover novel BGC classes compared with antiSMASH	https://pypi.org/project/deepbgc/0.0.2/
RODEO	Algorithm developed to identify ribosomally synthesized and post-translationally modified peptide BGCs	http://www.ripprodeo.org
RiPPER	Method for family-independent identification of the precursor peptides of RiPPs	https://github.com/streptomyces/ripper
IMG-ABC	Database of experimentally verified and predicted BGCs across 40 000 isolated microbial genomes	https://img.jgi.doe.gov/abc/
ATLAS	Database of all theoretical biochemical reactions based on known biochemical principles and compounds	http://lcsb-databases.epfl.ch/atlas/
RetroPath2.0	Open-source retrosynthesis workflow for pathway prediction based on generalized reaction rules	https://github.com/brsynth/rp2paths
MRE	Open web server for pathway design; this tool focuses on suggestions of which foreign enzymes and pathways could	http://www.cbrc.kaust.edu.sa/ mre/

	maintain effective activity in host cells by considering endogenous pathways.	
ASMPKS	Identification of PKS genes from genomic DNA, prediction of domain architecture, and visualization of predicted polyketide products. Also includes a database of known polyketides	http://gate.small-soft.co.kr:8008/pks/
Bagel2	Annotation of putative bacteriocins and lantibiotics from genomic DNA. Includes a database of validated bacteriocins	http://bagel2.molgenrug.nl/
CLUSEAN	Identification of domains and prediction of specificities for PKS and NRPS genes	https://bitbucket.org/tilmweber/clusean
ClustScan	Annotation and product structure prediction for PKS, NRPS, and hybrid NRPS/PKS genes from genomic DNA. Includes stereochemistry prediction	http://csdb.biooserv.pbf.hr/csdb/ClustScanWeb.html
NORINE	Database of more than 1100 peptides containing structural information as well as biological activity, producing organisms, and literature references	http://bioinfo.lifl.fr/norine/
NRPS-PKS	PKS, NRPS, and hybrid NRPS/PKS domain identification and specificity prediction, as well as a database of characterized gene clusters	http://www.nii.res.in/nrps-pks.html
NRPSpredictor2	Prediction of adenylation domain specificity based on 10-amino-acid Stachelhaus code and 8-angstrom signature. Includes bacterial and fungal prediction	http://nrps.informatik.uni-tuebingen.de/
PKS-NRPS Analysis Website	Identification of PKS and NRPS domains, as well as adenylation domain specificity prediction based on the 8-aminoacid code	http://nrps.igs.umaryland.edu/nrps/
SBSPKS	Structural modeling of PKS modules and identification of key residues in the interfaces between modular PKS subunits. Also includes all functionalities and database support of the NRPS-PKS tool	http://www.nii.ac.in/sbspks.html
SMURF	Annotation of polyketide, non-ribosomal peptide, NRPS-PKS hybrid, indole alkoloid, and terpene biosynthetic gene clusters from fungal genomic DNA, including prediction of borders	http://jcvi.org/smurf/index.Php
2metDB	Standalone (Mac) tool to mine PKS/NRPS gene clusters	http://secmetdb.sourceforge.net/
ClusterFinder	Standalone tool (LINUX and MacOS) to identify BGCs with an non-rule based approach	https://github.com/petercim/ClusterFinder
eSNaPD	Web application to mine metagenomic datasets for BGCs	http://esnapd2.rockefeller.edu/
EvoMining	Web application for phylogenomic approach of cluster identification	http://148.247.230.39/newevomining/new/evomining_web/index.html

GNP	Web application to mine and analyze BGCs, mainly PKS/NRPS	http://magarveylab.ca/gnp/#!/genome
MIDDAS-M	Web application to use transcriptome data to identify BGC coordinates in fungal genomes	http://133.242.13.217/MIDDAS-M/
MIPS-CG	Web application to identify BGC coordinates in fungal genomes without transcriptome data	http://www.fung-metb.net/
NaPDoS	Web application offering phylogenomic analysis of PKS-KS and NRPS-C domains	http://napdos.ucsd.edu/
LSI-based A-domain function predictor	Web application to predict A-domain specificities	http://bioserv7.bioinfo.pbf.hr/LSIpredictor/AdomainPrediction.jsp
NRPS/PKS substrate predictor	Web application to predict A-domain/AT-domain specificities	http://www.cmbi.ru.nl/NRPS-PKS-substrate-predictor/
NRPSsp	Web application to predict A-domain specificities	http://www.nrpssp.com/
PKS/NRPS Web Server/Predictive Blast Server	Web application to determine domain organization and A-domain specificities	http://nrps.igs.umaryland.edu/nrps/
SEARCHGTr	Web application to predict glycosyltransferase specificities	http://linux1.nii.res.in/~pankaj/gt/gt_DB/html_files/searchgtr.html
Antibioticome	Web accessible database on compounds, compound families and modes of action	http://magarveylab.ca/antibioticome
ChEBI	Web accessible database and ontology on compounds focused on small molecules	https://www.ebi.ac.uk/chebi/
ChEMBL	Web accessible database on bioactive compounds with druglike properties	https://www.ebi.ac.uk/chembl/
ChemSpider	Web accessible database on structures and properties of over 35 million structures	http://www.chemspider.com/
KNApSAcK database	Web accessible database on compounds; standalone version of KNApSAcK metabolite database available	http://kanaya.aist-nara.ac.jp/KNApSAcK/
Novel Antibiotics Database	Web accessible database on compounds	http://www.antibiotics.or.jp/journal/database/database-top.htm
PubChem	Web accessible database on compounds and bioactivities;	http://pubchem.ncbi.nlm.nih.gov/
StreptomeDB	Web accessible database on compounds produced by streptomycetes; download of compounds and metadata in SD format	http://www.pharmaceutical-bioinformatics.de/streptomedb

GNPS	Generic metabolomics portal to analyze MS/MS data (dereplication and molecular networking)	http://gnps.ucsd.edu/
GNP/iSNAP	Web application to automatically identify metabolites in MS/ MS data based on genomic data	http://magarveylab.ca/gnp/
NRPquest	Web application to correlate NRP tandem data with gene clusters	http://cyclo.ucsd.edu
Pep2Path	Standalone application to correlate peptide sequence tags with NRP and RiPP BGCs	http://pep2path.sourceforge.net

This table is adapted from tables published previously[17].

With the rapid development of gene sequencing technology and the related operating technologies such as molecular biology and genetics, new research ideas and approaches have been applied for the discovery of new NPs. The continuous reduction of sequencing costs has made it more and more convenient and feasible to obtain genome sequence information of different species through high-throughput genome sequencing technology. Through the analysis of these genomic data, it is possible to discover, screen and identify potential "silent" gene clusters related to active compounds with novel structures. Therefore, mining for novel active NPs based on massive genomic data has become the focus and hotspot of recent research. With the significant increase in the processing speed and accuracy of DNA sequence analysis, analysis tools based on a large amount of genomic analysis data has been established. For example, BLAST and FASTA can be used to infer the function of unknown genes [18]. In addition to blast based on sequence homology, Hidden Markov model (HMM) analysis of the protein family (Protein family, Pfam) has also been widely used [19]. Moreover, the statistical model antiSMASH, which specializes in analyzing and predicting the products of BGCs, has also been published [20] (Table 1). With the significant improvement of sequence processing technologies, the accuracy of protein function prediction has also been improved, and the evaluation of BGCs has become more accurate [21]. Therefore, the method of mining secondary metabolic biosynthetic gene clusters from genome sequence data is widely used.

