Liquid chromatography mass spectrometry for analysis of microbial metabolites

Klitgaard, Andreas; Nielsen, Kristian Fog; Andersen, Mikael Rørdam; Frisvad, Jens Christian

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Klitgaard, A., Nielsen, K. F., Andersen, M. R., & Frisvad, J. C. (2015). Liquid chromatography mass spectrometry for analysis of microbial metabolites. Kgs. Lyngby: Department of Systems Biology, Technical University of Denmark.

Liquid Chromatography Mass Spectrometry for Analysis of Microbial Secondary Metabolites

Andreas Klitgaard

PhD thesis

Technical University of Denmark

Department of Systems Biology

Eukaryotic Biotechnology

Preface

This study was carried out at the section for Eukaryotic Biotechnology, Department of Systems Biology, Technical University of Denmark (DTU), in the period 1 December 2011 to 30 November 2014. The study was supported by DTU and a grant (09-064967) from the Danish Council for Independent Research, Technology, and Production Sciences.

I would like to thank my three supervisors Kristian Fog Nielsen, Mikael Rørdam Andersen, and Jens Christian Frisvad. They have all three provided invaluable guidance, inspiration, and support whenever needed.

Further, I would also like to than Rasmus John Nordmand Frandsen, Jakob Blæsbjerg Nielsen, and Thomas Ostenfeld Larsen for good collaborations on several different projects. I would also like to thank Jakob and Thomas for their music recommendations, however crazy they might ever be.

Another thanks goes to Maria Månsson for introducing my to the world of marine bacteria, Lene Maj Petersen for all of the advice in and out of the lab, and Dorte Koefoed Holm for having the patience to repeatedly explaining what a gene is.

A big thank you goes to all the people at the center formerly known as CMB. It has been a fantastic three years of laughs, BBQ'ing, and occasional scientific discussions. I will miss you all very much.

Finally, I would like to thank my friends and family for helping and supporting me during my studies. I would especially like to thank Dorte - you made it all possible.

Summary

Filamentous fungi serve a very important role in Nature where they break down organic matter, releasing nutrients that can be used by other organisms. Fungi and other microorganisms also produce a wide array of bioactive compounds, the secondary metabolites(SMs), used for such diverse roles as signaling, defense, or pigmentation. Compounds from microorganisms have a dual impact on human society: they have been used as drugs, or as inspiration for the development of drugs for centuries. However, fungal infection of crops and the subsequent contamination by mycotoxins, continue to pose a threat to human health. Because of this, methods for detection and analysis of these compounds are vital. Estimates suggest that there are around 1.5 million different fungal species on Earth, dwarfing the number of plants estimated to 300,000, meaning that there potentially are many more interesting compounds are still to be discovered.

The main analytical technique used to investigate production of products from these diverse organisms is liquid-chromatography coupled to mass spectrometry (LC-MS). With the development of new and improved analytical instrumentation for chemical analysis, the time needed to perform a single analytical run has decreased, while the amount of information obtained from each of these analytical runs has increased drastically. Consequently, the limiting step in chemical analysis of a microorganism is no longer the analytical run itself, but rather analysis of the resulting data. Classical methods for manual interpretation of one single data file at a time are not sufficient to cope with this influx of data. Hence, there is a need for development of new methods for data analysis to extract valuable information in the data, and also speeding up the data analysis itself.

A prime goal of my PhD study was to develop methods that allow for high-throughput analysis of metabolite extracts from filamentous fungi and other microorganisms, and to reduce the time spent on manual interpretation of LC-MS data. This lead to development of a method that utilizes compound libraries to screen the recorded LC-MS data and annotate known compounds, a process we have named *aggressive dereplication*. By overlaying automatically generated extracted-ion chromatograms from detected compounds on the base peak chromatogram, all major potentially novel peaks can be visualized, allowing for fast dereplication of samples. This was further developed to include the use of recorded MS/MS data, allowing for greater confidence in matched compounds.

Another goal of the present study has been to develop methods that allow for faster coupling of SMs to their biosynthetic genes, as coupling of genes to metabolites is of large commercial interest for production of the bioactive compounds of the future. One part of my study focused on identification and elucidation of the biosynthesis of a nonribosomal peptide (NRP) nidulanin A from *Aspergillus nidulans*. Although the study was successful several analogs were not structure elucidated due to very low production titers. Instead a novel approach was developed for probing the biosynthesis of NRPs using stable isotope labeled (SIL) amino acids and subsequent analysis by MS/MS. Recorded MS/MS data were analyzed using molecular networking, coupling together compounds that exhibit similar MS/MS spectra. The combination of stable isotope labeling and molecular networking proved very effective for detection of structurally related NRPs. Labeling alone aided in determining the cyclic-peptide sequence, and may be used to provide information on biosynthesis of bioactive compounds.

In another study, the combined approach of targeted analysis methods and SIL precursors was used to elucidate the biosynthesis of yanuthone D in *A. niger,* and to determine compounds biosynthesized from the same precursor. Further studies on the biosynthesis of polyketides were conducted using feeding studies with SIL precursor in order to determine advantages and disadvantages of the approach. This led to determination of the biosynthetic origin of several compounds in *Fusarium* including antibiotic Y, and tentative identification of an intermediate in its biosynthetic pathway. Last, benzoic acid was identified as the precursor to asperrubrol in *A. niger*.

Finally, I have developed an integrated approach to evaluate the biosynthetic richness in bacteria and mine the associated chemical diversity. Here, 13 strains related to the marine bacterial species *Pseudoalteromonas luteoviolacea* were investigated in an untargeted metabolomics experiment and the results were correlated to whole-genome sequences of the strains. We found that 30 % of all chemical features and 24 % of the biosynthetic genes were unique to a single strain, while only 2 % of the features and 7 % of the genes were shared between all. The list of chemical features, originally comprising 2,000 features, was reduced to 50 discriminating features using a genetic algorithm combined with support vector machine evaluation. These features were efficiently dereplicated by molecular networking, which lead to tentative identification of several known antibacterial compounds, some of which had not previously been described from this organism. By combining metabolomics and genomics data, it was possible to link metabolites to chemical pathways at a very early stage in the discovery process.

Based on these results, the data analysis methods and methodologies developed during these studies have shown to be very effective and applicable to metabolite analysis of a wide range of microorganisms, and not restricted to fungi. The developed methods have revealed new insights into microbial SMs, and it is clear that even more discoveries can be made using these methods.