Identifications of genes encoding secondary metabolic biosynthetic enzymes is a classic method for mining new NPs. Although structures of secondary metabolites are rich and diverse, their biosynthetic mechanisms are relatively conservative, for example, the similarity of amino acid sequences of core enzymes is very high [22]. The scaffold core structure of many NPs is polyketide or peptide, which is controlled by the highly conserved polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS), respectively. By searching for the biosynthetic genes that control the structure of the scaffold, the biosynthetic gene clusters of NPs can be identified. With the accumulation of biosynthetic knowledge and the advancement of bioinformatics tools and databases, chemical scaffolds of metabolites synthesized by gene clusters can be predicted, and BGCs with new chemical scaffolds can be studied. For example, siphonazole is a NP isolated from *Herpetosiphon* species with anti-plasmodium properties [23]. To track its biosynthetic pathway, genome mining and imaging mass spectrometry technology were used and based on this it was suggested that it belongs to a hybrid polyketide synthase/non-ribosomal peptide synthetase (PKS/NRPS) pathway [23]. On the other hand, if microorganisms are difficult to separate and cultivate, it is not feasible to identify BGCs with the aforementioned methods. An alternative method is to first predict the scaffold structure through bioinformatics, and then chemically synthesize the compound. This method has been used to discover an antifungal peptide [24], proving the comprehensive ability of bioinformatics and chemical synthesis in the study of silent gene clusters. However, this strategy is limited to the study of silent gene clusters whose structure can be accurately predicted by bioinformatics software.

Another method to discover new NPs is to analyze not only the coding genes but also the regulatory genes and resistance genes of the BGCs. With the growing understanding of NP biosynthesis pathways, it has been discovered that BGCs not only contain biochemical enzymes, but also regulatory elements, transporters, and resistance genes [22]. Therefore, NP mining methods based on resistance or regulatory genes have been developed. For instance, microorganisms that can produce antibiotics have their own resistance system, which can effectively protect themselves from the synthetic antibiotics. The resistance mechanisms are diverse, including using efflux pumps and degrading enzymes to remove antibiotics, and modifying endogenous proteins to effectively prevent them from binding to antibiotics [25]. The required resistance gene often co-locates with the synthetic gene that encodes the production of antibiotics, so it can be used as a tag to discover putative antibiotics. Using this strategy, a new herbicide was identified by searching the dihydroxyacid dehydratase gene in the published fungal genomes [22].

The discovery of NPs can also be achieved through genome mining based on systematic evolution. The synthesis of new compounds is a very complex process. A recent study showed that by analyzing the evolutionary characteristics of 10,000 biosynthetic gene clusters, high-frequency of insertions, deletions and repetitive events occurred in secondary metabolic processes [26]. Two main research strategies based on systematic evolution can be used in the research of NPs: one strategy is to construct an evolutionary tree of strains based on conservative housekeeping genes or core genomes, then analyze the NPs produced by them and identify potential producing strains; The other strategy is to construct an evolutionary tree of genes based on genes or BGCs of secondary metabolites, and the evolutionary history of the synthetic genes or gene clusters of the products can be inferred from these gene trees. Compared with the method based on the similarity analysis of single sequence, the analysis of biosynthetic function of the enzymes can be more accurate [18].

The discovery of NPs has been boosted through the development of metagenome sequencing and single-cell sequencing. The metagenome includes all the genetic information of the microbial community of both culturable and unculturable microbes. The gene clusters of secondary metabolites are highly repetitive, and sometimes difficult to be analyzed. Through analysis of metagenomics, together with the analysis of the diversity and distribution of NPs in the living environment, it is also possible to discover novel substances and their biosynthetic pathways [27]. Moreover, the number of culturable microbes accounts for less than 5% of the total number of microbes. Because traditional microbial technology cannot obtain sufficient quantities of genome DNAs, it is difficult to obtain a large amount of diverse microbial genetic information through genome sequencing technology. The development of single-cell genome sequencing technology provides the possibility to solve this problem [28]. Different from the mass amount of data and complex analysis of metagenomics, single-cell genomes analysis only focus on the genetic characteristics of objects in the most basic biological unit [29]. Single-cell genomics and metagenomics research therefore could complement each other and work in synergy. Single-cell genome mining can directly and accurately discover the evolutionary characteristics and functions of a single cell, while metagenomic mining focuses on obtaining more genetic information on the environmental and evolutional basis. Gene fragments obtained by metagenomics are helpful for pathway prediction in the process of single-cell genome analysis [30]. With the continuous declining of sequencing costs and the continuous upgrading of sequencing technology, it will become easier to mine genetic information in single cells and complex environments, which is of great significance for revealing more putative BGCs.

In addition to identify putative synthesis pathway, creating pathways not found in nature is becoming a hot research field to discovery new NPs. Combinatorial biosynthesis is based on the versatility of the enzyme substrate, which produces new 'unnatural' NPs through the use of engineered catalytic enzymes or metabolic pathways [31,32]. In the process of assembling NPs, the diversity of monomers largely determines the diversity of their structures. The modular type I polyketide synthase (mPKSs) consists of continuous

catalytic modules, and each module has a different catalytic region to complete a cycle of C chain extension. For example, in *Streptomyces cinnamonensis*, the polyether antibiotic monensin is biosynthesized through the action of mPKS. The lipid acyltransferase region (AT region) is the fifth module in the monensin synthesis PKS complex, which can absorb unnatural malonic acid derivatives as monomers to synthesize new monensin precursor derivatives. Based on the computer model of the AT region, Bravo-Rodriguez *et al.* predicted that the AT active region of the enzyme can absorb a larger propynyl group as a substrate monomer, by adding a synthetic compound propargyl-malonyl-N-acetyl-cysteamine, the propynyl-monensin precursor compound was produced by *Streptomyces cinnamonensis* A495 [33]. In the process of polyketide biosynthesis, each module in class I PKS (mPKSs) catalyzes a specific reaction step, and then passes the mature product to the next module. This property makes it possible to design and synthesize new products by "permutation and combination" of these enzyme complexes. Modification of specific enzymes in mPKSs, and then changes the substrates catalyzed to produce new compound structures, has become a routine experimental method for combinatorial biosynthesis. For example, using the acetyltransferase region in the rapamycin synthesis pathway to replace the acetyltransferase region in the erythromycin synthesis system, 61 6-deoxyerythronolide B (6-DEB) analogs can be obtained, which contain a variety of new structures compound [34]. Carbohydrate compounds usually play an important role in the drug-target interaction, and the glycosylation process will have a significant impact on the solubility and biological activity of the drug. Therefore, the glycosylation of compounds provides a new direction for new drug discovery. The glycosylated NDP-d-digitoxose synthetic plasmid was expressed in *S. argillaceus* M3W1, and seven new structural analogues of mithramycin were obtained by changing the contour of the sugar molecule or changing its molecular contour and 3-side chain at the same time [35]. One of them (demycarosyl-3D-β-ddigitoxosyl-mithramycin SK) shows highly effective anti-tumor activity, but its cytotoxicity is much lower than that of mithramycin.

3. The build stage for natural product discovery

Putative BGCs can be evaluated either through activation in the native producer or expression in a heterologous host. In-situ activation of putative BGCs has been greatly advanced by recently developed multiplexed genome editing tools (Table 2). The new generation of genome editing tools based on CRISPR-Cas technology has the advantages of high efficiency, fast operation and high fidelity. It is the mainstream technology of multiplexed genome engineering at present, especially using the type II CRISPR-Cas9 or CRISPR-Cas12a (Cpf1) gene editing technology. For example, a CRISPR-Cas9 mediated knock-in of the kasO*p method for activation of silent BGCs has been reported and successfully discovered several novel pathways and NPs in streptomyces species [36]. Moreover, multiplexed automated genome engineering [37] has been reported with the capacity to simultaneously regulate the expression levels of twenty genes, and generate 4.3 billion combinatorial genomic mutations per day [38]. However, the use of CRISPR-Cas technology in most chassis cells requires the introduction of external sources of Cas protein, which can cause cytotoxicity. In order to solve this problem, researchers have developed genome editing technology based on the microbe's endogenous CRISPR-Cas system in multiple microorganisms, such as *Sulfolobus islandicus*, *Haloarcula hispanica*, *Clostridium tyrobutyricum*, *Clostridium pasteurianum*, *Lactobacillus crispatus* and *Zymomonas mobilis* [39]. These systems allow fast genome engineering including gene insertion, deletion, regulation and single base editing, with the editing efficiency in some cases reaching 100% , and will not be affected by the toxicity of exogenous Cas protein [40].