Sammenfatning

Filamentøse svampe udfylder en meget vigtig rolle i naturen hvor de nedbryder organisk materiale og derved frigiver næringsstoffer, som kan udnyttes af andre organismer. Svampe og andre mikroorganismer producerer derudover en bred vifte af bioaktive stoffer, de såkaldte sekundære metabolitter. Disse udfylder forskellige rolle såsom signalering, forsvar eller pigmentering. Produkter fra mikroorganismer har en todelt indflydelse på det menneskelige samfund: de er blevet benyttet som lægemidler eller som inspiration til udviklingen af lægemidler i århundreder. Samtidigt udgør svampeinfektioner i afgrøder, og den efterfølgende kontaminering med mykotoxiner, en fortsat trussel mod menneskers helbred. På grund af dette er metoder til at detektere og analysere disse stoffer vitale. Det er blevet skønnet at der eksisterer omkring 1,5 millioner forskellige svampe arter på Jorden, hvilket langt overstiger det skønnede antal af planter på 300.000, og dette betyder at der potentielt stadig findes mange uopdagede biologisk interessante stoffer.

Den primære analyseteknik der benyttes til at undersøge produktionen af stoffer fra disse forskellige organismer er væskekromatografi kombineret med massespektrometri (LC-MS). Med udviklingen af nye og forbedrede analyseinstrumenter til kemisk analyse er selve analysetiden blevet reduceret mens mængden af information, der opnås fra hver af disse analytiske undersøgelser, er steget drastisk. Som en konsekvens af dette er det begrænsende trin i analysen af mikroorganismer ikke længere selve den kemiske analyse, men i stedet analyse af data. Klassiske metoder, hvor datafiler analyseres enkeltvis og manuelt, er ikke længere tilstrækkelige til at håndtere de stigende mængder data. Der er derfor nødvendigt at udvikle nye metoder til at udvinde værdifuld information fra data, samt at øge hastigheden hvormed data analyseres.

Et af hovedmålene med mit PhD studium var at udvikle metoder der tillader analyse af store mængder data fra metabolit-ekstrakter fra filamentøse svampe og andre mikroorganismer, samt at reducere den tid der bruges på manuel tolkning af LC-MS data. Dette ledte til udvikling af en metode, kaldet aggressiv dereplikering, der benytter sig af metabolit-biblioteker til at screene LC-MS, for derved at annotere kendte stoffer. Ved at overlejre base peak kromatogrammer kunne potentielt nye toppe derved visualiseres, hvilket tillod hurtig dereplikering af data. Metoden blev yderligere udviklet til at benytte tandem MS data (MS/MS), hvilket øgede tilliden til identifikationen af fundne stoffer.

Et andet mål med mit studie har været at udvikle metoder der gør det muligt at koble sekundære metabolitter til de biosyntetiske gener der er ansvarlige for deres produktion. Denne kobling af metabolitter til gener er af stor kommerciel interesse med henblik på fremtidig produktion af bioaktive stoffer. En del af mit studie var fokuseret på identifikation samt udredning af biosyntesen af det ikke-ribosomale peptid (NRP) nidulanin A fra *Aspergillus nidulans*. Flere analoger til nidulanin A blev også fundet, men disse kunne ikke strukturopklares da blev produceret i meget små mængder. I stedet blev en ny fremgangsmåde udviklet til at undersøge NRPer ved hjælp af stabile isotopmærkede (SIL) aminosyrer og efterfølgende MS/MS analyse. Optagne MS/MS spektre blev analyseret ved at danne et molekylært netværk, som grupperede stoffer der udviste samme MS/MS spektre. Kombinationen af SIL og molekylære netværk viste sig at være meget effektivt til detektion af strukturelt relaterede NRPer. Ved at udnytte mærkning alene var det muligt at bestemme

sekvensen af det cykliske peptid, og metoden kan benyttes til at undersøge biosyntesen af andre bioaktive stoffer.

I et andet studie blev målrettet analyse kombineret med SIL udgangsstoffer brugt til at bestemme biosyntesen af stoffet yanuthone D, som produceres af *A. niger*. Metoden blev ydermere anvendt til at identificere andre stoffer, som bliver biosyntetiseret fra samme udgangsstof. Yderligere studier af polyketider blev foretaget, igen med brug af SIL udgangsstoffer for at undersøge fordele og ulemper ved fremgangsmåden. Disse studier ledte til bestemmelse af det biosyntetiske ophav af flere stoffer fra *Fusarium*, blandt andet antibiotic Y, samt til en tentativ identifikation af et intermediat i antibiotic Ys biosyntese. Ydermere blev benzoesyre bestemt til at være udgangsstoffet for stoffet asperrubrol i *A. niger*.

Afslutningsvist blev en fremgangsmåde udviklet til at evaluere det biosyntetiske potentiale i bakterier samt undersøge den kemiske diversitet. Til dette studie blev 13 forskellige stammer, relateret til den marine bakterie *Pseudoalteromonas luteoviolacea*, undersøgt i et ikke-målrettet (untargeted) metabolomics eksperiment, hvorefter de kemiske data blev korrelerede med fuld-genom sekvenser fra stammerne. Vi fandt derved at 30 % af de kemiske egenskaber samt 24 % af de biosyntetiske gener var unikke for den enkelte stamme, mens kun 2 % af kemiske detaljer (features) samt 7 % af generne var fælles mellem alle stammerne. Den oprindelige liste af 2.000 kemiske features blev reduceret til 50 særligt beskrivende kemiske detaljer ved hjælp af en genetisk algoritme som blev evalueret ved hjælp af en support vector machine. Disse kemiske detaljer blev effektivt dereplikeret ved brug af et molekylært netværk, og ledte til identifikation af flere kendte antibakterielle stoffer, flere af hvilke ikke tidligere var bestemt fra denne organisme. Ved at kombinere metabolomics samt genom-data var det da muligt at koble metabolitter til deres biosyntese på et meget tidligt tidspunkt i opdagelsesprocessen.

På basis af de opnåede resultater, må det konkluderes af de udviklede metoder og metodikker er meget effektive samt anvendelige til analyse af metabolitter fra en bred vifte af mikroorganismer. De udviklede metoder har ledt til ny indsigt i mikrobielle sekundære metabolitter, og det står klart at stadig flere opdagelser kan gøres ved brug af disse metoder.

List of papers and other publications

Paper 1 **Klitgaard, A.**, Iversen, A., Andersen, M. R., Larsen, T. O., Frisvad, J. C., & Nielsen, K. F. (2014). Aggressive dereplication using UHPLC-DAD-QTOF: screening extracts for up to 3000 fungal secondary metabolites. Analytical and Bioanalytical Chemistry, 406(7), 1933–1943.

Paper 2 Kildgaard, S., Mansson, M., Dosen, I., **Klitgaard, A**., Frisvad, J. C., Larsen, T. O., & Nielsen, K. F. (2014). Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. Marine Drugs, 12(6), 3681–3705.

Paper 3 Holm, D. K., Petersen, L. M., **Klitgaard, A**., Knudsen, P. B., Jarczynska, Z. D., Nielsen, K. F., Mortensen, U. H. (2014). Molecular and Chemical Characterization of the Biosynthesis of the 6-MSA-Derived Meroterpenoid Yanuthone D in *Aspergillus niger*. Chemistry & Biology, 21(4), 1–11.

Paper 4 **Klitgaard, A.**, Frandsen, R. J. N., Holm, D. K., Knudsen, P. B., Frisvad, J. C., Nielsen, K. F. (2015). Combining UHPLC-high resolution MS and feeding of stable isotope labeled polyketide intermediates for linking precursors to end products. Journal of Natural Products.

Paper 5 Andersen, M. R., Nielsen, J. B., **Klitgaard, A**., Petersen, L. M., Zachariasen, M., Hansen, T. J., Mortensen, U. H. (2013). Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proceedings of the National Academy of Sciences of the United States of America, 110(1), E99–107.

Paper 6 **Klitgaard, A.**, Nielsen, J. B., Frandsen, R. J. N., Andersen, M. R., Nielsen, K.F. (2015). Combining stable isotope labeling and molecular networking for biosynthetic pathway characterization. Analytical Chemistry, 87(13), 6520-6526.

Paper 7 Månsson, M., Vynne, N. G., **Klitgaard, A.**, Nybo, J. L., Melchiorsen, J., Ziemert, N., Dorrestein, P. C., Andersen, M. R., Gram, L. (2014). Integrated Metabolomic and Genomic Mining of the Biosynthetic Potential of the Marine Bacterial *Pseudoalteromonas luteoviolacea* species. (**Draft**).

Other publications:

Poster presentation at the Metabolomics 2014 conference, Tsuruoka, 2012: *Study of the plasticity of secondary metabolites in the black Aspergilli using UHPLC-qTOF molecular networking*, **Klitgaard, A.**, Månsson, M., Gezgin, Y., Lamboni, Y., and Nielsen, K. F..

Oral presentation at the Danish Society of Mass Spectrometry yearly meeting, Svendborg, 2014: *UHPLC-HR MS and MS/MS in fungal biosynthetic pathway discovery*, **Klitgaard, A..**

Poster presentation at the Gordon Conference on Marine Natural Products, 2014, Ventur, CA: *Comparative Genomic and Metabolomic Analysis of Twelve Strains of Pseudoalteromonas luteoviolacea*, Månsson, Maria; Vynne, N. G., **Klitgaard, A.,** Melchiorsen, J., Dorrestein, P. C., Gram, L..

Poster presentation at the International Congress on Natural Products Research on Global Change, Natural Products and Human Health/8th Joint Meeting of AFERP, ASP, GA, PSE and SIF, New York, 2012: *Screening of Aspergillus nidulans metabolites from habitat mimicking media using LC-DAD-TOFMS system*, **Klitgaard, A.**, Holm, D. K., Frisvad, J. C., Nielsen, K. F..

Poster presentation at the International Congress on Natural Products Research on Global Change, Natural Products and Human Health/8th Joint Meeting of AFERP, ASP, GA, PSE and SIF, New York, 2012: *Elucidation of the biosynthesis of meroterpenoid yanuthone D in Aspergillus niger*, Holm, D. K., Petersen, L. M., **Klitgaard, A.**, Jarczynska, Z. D., Larsen, T. O., Mortensen, U. H..

Oral presentation at the conference Directing Biosynthesis III, 2012, Nottingham: *Chemical analysis of a genome wide polyketide synthase gene deletion library in Aspergillus nidulans*, Larsen, T. O., Klejnstrup, M. L. ; Nielsen, J. B. ; Holm, D. K., Petersen, L. M., **Klitgaard, A.**, ; Nielsen, K. F., Andersen, M. R., Mortensen, U. H..

Abbreviations

6-MSA	6-methyl salicylic acid
AA	Amino acid
ACP	Acyl carrier protein
antiSMASH	Antibiotics & secondary metabolite analysis shell
AT	Acyl transferase
Bell’s	Bell’s medium
CoA	Coenzyme A
Conc.	Concentration
DAD	Diode array detector
DFM	Defined *Fusarium* medium
EIC	Extracted ion chromatogram
eV	Electron Volt
FT-ICR	Fourier transform ion cyclotron resonance
FWHM	Full width at half maximum
GA	Genetic algorithm
GC	Gas chromatography
GnPS	Global natural products social molecular networking
HPLC	High pressure liquid chromatography
HR	High-resolution
ISCID	In-source collision-induced dissociation
KS	β-ketoacyl CoA synthase
LC-MS	Liquid chromatography – mass spectrometry
LTQ	Linear ion trap quadrupole
MFE	Molecular feature extraction
ND	No incorporation detected
NRP	Nonribosomal peptide

NRPS	Nonribosomal peptides synthase
OBU	Operational biosynthetic unit
Phe	Phenylalanine
PK	Polyketide
PKS	Polyketide synthase
Ppm	Parts-per-million
Q	Quadrupole
SIL	Stable isotope labeled
SILAA	Stable isotope labeled amino acid
SM	Secondary metabolite
SVM	Support vector machine
SWATH	Sequential windowed acquisition of all theoretical mass spectra
TOF	Time-of-flight
Trp	Tryptophan
Tyr	Tyrosine
UHPLC	Ultra high performance liquid chromatography
UV/Vis	Ultraviolet/visible light
Val	Valine