Table 2. Experimental strategies in the construction of chassis cells for natural products.

DNA assembly	Reference

Golden Gate assembly	Scarless method that can assemble multiple DNA fragments into specific plasmids using type IIs restriction enzymes	[41]
Start-Stop assembly	Optimized Golden Gate assembly method, which can assemble 60 DNA parts in one destination vector	[42]
One-step SLIC	Scarless and one-step method for DNA assembly based on 3′-to-5′ exonuclease activity of T4 DNA polymerase	[43]
Gibson assembly	Scarless, one-step, and isothermal DNA assembly method, which can assemble multiple DNA fragments into any plasmid using T5 exonuclease, Taq DNA ligase, and Pfu DNA polymerase	[44]
TEDA	Optimized Gibson assembly method, which requires only T5 exonuclease; thus, costs were significantly reduced (~1200-fold)	[45]
LCR assembly	DNA assembly method that can assemble 20 DNA fragments in one step by introducing single-stranded bridging oligos between two neighboring DNA parts	[46]
TAR	Large DNA fragment capture and cloning method depending on the highly efficient homologous recombination system of Saccharomyces cerevisiae	[47]
CATCH	One-step targeted clone method, which can capture 100-kb DNA genomic sequences based on Cas9 and Gibson	[48]
Programmable genome engineering	Genome assembly method, which can rearrange 1.55-Mb genome sequences by combining Cas9 and lambda-red recombination	[49]
Genome editing tools		
MAGE	Simultaneous editing of multiple genes using short single-stranded oligonucleotides (ssDNA); capable of simultaneously targeting multiple genes with moderate efficiency but has extensive off-target mutagenesis and low portability	[50]
TALLEN	Simultaneous editing of multiple genes using TALENs; has high portability and moderate off-target effects but low multiplex ability	[38]
CRISPR/Cas	Genomic DNA is specifically cleaved under the guidance of RNA; has a simpler manipulation process and higher efficiency	[51]

CRISPRi	Represses the transcription of a gene using guide RNA and inactive Cas	[52]
CRISPR-AID	Trifunctional system that can simultaneously achieve gene deletion, transcriptional activation, and repression	[53]

This table is adapted from tables published previously[17].

For microbes with harsh cultivation requirements or lack of genetic manipulation tools, heterologous expression of putative BGCs can be used to discovery NPs. Heterologous hosts have many advantages, such as clean background, fast growth, and mature genetic manipulation tools. Given that BGCs usually include all genes required for biosynthesis of the target NP, cloning of the entire BGC for heterologous expression is of great interest [54]. However, because putative BGCs may have high G+C content, high sequence similarities, and generally large size in many cases reaching over 100 kb, selection of the suitable cloning method is crucial. The selection of cloning method depends on the size and complexity of the BGC, whether refactoring is needed, the target NP and the expression host. For example, PCR and Gibson assembly-based cloning and refactoring of a streptophenazine BGCs in *S. coelicolor* M1146 resulted in detection of over 100 streptophenazines [54].

Heterologous cloning for NP discovery mainly includes DNA fragmentation, cloning, expression and analysis. According to the method of DNA fragmentation, heterologous cloning can be divided into random library cloning and direct cloning. The random library cloning method constructs expression libraries on random spliced genomes from mixed populations (such as environmental DNAs) or pure cultures, and screens for novel NPs [55,56]. Both sequenced and unsequenced genomes can be used to construct random libraries. The genome is broken into gene fragments ranging from 10 to 200 kb by partial restriction endonuclease cleavage or mechanical shearing force, that can well cover the size of NP BGCs. However, random library cloning has high chances of disruptions of BGCs, in order to obtain clones containing enough BGCs, it is necessary to obtain a genome coverage of 10-20 times [57]. This need can be solved by optimizing the cloning process, for example, increase the efficiencies and capacities of DNA extraction, fragmentation, cloning and transformation, avoiding degradations of large DNA fragments, and normalize the cloning and transformation efficiencies among BGCs with varied sizes. After genome extraction and fragmentation, all gene fragments are assembled and transformed into the cloning host for screening [58]. Many NPs have been discovered through random library cloning based on different library construction strategies (such as fosmid, cosmid, phage artificial chromosome, bacterial artificial chromosome (BAC) [55,56]. The random library cloning method is particularly useful when the genome information and features are insufficient. This method has the advantage of covering to the entire genetic material, including potential BGCs even identified [59]. However, this method must rely on high-throughput screening and analysis platforms.

The direct cloning method relies on genome sequencing and bioinformatics analysis to predict BGCs. Target BGCs will then be captured through *in vitro* CRISPR based digestion or PCR amplification and cloned into target host strains. This method can directly isolate target gene clusters, bypassing libraries construction and the time-consuming and labor-intensive screening process. If refactoring is needed, ORFs of cDNAs from target BGCs will be amplified through PCR or reverse PCR, respectively, and assembled with promoters and terminators from the host strains [60]. If refactoring is not needed, the target BGCs can be achieved through in vitro CRISPR based digestion. This "molecular scissor" system can specifically recognize target DNAs through user-defined guide RNAs, and achieve efficient and precise cutting [61]. In the past ten years, the direct cloning method has made great progress [56,62-64]. However, the direct cloning method relies heavily on the quality of genome sequencing and annotation, and can only analyze few gene clusters at one reaction, which greatly reduces the efficiency of NP discovery. With

the advancement of synthetic biology tools, it will be ideal that all putative BGCs in one target genome can be cloned at one reaction.

According to the method of DNA assembly, heterologous cloning for NP discovery can be divided into the methods for cloning BGCs with or without the need of refactoring (Table 2). The widely used methods for assembly BGCs without refactoring specialize for cloning small fragment number but large fragment size, including Gibson [44], Cas9-facilitated homologous recombination assembly (CasHRA) method [65], and transformation-associated recombination (TAR) [47,66]. Gibson is widely used because of its simple operation and seamless splicing, and it could realize the assemble of 4 large fragments with the sizes over 100 kb [44]. Compared with the in vitro Gibson assembly, the efficient assembly method of multiple large DNA fragments based on the principle of homologous recombination in vivo has also been popular in commonly used microbial cell factories, including *Saccharomyces cerevisiae*, *Escherichia coli* and *Bacillus subtilis*. For example, the Cas9-facilitated homologous recombination assembly (CasHRA) method has been successfully used in the assembly of large fragments of *E. coli* with a total length of 1.05 Mb, which includes 449 essential genes and 267 growth-related genes [65]. Similarly, TAR has also been used to identify several novel NPs, including orphan cosmomycin [67], thiostreptamide and scleroic acid [64]. The widely used methods for assembly BGCs with the need for refactoring features in cloning multiple fragments with high efficiency and fidelity number, including the Golden Gate assembly [41] and LCR assembly[46]. Golden Gate is based on type IIs restriction endonucleases, which can realize combinatorial assembly of 27 components by constructing modular libraries without leaving "scar" sequences [41]. LCR assembly method can realize 20 DNA fragments assembly in one step by using single-stranded bridging oligos between two adjacent fragments of DNA [46]. Combining CRISPR technology with Golden Gate, Gibson or TAR methods has significantly improved the assembly efficiency of large fragments and the size of DNA fragments (1.55 Mb) [31,49], it will be of great significance to develop more vectors and methods to clone large eukaryotic gene clusters for heterologous expression [68].

4. The test stage for natural product discovery

The commonly used method for studying NPs is to determine the biologically active compounds from the "crude" extract of the fermentation broth, and then fractionate to further separate them. Since this method is time-consuming and labor-intensive, automated liquid handling systems have been developed for high-throughput screening of library-based pre-fractionated crude extracts. Moreover, metabolomics analysis has been developed for simultaneously analyzing large numbers of metabolites in biological samples. The isomers present in NP extracts can be analyzed using NMR spectroscopy, high resolution mass spectrometry (HRMS), and the LC-HRMS method [69,70]. The advancement of analytical instruments used in NP research, coupled with annotation and calculation methods that can analyze putative NP structures [71], makes the "omics" method more effective .