Table of Contents

1 Introduction

1.1 Introduction to work performed in the thesis

One of the primary aims of this thesis was to develop methods for high-throughput analysis of metabolite extracts from filamentous fungi and other microorganisms using liquid chromatography-mass spectrometry (LC-MS) for investigation of secondary metabolites (SMs), with a particular focus on reducing the amount of manual inspection of the resulting data. The second aim was to investigate the biosynthesis of selected SMs, and couple these to the biosynthetic genes responsible for their production. At present, only few fungal biosynthetic synthases have been linked to a product. Increasing the pool of links between synthase genes and their products will aid in future computational prediction of products from newly sequenced fungi. This knowledge will aid in identification of potential mycotoxins in food and feed, or could be used for identifying potential new drug candidates. Increasing the pool of links between synthase genes and their products will also aid in identification of conserved characteristics that are important for the specific activities displayed by the synthases. This knowledge may be used to engineer novel synthases that produce a compound of interest e.g. a drug candidate precursor with or without specific pharmacophores, or biologically active structural motifs. This also applies to elucidation of the specific biosynthetic steps involved in biosynthesis of a given compound, as many different reactions take place in order to synthesize fungal SMs from a given precursor. These reactions are catalyzed by tailoring enzymes, which are most often very substrate specific. Most tailoring enzymes can only be predicted by their overall activity e.g. oxidation, dehydration etc., however, enzymes within the same class can catalyze a multitude of different reactions, using different substrates. Increasing the pool of links between enzymes and substrates can lead to a more accurate prediction of activity, based on the enzyme secondary structure alone. This knowledge is invaluable for de novo design of novel drugs using a given precursor.

The work performed in this thesis has focused on three specific themes; targeted analysis, untargeted analysis, and isotopic labeling for the study of biosyntheses. Publications resulting from this work have been categorized according to the themes covered as illustrated in Figure 1.

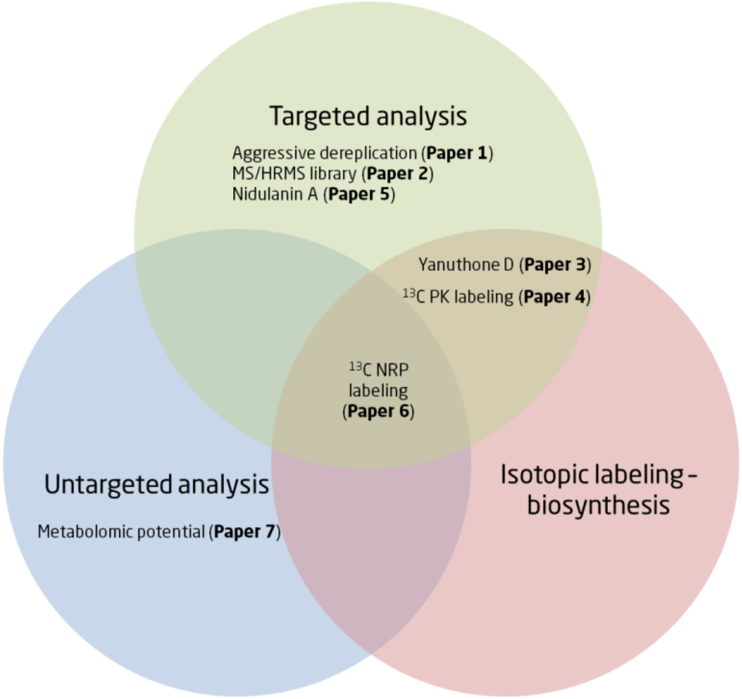

Targeted analysis
Aggressive dereplication (Paper 1)
MS/HRMS library (Paper 2)
Nidulanin A (Paper 5)
Yanuthone D (Paper 3)
13C PK labeling (Paper 4)
13C NRP labeling (Paper 6)
Untargeted analysis
Metabolomic potential (Paper 7)
Isotopic labeling - biosynthesis

Figure 1 - Venn diagram of my papers according to themes covered.

The results section has also been divided into these three sections; describing and discussing the results obtained through the use of these methods:

In section 2.1, cases for targeted analysis will be outlined, and the two methods developed for targeted analysis will be described and compared. Both methods were based on the use of compound libraries for fast screening of LC-MS data to identify compounds of interest.

Section 2.2 presents the methodologies developed for investigation of fungal metabolite biosynthesis using stable isotope labeled precursors, including investigation of PK- and NRP-derived metabolites. Data from these experiments were investigated using both targeted and untargeted analysis.

In section 2.3, an untargeted approach, developed for investigation of the chemical diversity of marine bacteria is presented. The developed metabolomics analysis was used to prioritize strains for further targeted investigation of metabolites.

Subsequently, perspectives on the development within the field of research and analysis methods are presented, and finally the overall results obtained in my study are summarized.

1.2 Importance of natural products

Filamentous fungi play an important role in Nature where they decompose organic matter releasing nutrients for themselves and for other organisms. Fungi are also hugely important in Nature because of the compounds that they produce, especially those referred to as secondary metabolites (SM). There is not one conclusive definition of a SM, however, one definition is that "a SM is a metabolite that is not essential for growth of the organism", in contrast to the primary metabolites. Still, as SMs seem to fulfill a multitude of different roles including signaling and regulation, defense against predators (Kempken and Rohlfs 2010), and protection against UV radiation, the definition of a SM could be expanded to "not being essential for growth in an ideal and uncontested environment" (Demain and Fang 2000). With such a broad spectrum of activities, it comes as no surprise that many pharmaceuticals are derived or partially derived from fungal SMs, including the famous antibacterial penicillins (Cragg and Newman 2013). In fact, in 1995 around 22% of the then known antibiotics could be produced by filamentous fungi (Adrio and Demain 2003). Other important compounds produced by fungi are the immunosuppressive agents cyclosporine (Borel et al. 1994) and mycophenolic acid, the cholesterol lowering statins (Endo 1985), as well as industrially important chemicals such as citric and maleic acid (Bennett and Klich 2003).

Unfortunately, not all compounds produced by fungi are beneficial to human health or industry. Numerous toxic compounds, also referred to as mycotoxins, are produced as well. Among some of the most well-known mycotoxins are the aflatoxins (Nesbitt et al. 1962) - the most carcinogenic compounds known, the ochratoxins (van der Merwe et al. 1965), trichothescenes (Bennett and Klich 2003; Frisvad et al. 2009), zearalenones (Christensen et al. 1965; Urry et al. 1966), and fumonisins (Bezuidenhout 1988). Fungi can also infect crops, leading to mycotoxins in the produce. This may result in adverse health effects in animals and humans because of the mycotoxins produced by the fungi, and further lead to severe economic loss in both the agricultural, feed, and food industry. Because of fungi's ability to produce beneficial as well as very toxic compounds, detection and identification of known compounds, as well as characterization of new compounds, is very important.

1.3 Biosynthesis

SMs are categorized based on their biosynthetic origin, where the major classes are the polyketides (PKs) (Hertweck 2009), nonribosomal peptides (NRPs) (Finking and Marahiel 2004), and terpenoids (Keller et al. 2005). They are all produced by synthases/synthetases encoded by genes that are often part of complex biosynthetic gene clusters, and many examples of mixed biosynthetic pathways of two or even all three are known. Examples of some of the different classes of fungal metabolites are shown in Figure 2, illustrating the diversity of fungal SMs.

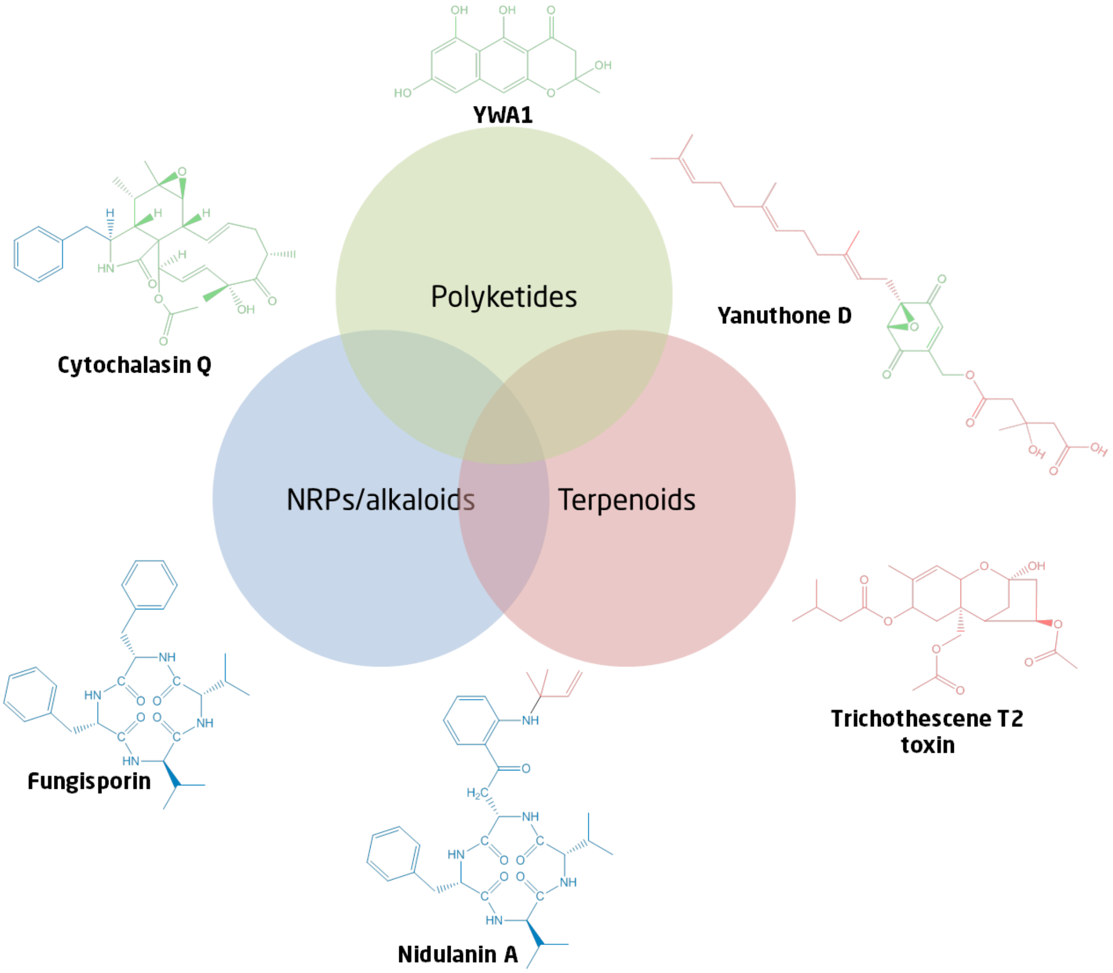


Figure 2 - Compounds representative of the three major biosynthetic classes as well as hybrids between the three, with atoms colored according to biosynthetic origin.

Several of the compounds in Figure 2 were investigated and will be presented in the results section of this thesis. During my studies I have primarily worked with compounds of PK and NRP origin, as well as hybrids such as the meroterpenoids. The focus has been on identifying biosynthetically related compounds, and development of methods for investigation of biosynthesis using LC-MS. As such, I have not focused on elucidation of the biosynthetic mechanisms involved in production of metabolites.

Coupling of biosynthetic genes to metabolite products has traditionally been a very labor intensive process. Currently, the process requires full genome sequenced organisms, and specially prepared fungal or bacterial strains that allow for easy gene deletion and up-regulation. In my studies, I have worked on development of methods for investigation of biosynthesis of fungal metabolites using stable isotope labeled precursors. In order to explain some of the reasoning behind the applied methods, a short introduction to the biosynthesis of fungal metabolites is given below.

1.3.1 Polyketides

PKs represent a very diverse class of compounds that fulfill a multitude of roles for the producer organism. Although very diverse in structure, PKs are biosynthesized from the same precursors or starter units, such as acetyl coenzyme A (CoA) or malonyl-CoA (Simpson and Cox 2012).

PKs are biosynthesized by large enzyme complexes called polyketide synthases (PKS), for which several different types exist (Hertweck 2009). These are made up of several different types of catalytic domains comprising a minimum of three domains: the acyltransferase (AT), β-ketoacyl CoA synthase (KS), and acyl carriers protein (ACP) domains (Keller et al. 2005). In short, the AT domain is responsible for selecting and providing an extender unit (building block), and the KS domain is responsible for catalyzing the Claisen-like condensation reaction that joins the extender unit and the growing PK chain. Lastly, the ACP domain is responsible for covalent attachment of the PK chain, and maneuvering between catalytic domains, while building the PK chain.

In fungi, the PKSs are usually of a configuration called type I iterative PKSs. The term iterative refers to the way the biosynthesis is carried out: repeating cycles of extension re-using the same catalytic domains, while type I refers to a linear arrangement of catalytic domains unlike having domains present in a complex of discrete enzymes (type II). Because of the iterative nature of the PKS, it is difficult to predict the product of a such, as the number of reduction reactions, the identity of the extender unit, the methylation pattern, and possible cyclization can result in very different products (Walsh and Fischbach 2010).