In order to identify NPs with new structures and new activities, it is very important to deduplicate the metabolites in biological extracts by determining their molecular weights and molecular formulas, or comparing them with databases with detailed classification information. However, there are still challenges in data mining and the use of various workflows and web-based tools to identify metabolites with novel structures [72]. These metadata are sometimes difficult to query from literatures and databases. In this regard, a molecular network platform called the Global Natural Products Society (GNPS) has been developed by Dorrestein laboratories, and becomes the significant supplement to the toolbox of NPdiscovery [73]. This network has a large amount of MS/MS data and can visualize the gene cluster of corresponding molecules. Based on these methods, a large number of theoretical NP spectral databases have been created and applied to deduplication [74]. Similarly, METLIN is the platform containing a high-resolution MS/MS database to search for metabolites by analyzing similar structural features derived from reported compound [75]. Combining metabolome data and the results of biological activity analysis

in the extract can accelerate the identification of biologically active NPs in the extract [76]. Some chemometric methods, such as multivariate data analysis, can be used to correlate the signals detected in NMR and MS spectra to track active metabolites in complex mixtures. However, the current platforms still have limitations, such as the applicability of certain categories of NPs better than other categories, and ambiguous structure predictions and assignments over certain candidates. Efforts to solve these problems are underway [77], including covering the molecular network of large-scale NP extraction libraries with classification information to improve the credibility of annotation, and the development of comprehensive LC-MS/MS databases to support the NP analysis [78]. Acharya team used this method to characterize the NP-mediated interaction between different species [79]. In general, molecular networks are mainly used to strengthen the deduplication process to better determine the separation priority of unknown compounds and the elucidation of the relationship between NP analogs, and the rigorous structural elucidation of the NPs of interest also need to be taken seriously.

Because the production level of NP is relatively low in most cases, technologies have been developed to increase the detection specificity and tractability of current platforms. With the advancement of N-Methyl-2-Pyrrolidone (NMP) instruments and probe technology, it is possible to analyze a very small amount (less than 10 ug) of the analyte to determine the structure of NP [80]. Analyzing the response of biologically active compounds at the single cell level can also accelerate the discovery of NP drugs. In order to accurately determine the structure of small molecules, a microcrystal electron diffraction (MicroED) based on cryo-electron microscopy has recently been developed. Moreover, combining metabolomics data with transcriptome or proteomics data can also accelerate the identification of NPs [81].

The biological element such as enzyme, reporter gene, promoter or RBS, or logical circuits and modular metabolic pathways exist a lot of mutants after rational or irrational designing. Therefore, efficient, accurate and economical detection method is critical to the selection of the best biological components and combinations, such as expression, purification and activity test of enzyme elements, testing of transcription or translation components and unnatural pathways in vitro or in vivo, spatial and temporal regulation after cell factory modification and its effects on growth and metabolism. Traditional detection methods cannot meet the requirements for large-scale quantification about biological elements, logic circuits, metabolic and regulatory pathway combinations. The development and utilize a variety of high-throughput or automatic screening and detection technologies to improve the efficiency of testing is highly needed. Biology phenotype chip, microplate high-throughput screening, microfluidics, fluorescence activated droplet sorting system (FADS), raman light spectrum, fourier transform infrared spectroscopy or near-infrared spectroscopy and advanced spectral sensor have already been used for strain screening and phenotyping technology [82,83]. For example, Irish research team combined flow cytometry technology, single-cell chemical biology and cell barcoding with metabolomics to develop a high-throughput platform for bioactive metabolomics analysis. Using this platform, polyketides with new biological activities have been identified [84].

5. The learn stage for natural product discovery

The learning process is a critical part of the DBTL of synthetic biology, which provides important feedbacks for the improvement of the next cycle. The learning process involves data collection and integration, data analysis, result visualization, modeling analysis, including the different levels of omics analysis on gene-RNA-protein-metabolism-phenotype, construction of a knowledge map of genotype-phenotype and metabolic regulation network, etc. A large number of omics data and process detection data has been accumulated rapidly, and dedicated public database has provided great convenience for data sharing and storage (Table 1). Meanwhile, databases also provide automated data download programs or scripts, which greatly facilitates the data collection process (Table 1). At present, the data related to non-model microorganisms is developed far less than

that of model microorganisms, while the processing and learning of data related to model microorganisms can provide a valuable proof of concept for the study of non-model microorganisms.

The large amount of collected data can be analyzed using bioinformatics, artificial intelligence/ machine learning, especially deep learning, and mathematical models, such as genome-scale metabolic network models and whole-cell models [85,86]. Machine learning obtains capabilities from empirical data and has been widely used in computational biology, metabolic modeling, gene and protein network analysis to guide the design of microbial cell factories. For example, Alphafold developed by machine learning can be used for the de novo prediction of protein structure [87]. Genome-scale metabolic network model, as an effective system biology learning tool, has also been widely used in recent years to quickly understand metabolism of industrial microorganisms in vivo to find targets for modification [86], or to predict enzyme functions, and simulate cellular interactions. and related databases are gradually established, such as BioModels, BiGG Models and Kbase [88-90]. These results can be feedback to the design stage to further promote the development of synthetic biology through association, centralized query and visualization. Web-based interactive data visualization is pursued by scientific researchers for its convenience and practicality. At present, many databases provide web-based visualization results, such as BioCyc, which provides tens of thousands of sequenced microbial genomes and their metabolic pathways. The integration of analysis tools and result display, is an excellent example of web page visualization based on the web [54]. Whole-cell models have also been developed. For example, the whole-cell model of Mycoplasma genitalium [91] and a special visualization platform WholeCellViz [92] have been established to dynamically display its simulation process, intuitively understand the internal process, and facilitate learning and summary. The rules and conclusions summarized in this stage of “learning” can guide the modules of other stages of synthetic biology DBTL, and optimize the establishment of a more efficient and streamlined synthetic biology workflow for the construction and performance optimization of different chassis cells.

6. Conclusion and Future Perspectives

The rapid development of genome sequencing and genome mining have become important technologies for NP discovery and drug development [93]. The biosynthetic pathway of NPs involves delicate catalytic mechanisms, regulatory mechanisms and complex metabolic environments. Using bioinformatics analysis and calculation tools can carry out complex data analysis and design to develop NP biosynthetic pathways [94]. Vast genome sequencing data of non-cultivable and cultivable microorganisms, continuous improvement of omics analysis and machine learning technologies, and tools developments of systems biology and synthetic biology will continuously promote the discovery of new types of NPs with biological activity (Figure 2) [95]. In addition, the combination of artificial intelligence and computer methods has been successfully used to predict synthetic modules and pathways [17].