Further modification of the PK products often takes place in many different post-PKS synthesis steps. Products can undergo cyclizations, carbon bond cleavages, and rearrangement reactions resulting in the formation of carba- and heterocycles. Tailoring reactions such as glycosylation, alkylations, acyl transfers, and hydroxylations can also take place, providing an immense diversity of products (Hertweck 2009). In my studies I have worked extensively with the PKs for investigation of biosyntheses. This includes investigations into the biosynthesis of yanuthone D from a 6-methyl salicylic acid (6-MSA) precursor (**Paper 3**) described in in section 2.2.2, as well as investigation of the PK YWA1 and the biosynthesis of compounds derived thereof (**Paper 4**), as described in chapter 2.2.3.

1.3.2 Nonribosomal peptides

Another large group of compounds found in microorganisms are the NRPs. These are biosynthesized from amino acids (AAs) by multidomain, multimodular enzymes called nonribosomal peptide synthases (NRPSs). Unlike the fungal PKSs, the NRPS are not iterative i.e. the catalytic domains of the NRPS are not re-used. Instead, the NRPS contains several so-called modules, and each of the modules in the NRPS contains all the domains that allow for recognition, activation, and binding of a specific AA. The AA is then covalently bound to the NRPS as a thioester, after which peptide bonds are formed between the selected AAs. Other catalytic functions may be present in the NRPS, including epimerases, that catalyze conversion from L-to D-forms of AAs (Finking and Marahiel 2004; Keller et al. 2005).

Advances in bioinformatics have made it possible to predict the products encoded by NRPSs in microorganisms, however, these prediction tools are not yet perfect and can at best be used as guidelines for a specific trend: in fungi, they may be used to suggest possible AAs present in the final product, however, the predictions are tentative and can in some cases only be used to predict that one AA should be aromatic etc. (Challis et al. 2000). The biosynthesis of the NPRs nidulanin A and the related fungisporin were investigated (**Paper 6)** (section 2.2.4) and illustrate that we are not yet able to predict all products of NRPSs.

1.3.3 Hybrid metabolites

Hybrid metabolites are metabolites of mixed biosynthetic origin. Examples of hybrid metabolites are the meroterpenoids - hybrid metabolites comprising terpenoid part as well as a non-terpenoid part (Geris and Simpson 2009). In this study the meroterpenoids are exemplified by the yanuthones (**Paper 3**) and asperrubrol (**Paper 4**). Another example of a hybrid metabolite is nidulanin A (**Paper 5**), which is a cyclic tetrapeptide of NRP origin, as well as a prenyl-group biosynthesized as part of the terpenoid pathway.

1.4 Chemical analysis

The work described in this thesis has been conducted using ultra high pressure liquid chromatography diode array detection quadrupole time-of-flight (UHPLC-DAD-QTOF) hyphenated instruments. These are very versatile instruments allowing for a wide range of different experiments. More importantly TOF-type instruments allow for full-scan acquisition of data. This means that instrument is able to record all ions in a wide mass-range in a single analytical run. Data recorded using an LC-MS system is therefore two-dimensional, as seen in Figure 3.

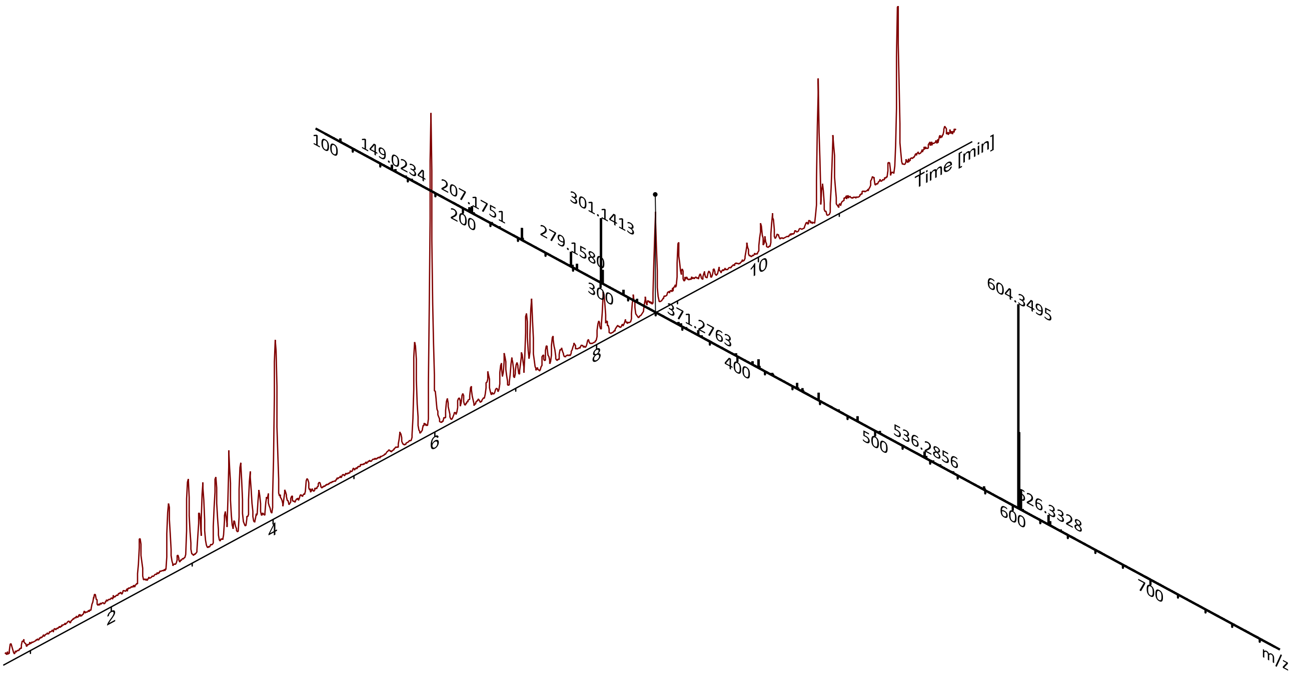


Figure 3 - LC-MS data obtained from a full-scan acquisition using a TOF-based instrument. Many different ions are detected at the same RT as the instrument scans in the 100-1,000 *m/z* range.

By using a hyphenated technique like LC-MS, it is possible to analyze complex samples, as compounds can be separated based on their chemical properties in the LC system before entering the MS system. Because of this, hyphenation with LC not only leads to simplified mass spectra, by reducing or eliminating co-eluting compounds, it also provides information on the chemical properties of the compound. Based on the stationary and mobile phases used in the LC, the RT a compound can be correlated to the logD providing additional information about the compound (K. F. Nielsen et al. 2011).

For the types of chemical analysis performed for this thesis, full-scan instruments are a requirement for effective analysis. Several types of instruments can be used to perform full-scan acquisition of data. Although quadrupole based MS systems such as triple quadrupoles MS are technically able to perform full-scan acquisition of data, the mass accuracy and isotopic pattern recorded is insufficient for use in dereplication. Another option is the Fourier transform ion cyclotron resonance (FT-ICR) systems that offer unprecedented mass accuracy and determination of isotopic pattern. It is possible to interface LC and FT-ICR, however the low scan speed of the FT-ICR makes it unsuitable for the narrow peaks obtained from UHPLC analysis. FT-ICRs are therefore often used for analysis of few very complex samples as opposed to larger screening regiments. Other disadvantages of the instrument are the very high price and the complexity of operation (Brown et al. 2005; J. Zhang et al. 2005).

The best suited instrument types for interfacing with LC for analysis of complex samples are thus the TOF (Mamyrin 2001) and orbitrap (Strife 2011; Zubarev and Makarov 2013) based MS-systems, and these are also the most widely used instruments fulfilling the mentioned criteria. I will not give a detailed description of the different instrument types in the present thesis, however, one of the key differences between the instruments is the ability of some orbitrap instruments, those fitted with ion-traps, to be used for tandem MS^n (MS to the power of *n*), where the TOF based instruments can only do MS/MS (MS^2). A comparison of some of the key specifications of the two instrument types is given in Table 1.

Table 1 - Comparison of TOF and orbitrap mass analyzers. Typical values for an *m/z* range of 300-400 are given. Data from (Andrews et al. 2011; Honoré et al. 2013; Jones et al. 2013; Junot et al. 2014; Krauss et al. 2010)**.**

Mass spectrometer	Resolving power (FWHM)	Mass accuracy (ppm)	Linear dynamic range	Scan speed	Mass range	Isotope ratio accuracy (A+1)	Sensitivity (absolute mass)
QTOF	60,000	1-2	10^4-10^5	Up to 100 Hz	Up to 40,000 *m/z*	<2 %	Picogram (full scan)
Orbitrap	240,000	2	10^3-10^4	Up to 12 Hz	Up to 4,000 *m/z*	3-10 %	Femto- to picogram (full scan)

The two instrument types can be used for many types of analysis. One type is targeted analysis, which can refer to several different analytical techniques. In this thesis the term is used to describe a method where a specific compound is being analyzed. However, the analysis is performed using a standardized method, and not methods optimized for the specific compounds. Here, the term thus refers to the retrospective analysis of recorded data to determine if a specific compound is present. In my studies, targeted analysis has been used for investigation of compounds from a specific biosynthetic pathway, by making it very easy to investigate any changes in intensities or isotopic patterns.

Several studies comparing the performance TOF- and orbitrap based instruments have been published. For metabolomics, the performance of the two instrument types was found to be comparable, and both instrument types were found to be well suited for use in metabolomics (Glauser et al. 2012). Many new types of hybrid TOF-based instruments have been developed in the last decade, enabling new forms of analysis. One of these is the TripleTOF, consisting of a hybrid quadrupole TOF platform working at a very high signal acquisition rate, with the speed and sensitivity of a TOF and quantification capabilities of a QqQ-based system (Andrews et al. 2011; Jones et al. 2013). Another hybrid instrument is the ion mobility TOF system, where ions are separated based on their flight time through a gas chamber, thereby separating ions based on their cross-section in addition to their accurate mass, allowing for separation in an orthogonal dimension (Kanu et al. 2008; Sysoev et al. 2013; Wolfender et al. 2014).

1.4.1 Tandem mass spectrometry

All experiments performed in my studies were performed using QTOF instruments. Because of the addition of the quadrupole, QTOF instruments can be used to perform several different MS/MS or tandem MS experiments.

Traditionally, MS/MS was performed by making a method, for which a specific ion was selected for study. This is referred to as targeted MS/MS, and is illustrated in Figure 4. In this experiment the Q is used to select ions with a specific *m/z*-ratio. These ions are then transferred to the collision cell, where the ions are fragmented, followed by detection. The result of this is a list of fragment ions formed by the targeted ion, as well as their abundances. Using targeted MS/MS, rather than single MS, a better selectivity can be achieved, and by matching the formed fragments against a database, the identity of compounds can be determined with higher certainty (**Paper 2**) (de Hoffmann and Stroobant 2007; Ding et al. 2013; Vaclavik et al. 2014).

This type of analysis is typically performed to quantify compounds, and is routinely employed using QqQ-instruments for screening of drugs, food, and feed for toxins, pesticides etc., as QqQ instruments have the highest selectivity (Kaufmann 2011). Advances in electronics and software had also made it possible to analyze samples using so-called data-dependent acquisition. In this mode MS/MS spectra of compounds are recorded at different fragmentation energies, based on the compound's *m/z*-ratio. In theory this makes it possible to record MS/MS spectra of all compounds in a sample, if they are chromatographically resolved to a degree that allows scanning of all concurrently eluting compounds, without making specific methods for each compound to be analyzed. This can be performed using both QTOF, orbitrap instruments and Q-Exactive instruments (Konishi et al. 2007; Lehner et al. 2011).