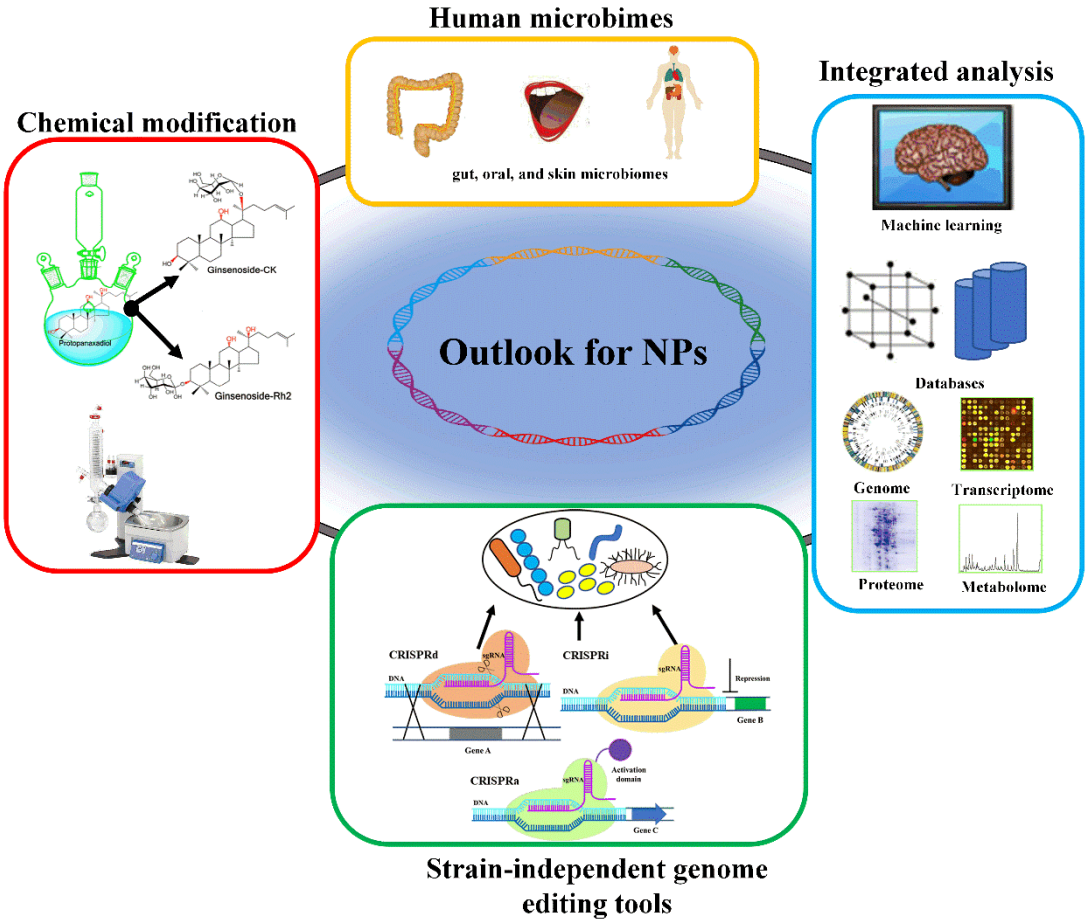

Figure 2. Future perspectives for NPs in drug discovery. In the future, NPs repertoire should be derived from human microbimes. Existing machine learning methods and databases should be optimized and integrated with the existing omics data, thereby improving the accuracy of the design and accelerating strain development. Strain-independent genome editing tools should be developed to enable efficient gene editing of nonconventional microbes as well as natural overproducers. Chemical modification should be used to optimize the properties of existing NPs.

In the method of excavating new structures and new activities of NPs by heterologous expression activation of silent gene clusters, the selection of a suitable host is also an important factor. Although many unconventional microorganisms have been successfully used in chassis cells, the available modification strategies and tools still lag far behind conventional chassis cells. The newly developed CRISPR system revolutionizes genome engineering of conventional and unconventional chassis cells, and the development of strain-independent genome editing tools plays a very important role in the mining of natural products (Figure 2). The research strategy of genome mining is constantly developing and becoming diversified. In addition to the classic genome mining strategy based on enzyme function analysis, new research strategies such as mining based on system

evolution, resistance gene, regulators, culture-independent single-cell and metagenomic have emerged one after another [22]. Nowadays, the acquisition of big data has become easier and easier, but the research on the biological significance behind these data remains a challenge. Effective analysis tools and algorithms will better guide the development of corresponding experiments.

In the mining of BGCs, metagenomics revealed that uncultured organisms can produce NPs with complex structures and biological activities. Using advanced methods and strategies to analyze the human microbiome (including gut microbes, oral microbes and skin microbes) with great biosynthetic potential will be an important research direction in the future [96] (Figure 2). In particular, the gut microbiota, which is considered to play an important role in health and disease, will be an emerging field for NP drug discovery [97].

Guided by genome mining, in-depth study of the biosynthetic mechanism, regulation mechanism, and key enzyme reaction mechanism is a necessary step to accelerate the research of NPs. Through genetic manipulation of metabolic pathways in microorganisms, not only more new structures and highly active compounds can be obtained, but also biosynthetic pathways of different kinds of compounds can be discovered, which will provide a brand-new way for the development of new drugs. In the field of drug discovery and development, combinatorial biosynthesis based on the biosynthesis and metabolism of NPs, can rationalize the genetic modification and reorganization of the biosynthetic pathway at the molecular level, and establish a library of NP analogs with complex structures [98]. Unmodified NPs usually exhibit suboptimal absorption, metabolism, excretion and toxicity properties, and superior analogues with improved pharmacological properties, such as higher specific activity, lower toxicity and better pharmacokinetics need to be acquired in order to yield valuable new drugs [93]. NP analogues can be accessed through the introduction of chemical modifications [99](Figure 2). From this, drugs with more clinical application values will be developed.

Author Contributions: Conceptualization, J.W. and Z.L.; writing—original draft preparation, J.W. and Z.L.; writing—review and editing, J.W., Z.L. and J.N.; funding acquisition, Z.L. and J.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program (2020YFA0712100), the National Natural Science Foundation of China (22078012), the Novo Nordisk Foundation (NNF10CC1016517), the Knut and Alice Wallenberg Foundation, and Beijing Advanced Innovation Center for Soft Matter Science and Engineering.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no competing financial interest.

References

1. Newman, D.J.; Cragg, G.M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770-803, doi:10.1021/acs.jnatprod.9b01285.
2. Cantrell, C.L.; Dayan, F.E.; Duke, S.O. Natural products as sources for new pesticides. *J. Nat. Prod.* **2012**, *75*, 1231-1242, doi:10.1021/np300024u.
3. Hernando-Amado, S.; Coque, T.M.; Baquero, F.; Martinez, J.L. Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat. Microbiol.* **2019**, *4*, 1432-1442, doi:10.1038/s41564-019-0503-9.
4. Zhang, M.Z.M.; Qiao, Y.; Ang, E.L.; Zhao, H.M. Using natural products for drug discovery: the impact of the genomics era. *Expert. Opin. Drug. Dis.* **2017**, *12*, 475-487, doi:10.1080/17460441.2017.1303478.
5. Pham, J.V.; Yilma, M.A.; Feliz, A.; Majid, M.T.; Maffetone, N.; Walker, J.R.; Kim, E.; Cho, H.J.; Reynolds, J.M.; Song, M.C.; et al. A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Front. Microbiol.* **2019**, *10*, 1404, doi:10.3389/fmicb.2019.01404.
6. Nielsen, J.C.; Grijseels, S.; Prigent, S.; Ji, B.; Dainat, J.; Nielsen, K.F.; Frisvad, J.C.; Workman, M.; Nielsen, J. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. *Nat. Microbiol.* **2017**, *2*, 17044, doi:10.1038/nmicrobiol.2017.44.
7. Rogers, J.K.; Church, G.M. Multiplexed Engineering in Biology. *Trends Biotechnol.* **2016**, *34*, 198-206, doi:10.1016/j.tibtech.2015.12.004.
8. Carbonell, P.; Currin, A.; Jervis, A.J.; Rattray, N.J.; Swainston, N.; Yan, C.; Takano, E.; Breitling, R. Bioinformatics for the synthetic biology of natural products: integrating across the Design-Build-Test cycle. *Nat. Prod. Rep.* **2016**, *33*, 925-932, doi:10.1039/c6np00018e.
9. Casini, A.; Storch, M.; Baldwin, G.S.; Ellis, T. Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 568-576, doi:10.1038/nrm4014.
10. Csorgo, B.; Nyerges, A.; Posfai, G.; Feher, T. System-level genome editing in microbes. *Curr. Opin. Microbiol.* **2016**, *33*, 113-122, doi:10.1016/j.mib.2016.07.005.
11. Smanski, M.J.; Bhatia, S.; Zhao, D.; Park, Y.; L, B.A.W.; Giannoukos, G.; Ciulla, D.; Busby, M.; Calderon, J.; Nicol, R.; et al. Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **2014**, *32*, 1241-1249, doi:10.1038/nbt.3063.
12. Chae, T.U.; Choi, S.Y.; Kim, J.W.; Ko, Y.S.; Lee, S.Y. Recent advances in systems metabolic engineering tools and strategies. *Curr. Opin. Biotechnol.* **2017**, *47*, 67-82, doi:10.1016/j.copbio.2017.06.007.
13. Liu, R.; Bassalo, M.C.; Zeitoun, R.I.; Gill, R.T. Genome scale engineering techniques for metabolic engineering. *Metab. Eng.* **2015**, *32*, 143-154, doi:10.1016/j.ymben.2015.09.013.
14. Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **2017**, *45*, D380-D388, doi:10.1093/nar/gkw952.
15. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353-D361, doi:10.1093/nar/gkw1092.
16. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2016**, *44*, D471-480, doi:10.1093/nar/gkv1164.
17. Xu, X.; Liu, Y.; Du, G.; Ledesma-Amaro, R.; Liu, L. Microbial Chassis Development for Natural Product Biosynthesis. *Trends Biotechnol.* **2020**, *38*, 779-796, doi:10.1016/j.tibtech.2020.01.002.
18. Ziemert, N.; Alanjary, M.; Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **2016**, *33*, 988-1005, doi:10.1039/c6np00025h.

19. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279-285, doi:10.1093/nar/gkv1344.
20. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H.U.; Bruccoleri, R.; Lee, S.Y.; Fischbach, M.A.; Muller, R.; Wohlleben, W.; et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **2015**, *43*, W237-243, doi:10.1093/nar/gkv437.
21. Machado, H.; Tuttle, R.N.; Jensen, P.R. Omics-based natural product discovery and the lexicon of genome mining. *Curr. Opin. Microbiol.* **2017**, *39*, 136-142, doi:10.1016/j.mib.2017.10.025.
22. Scherlach, K.; Hertweck, C. Mining and unearthing hidden biosynthetic potential. *Nat. Commun.* **2021**, *12*, 3864, doi:10.1038/s41467-021-24133-5.
23. Mir Mohseni, M.; Hover, T.; Barra, L.; Kaiser, M.; Dorrestein, P.C.; Dickschat, J.S.; Schaberle, T.F. Discovery of a Mosaic-Like Biosynthetic Assembly Line with a Decarboxylative Off-Loading Mechanism through a Combination of Genome Mining and Imaging. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 13611-13614, doi:10.1002/anie.201606655.
24. Vila-Farres, X.; Chu, J.; Inoyama, D.; Ternei, M.A.; Lemetre, C.; Cohen, L.J.; Cho, W.; Reddy, B.V.; Zebroski, H.A.; Freundlich, J.S.; et al. Antimicrobials Inspired by Nonribosomal Peptide Synthetase Gene Clusters. *J. Am. Chem. Soc.* **2017**, *139*, 1404-1407, doi:10.1021/jacs.6b11861.
25. Cox, G.; Wright, G.D. Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *Int. J. Med. Microbiol.* **2013**, *303*, 287-292, doi:10.1016/j.ijmm.2013.02.009.
26. Medema, M.H.; Cimermancic, P.; Sali, A.; Takano, E.; Fischbach, M.A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **2014**, *10*, e1004016, doi:10.1371/journal.pcbi.1004016.
27. Katz, L.; Baltz, R.H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **2016**, *43*, 155-176, doi:10.1007/s10295-015-1723-5.
28. Hutchison, C.A., 3rd; Venter, J.C. Single-cell genomics. *Nat. Biotechnol.* **2006**, *24*, 657-658, doi:10.1038/nbt0606-657.
29. Hedlund, B.P.; Dodsworth, J.A.; Murugapiran, S.K.; Rinke, C.; Woyke, T. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles* **2014**, *18*, 865-875, doi:10.1007/s00792-014-0664-7.
30. Dodsworth, J.A.; Blainey, P.C.; Murugapiran, S.K.; Swingley, W.D.; Ross, C.A.; Tringe, S.G.; Chain, P.S.; Scholz, M.B.; Lo, C.C.; Raymond, J.; et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat. Commun.* **2013**, *4*, 1854, doi:10.1038/ncomms2884.
31. Kuwahara, H.; Alazmi, M.; Cui, X.; Gao, X. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. *Nucleic Acids Res.* **2016**, *44*, W217-225, doi:10.1093/nar/gkw342.
32. Yang, P.; Wang, J.; Pang, Q.; Zhang, F.; Wang, J.; Wang, Q.; Qi, Q. Pathway optimization and key enzyme evolution of N-acetylneuraminate biosynthesis using an in vivo aptazyme-based biosensor. *Metab. Eng.* **2017**, *43*, 21-28, doi:10.1016/j.ymben.2017.08.001.
33. Bravo-Rodriguez, K.; Ismail-Ali, A.F.; Klopries, S.; Kushnir, S.; Ismail, S.; Fansa, E.K.; Wittinghofer, A.; Schulz, F.; Sanchez-Garcia, E. Predicted incorporation of non-native substrates by a polyketide synthase yields bioactive natural product derivatives. *ChemBioChem* **2014**, *15*, 1991-1997, doi:10.1002/cbic.201402206.
34. McDaniel, R.; Thamchaipenet, A.; Gustafsson, C.; Fu, H.; Betlach, M.; Ashley, G. Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 1846-1851, doi:10.1073/pnas.96.5.1846.
35. Nunez, L.E.; Nybo, S.E.; Gonzalez-Sabin, J.; Perez, M.; Menendez, N.; Brana, A.F.; Shaaban, K.A.; He, M.; Moris, F.; Salas, J.A.; et al. A novel mithramycin analogue with high antitumor activity and less toxicity generated by combinatorial biosynthesis. *J. Med. Chem.* **2012**, *55*, 5813-5825, doi:10.1021/jm300234t.

36. Doroghazi, J.R.; Albright, J.C.; Goering, A.W.; Ju, K.S.; Haines, R.R.; Tchalukov, K.A.; Labeda, D.P.; Kelleher, N.L.; Metcalf, W.W. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **2014**, *10*, 963-968, doi:10.1038/nchembio.1659.
37. M20; Liu, Q.; Yang, Q.; Sun, W.; Vogel, P.; Heydorn, W.; Yu, X.Q.; Hu, Z.; Yu, W.; Jonas, B.; et al. Discovery and characterization of novel tryptophan hydroxylase inhibitors that selectively inhibit serotonin synthesis in the gastrointestinal tract. *J. Pharmacol. Exp. Ther.* **2008**, *325*, 47-55, doi:10.1124/jpet.107.132670.
38. Wang, H.H.; Isaacs, F.J.; Carr, P.A.; Sun, Z.Z.; Xu, G.; Forest, C.R.; Church, G.M. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **2009**, *460*, 894-898, doi:10.1038/nature08187.
39. Zheng, Y.; Li, J.; Wang, B.; Han, J.; Hao, Y.; Wang, S.; Ma, X.; Yang, S.; Ma, L.; Yi, L.; et al. Endogenous Type I CRISPR-Cas: From Foreign DNA Defense to Prokaryotic Engineering. *Front. Bioeng. Biotechnol.* **2020**, *8*, 62, doi:10.3389/fbioe.2020.00062.
40. Zheng, Y.; Han, J.; Wang, B.; Hu, X.; Li, R.; Shen, W.; Ma, X.; Ma, L.; Yi, L.; Yang, S.; et al. Characterization and repurposing of the endogenous Type I-F CRISPR-Cas system of Zymomonas mobilis for genome engineering. *Nucleic Acids Res.* **2019**, *47*, 11461-11475, doi:10.1093/nar/gkz940.
41. Engler, C.; Gruetzner, R.; Kandzia, R.; Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **2009**, *4*, e5553, doi:10.1371/journal.pone.0005553.
42. Taylor, G.M.; Mordaka, P.M.; Heap, J.T. Start-Stop Assembly: a functionally scarless DNA assembly system optimized for metabolic engineering. *Nucleic Acids Res.* **2019**, *47*, e17, doi:10.1093/nar/gky1182.
43. Jeong, J.Y.; Yim, H.S.; Ryu, J.Y.; Lee, H.S.; Lee, J.H.; Seen, D.S.; Kang, S.G. One-step sequence- and ligation-independent cloning as a rapid and versatile cloning method for functional genomics studies. *Appl. Environ. Microbiol.* **2012**, *78*, 5440-5443, doi:10.1128/AEM.00844-12.
44. Gibson, D.G.; Young, L.; Chuang, R.Y.; Venter, J.C.; Hutchison, C.A., 3rd; Smith, H.O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **2009**, *6*, 343-345, doi:10.1038/nmeth.1318.
45. Xia, Y.; Li, K.; Li, J.; Wang, T.; Gu, L.; Xun, L. T5 exonuclease-dependent assembly offers a low-cost method for efficient cloning and site-directed mutagenesis. *Nucleic Acids Res.* **2019**, *47*, e15, doi:10.1093/nar/gky1169.
46. de Kok, S.; Stanton, L.H.; Slaby, T.; Durot, M.; Holmes, V.F.; Patel, K.G.; Platt, D.; Shapland, E.B.; Serber, Z.; Dean, J.; et al. Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* **2014**, *3*, 97-106, doi:10.1021/sb4001992.
47. Kim, J.H.; Dilthey, A.T.; Nagaraja, R.; Lee, H.S.; Koren, S.; Dudekula, D.; Wood Iii, W.H.; Piao, Y.; Ogurtsov, A.Y.; Utani, K.; et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res.* **2018**, *46*, 6712-6725, doi:10.1093/nar/gky442.
48. Jiang, W.; Zhu, T.F. Targeted isolation and cloning of 100-kb microbial genomic sequences by Cas9-assisted targeting of chromosome segments. *Nat. Protoc.* **2016**, *11*, 960-975, doi:10.1038/nprot.2016.055.
49. Wang, K.; de la Torre, D.; Robertson, W.E.; Chin, J.W. Programmed chromosome fission and fusion enable precise large-scale genome rearrangement and assembly. *Science* **2019**, *365*, 922-926, doi:10.1126/science.aay0737.
50. Sun, N.; Zhao, H. Transcription activator-like effector nucleases (TALENs): a highly efficient and versatile tool for genome editing. *Biotechnol. Bioeng.* **2013**, *110*, 1811-1821, doi:10.1002/bit.24890.
51. Peters, J.M.; Colavin, A.; Shi, H.; Czarny, T.L.; Larson, M.H.; Wong, S.; Hawkins, J.S.; Lu, C.H.S.; Koo, B.M.; Marta, E.; et al. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell* **2016**, *165*, 1493-1506, doi:10.1016/j.cell.2016.05.003.
52. Kim, S.K.; Han, G.H.; Seong, W.; Kim, H.; Kim, S.W.; Lee, D.H.; Lee, S.G. CRISPR interference-guided balancing of a biosynthetic mevalonate pathway increases terpenoid production. *Metab. Eng.* **2016**, *38*, 228-240, doi:10.1016/j.ymben.2016.08.006.
53. Lian, J.; HamediRad, M.; Hu, S.; Zhao, H. Combinatorial metabolic engineering using an orthogonal tri-functional CRISPR system. *Nat. Commun.* **2017**, *8*, 1688, doi:10.1038/s41467-017-01695-x.

54. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; Kothari, A.; Keseler, I.M.; Krummenacker, M.; Midford, P.E.; Ong, Q.; et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **2019**, *20*, 1085-1093, doi:10.1093/bib/bbx085.
55. Nara, A.; Hashimoto, T.; Komatsu, M.; Nishiyama, M.; Kuzuyama, T.; Ikeda, H. Characterization of bafilomycin biosynthesis in Kitasatospora setae KM-6054 and comparative analysis of gene clusters in Actinomycetales microorganisms. *J. Antibiot.* **2017**, *70*, 616-624, doi:10.1038/ja.2017.33.
56. Lin, Z.; Nielsen, J.; Liu, Z. Bioprospecting Through Cloning of Whole Natural Product Biosynthetic Gene Clusters. *Front. Bioeng. Biotechnol.* **2020**, *8*, 526, doi:10.3389/fbioe.2020.00526.
57. Bok, J.W.; Ye, R.; Clevenger, K.D.; Mead, D.; Wagner, M.; Krerowicz, A.; Albright, J.C.; Goering, A.W.; Thomas, P.M.; Kelleher, N.L.; et al. Fungal artificial chromosomes for mining of the fungal secondary metabolome. *BMC Genom.* **2015**, *16*, 343, doi:10.1186/s12864-015-1561-x.
58. Karas, B.J.; Suzuki, Y.; Weyman, P.D. Strategies for cloning and manipulating natural and synthetic chromosomes. *Chromosome Res.* **2015**, *23*, 57-68, doi:10.1007/s10577-014-9455-3.
59. Zhang, J.J.; Tang, X.; Moore, B.S. Genetic platforms for heterologous expression of microbial natural products. *Nat. Prod. Rep.* **2019**, *36*, 1313-1332, doi:10.1039/c9np00025a.
60. Cobb, R.E.; Ning, J.C.; Zhao, H. DNA assembly techniques for next-generation combinatorial biosynthesis of natural products. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 469-477, doi:10.1007/s10295-013-1358-3.
61. Hsu, P.D.; Lander, E.S.; Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **2014**, *157*, 1262-1278, doi:10.1016/j.cell.2014.05.010.
62. Wang, H.; Li, Z.; Jia, R.; Yin, J.; Li, A.; Xia, L.; Yin, Y.; Muller, R.; Fu, J.; Stewart, A.F.; et al. ExoCET: exonuclease in vitro assembly combined with RecET recombination for highly efficient direct DNA cloning from complex genomes. *Nucleic Acids Res.* **2018**, *46*, e28, doi:10.1093/nar/gkx1249.
63. Song, C.; Luan, J.; Cui, Q.; Duan, Q.; Li, Z.; Gao, Y.; Li, R.; Li, A.; Shen, Y.; Li, Y.; et al. Enhanced Heterologous Spinosad Production from a 79-kb Synthetic Multioperon Assembly. *ACS Synth. Biol.* **2019**, *8*, 137-147, doi:10.1021/acssynbio.8b00402.
64. Alberti, F.; Leng, D.J.; Wilkening, I.; Song, L.; Tosin, M.; Corre, C. Triggering the expression of a silent gene cluster from genetically intractable bacteria results in scleric acid discovery. *Chem. Sci.* **2019**, *10*, 453-463, doi:10.1039/c8sc03814g.
65. Zhou, J.; Wu, R.; Xue, X.; Qin, Z. CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA. *Nucleic Acids Res.* **2016**, *44*, e124, doi:10.1093/nar/gkw475.
66. Wang, J.W.; Wang, A.; Li, K.; Wang, B.; Jin, S.; Reiser, M.; Lockey, R.F. CRISPR/Cas9 nuclease cleavage combined with Gibson assembly for seamless cloning. *BioTechniques* **2015**, *58*, 161-170, doi:10.2144/000114261.
67. Larson, C.B.; Crusemann, M.; Moore, B.S. PCR-Independent Method of Transformation-Associated Recombination Reveals the Cosmomycin Biosynthetic Gene Cluster in an Ocean Streptomycete. *J. Nat. Prod.* **2017**, *80*, 1200-1204, doi:10.1021/acs.jnatprod.6b01121.
68. Liao, L.; Su, S.; Zhao, B.; Fan, C.; Zhang, J.; Li, H.; Chen, B. Biosynthetic Potential of a Novel Antarctic Actinobacterium Marisediminicola antarctica ZS314(T) Revealed by Genomic Data Mining and Pigment Characterization. *Mar. Drugs.* **2019**, *17*, doi:10.3390/md17070388.
69. Stavrianidi, A. A classification of liquid chromatography mass spectrometry techniques for evaluation of chemical composition and quality control of traditional medicines. *J. Chromatogr. A.* **2020**, *1609*, 460501, doi:10.1016/j.chroma.2019.460501.
70. Wolfender, J.L.; Marti, G.; Thomas, A.; Bertrand, S. Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A.* **2015**, *1382*, 136-164, doi:10.1016/j.chroma.2014.10.091.
71. Allard, P.M.; Genta-Jouve, G.; Wolfender, J.L. Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Curr. Opin. Chem. Biol.* **2017**, *36*, 40-49, doi:10.1016/j.cbpa.2016.12.022.

72. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S.S.; Wohlgemuth, G.; Barupal, D.K.; Showalter, M.R.; Arita, M.; et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **2018**, *37*, 513-532, doi:10.1002/mas.21535.
73. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kapono, C.A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828-837, doi:10.1038/nbt.3597.
74. Allard, P.M.; Peresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V.C.; Roussi, F.; Litaudon, M.; Wolfender, J.L. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal. Chem.* **2016**, *88*, 3317-3323, doi:10.1021/acs.analchem.5b04804.
75. Fox Ramos, A.E.; Pavesi, C.; Litaudon, M.; Dumontet, V.; Poupon, E.; Champy, P.; Genta-Jouve, G.; Beniddir, M.A. CANPA: Computer-Assisted Natural Products Anticipation. *Anal. Chem.* **2019**, *91*, 11247-11252, doi:10.1021/acs.analchem.9b02216.
76. M87; Schreier, B.; Stumpp, C.; Wiesner, S.; Hocker, B. Computational design of ligand binding is not a solved problem. *Proc Natl Acad Sci U S A* **2009**, *106*, 18491-18496, doi:10.1073/pnas.0907950106.
77. da Silva, R.R.; Wang, M.; Nothias, L.F.; van der Hooft, J.J.J.; Caraballo-Rodriguez, A.M.; Fox, E.; Balunas, M.J.; Klassen, J.L.; Lopes, N.P.; Dorrestein, P.C. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **2018**, *14*, e1006089, doi:10.1371/journal.pcbi.1006089.
78. Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; Bisson, J.; Bagheri, M.; Saesong, T.; Ebrahimi, S.N.; Ingkaninan, K.; Wolfender, J.L.; Allard, P.M. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Front. Plant. Sci.* **2019**, *10*, 1329, doi:10.3389/fpls.2019.01329.
79. M89; Park, S.; Kang, K.; Lee, S.W.; Ahn, M.J.; Bae, J.M.; Back, K. Production of serotonin by dual expression of tryptophan decarboxylase and tryptamine 5-hydroxylase in Escherichia coli. *Appl. Microbiol. Biotechnol.* **2011**, *89*, 1387-1394, doi:10.1007/s00253-010-2994-4.
80. M91; Dempsey, D.R.; Jeffries, K.A.; Handa, S.; Carpenter, A.M.; Rodriguez-Ospina, S.; Breydo, L.; Merkler, D.J. Mechanistic and Structural Analysis of a Drosophila melanogaster Enzyme, Arylalkylamine N-Acetyltransferase Like 7, an Enzyme That Catalyzes the Formation of N-Acetylarylalkylamides and N-Acetylhistamine. *Biochemistry* **2015**, *54*, 2644-2658, doi:10.1021/acs.biochem.5b00113.
81. Weber, T.; Kim, H.U. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synth. Syst. Biotechnol.* **2016**, *1*, 69-79, doi:10.1016/j.synbio.2015.12.002.
82. Ma, X.; Huo, Y.X. The application of microfluidic-based technologies in the cycle of metabolic engineering. *Synth. Syst. Biotechnol.* **2016**, *1*, 137-142, doi:10.1016/j.synbio.2016.09.004.
83. Sarnaik, A.; Liu, A.; Nielsen, D.; Varman, A.M. High-throughput screening for efficient microbial biotechnology. *Curr. Opin. Biotechnol.* **2020**, *64*, 141-150, doi:10.1016/j.copbio.2020.02.019.
84. M90; Williams, B.B.; Van Benschoten, A.H.; Cimermancic, P.; Donia, M.S.; Zimmermann, M.; Taketani, M.; Ishihara, A.; Kashyap, P.C.; Fraser, J.S.; et al. Discovery and characterization of gut microbiota decarboxylases that can produce the neurotransmitter tryptamine. *Cell Host Microbe* **2014**, *16*, 495-503, doi:10.1016/j.chom.2014.09.001.
85. Zhou, Y.; Li, G.; Dong, J.; Xing, X.H.; Dai, J.; Zhang, C. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae. *Metab. Eng.* **2018**, *47*, 294-302, doi:10.1016/j.ymben.2018.03.020.
86. Oyetunde, T.; Bao, F.S.; Chen, J.W.; Martin, H.G.; Tang, Y.J. Leveraging knowledge engineering and machine learning for microbial bio-manufacturing. *Biotechnol. Adv.* **2018**, *36*, 1308-1315, doi:10.1016/j.biotechadv.2018.04.008.
87. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706-710, doi:10.1038/s41586-019-1923-7.

88. Li, C.; Donizelli, M.; Rodriguez, N.; Dharuri, H.; Endler, L.; Chelliah, V.; Li, L.; He, E.; Henry, A.; Stefan, M.I.; et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* **2010**, *4*, 92, doi:10.1186/1752-0509-4-92.
89. Arkin, A.P.; Cottingham, R.W.; Henry, C.S.; Harris, N.L.; Stevens, R.L.; Maslov, S.; Dehal, P.; Ware, D.; Perez, F.; Canon, S.; et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **2018**, *36*, 566-569, doi:10.1038/nbt.4163.
90. Norsigian, C.J.; Pusarla, N.; McConn, J.L.; Yurkovich, J.T.; Drager, A.; Palsson, B.O.; King, Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **2020**, *48*, D402-D406, doi:10.1093/nar/gkz1054.
91. Karr, J.R.; Sanghvi, J.C.; Macklin, D.N.; Gutschow, M.V.; Jacobs, J.M.; Bolival, B., Jr.; Assad-Garcia, N.; Glass, J.I.; Covert, M.W. A whole-cell computational model predicts phenotype from genotype. *Cell* **2012**, *150*, 389-401, doi:10.1016/j.cell.2012.05.044.
92. Lee, R.; Karr, J.R.; Covert, M.W. WholeCellViz: data visualization for whole-cell models. *BMC Bioinformatics* **2013**, *14*, 253, doi:10.1186/1471-2105-14-253.
93. Atanasov, A.G.; Zotchev, S.B.; Dirsch, V.M.; International Natural Product Sciences, T.; Supuran, C.T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **2021**, *20*, 200-216, doi:10.1038/s41573-020-00114-z.
94. Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20, doi:10.1186/s13321-020-00424-9.
95. Palazzotto, E.; Weber, T. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. *Curr. Opin. Microbiol.* **2018**, *45*, 109-116, doi:10.1016/j.mib.2018.03.004.
96. Zipperer, A.; Konnerth, M.C.; Laux, C.; Berscheid, A.; Janek, D.; Weidenmaier, C.; Burian, M.; Schilling, N.A.; Slavetinsky, C.; Marschal, M.; et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **2016**, *535*, 511-516, doi:10.1038/nature18634.
97. Lynch, S.V.; Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N. Engl. J. Med.* **2016**, *375*, 2369-2379, doi:10.1056/NEJMra1600266.
98. Shaeer, K.M.; Zmarlicka, M.T.; Chahine, E.B.; Piccicacco, N.; Cho, J.C. Plazomicin: A Next-Generation Aminoglycoside. *Pharmacotherapy* **2019**, *39*, 77-93, doi:10.1002/phar.2203.
99. Tevyashova, A.N.; Olsufyeva, E.N.; Solovieva, S.E.; Printsevskaya, S.S.; Reznikova, M.I.; Trenin, A.S.; Galatenko, O.A.; Treshalin, I.D.; Pereverzeva, E.R.; Mirchink, E.P.; et al. Structure-antifungal activity relationships of polyene antibiotics of the amphotericin B group. *Antimicrob. Agents Chemother.* **2013**, *57*, 3815-3822, doi:10.1128/AAC.00270-13.