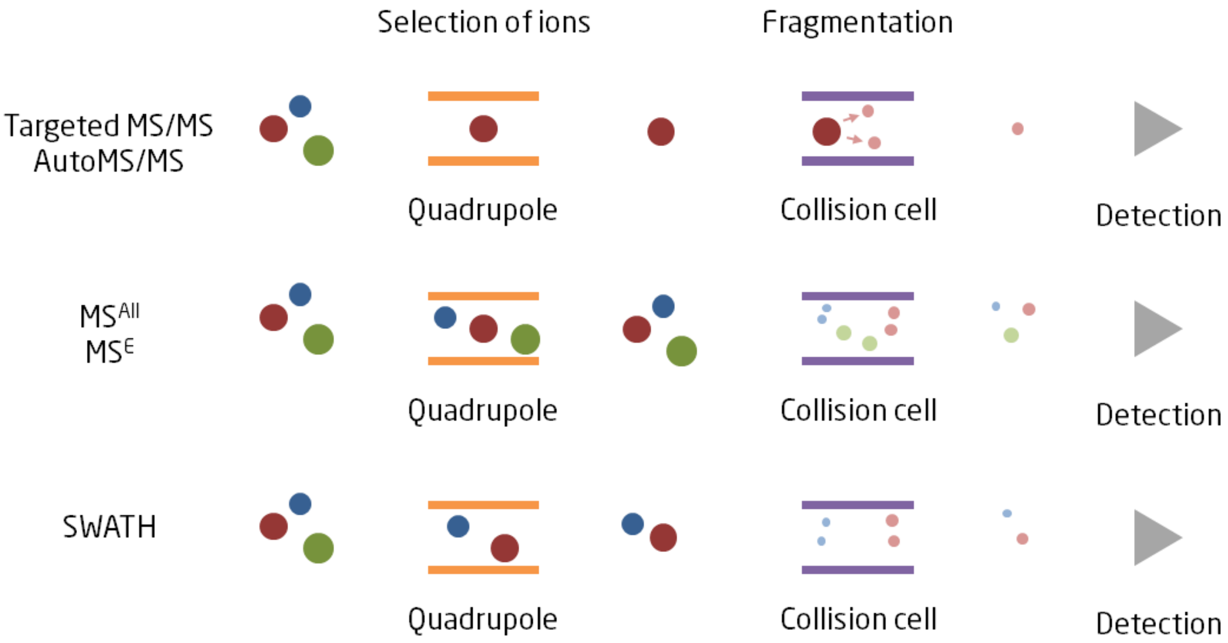


Figure 4 - Overview of the different MS/MS techniques.

Other ways of recording MS/MS data include the MS^{All} or MS^{E} methods where all ions entering the mass spectrometer are fragmented, and the resulting fragments are then detected. This can be used to reveal structural information about both known and unknown compounds (Figure 4) (Bijlsma et al. 2011). By building libraries of known fragments, these can be used to predict the structures of unknown compounds by matching known losses against the libraries, aiding identification of compounds (Hufsky et al. 2014a; Wolf et al. 2010). Several different methods relying on different informatics procedures have been developed for this prediction of MS/MS spectra and chemical structures (Hufsky et al. 2014b).

A relatively new method for acquisition of MS/MS data is sequential windowed acquisition of all theoretical mass spectra (SWATH), which can be performed using TripleTOF instruments (Collins et al. 2013; Röst et al. 2014; X. Zhu et al. 2014). This technique is a compromise between the targeted MS/MS and MS^{All}, where a narrower window of ions is passed into the collision cell compared to MS^{All} (Figure 4). This allows for recording of more specific mass spectra while still allowing for recording data for all compounds.

During my studies, I have used MS/MS data acquisition for several of the studies I was involved in. Firstly, a method for dereplication of metabolites based on MS/MS data was developed (chapter 2.1) (**Paper 2**). Other examples include the yanuthone D study (chapter 2.2.2) (**Paper 3**) where MS/MS spectra were recorded of the different yanuthones for aiding in linking them to a biosynthetic pathway. In the study of nidulanin A (chapter 2.2.4) (**Papers 5 and 6**) it was used to link biosynthetic analogs but also to elucidate the structure of the compounds. Finally, it was used for dereplication of compounds from marine bacteria (chapter 2.3) (**Paper 7**).

1.4.2 Methods for untargeted analysis

The term untargeted analysis refers to studies where there is no explicit target. Although in chemistry the term is often equated to metabolomics, in principle it may refer to any analysis form that is not based on measurement of a specific target. Several methods for untargeted analysis of samples can be used depending on the object of the analysis. For the work performed in this thesis, the object of analyses has most often been to find new compounds, or to find compounds that were only present in a subset of samples (America and Cordewener 2008). Traditionally, samples have been investigated using comparative analysis, where the BPCs of two samples have been compared against each other to identify any differences, as seen in Figure 5.

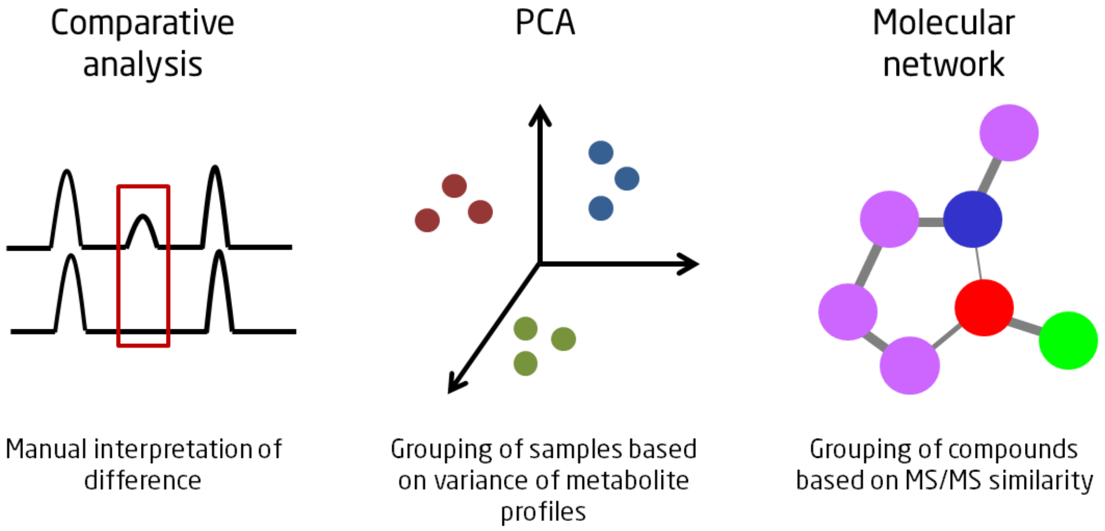

Comparative analysis
PCA
Molecular network
Manual interpretation of difference
Grouping of samples based on variance of metabolite profiles
Grouping of compounds based on MS/MS similarity

Figure 5 - Different types of untargeted data analysis.

As this method requires manual investigation of data files it is extremely labor-intensive and unfeasible to use for analysis of large datasets.

Principal component analysis (PCA) is traditionally the method of choice to group microorganisms on the basis of their production of small molecules as it provides a nice visual representation of the variance between LC-MS profiles (Figure 5) (Forner et al. 2013; Hou et al. 2012). While PCA can be good for a first exploratory step in the data analysis, it can become problematic with data of high dimensionality like metabolomics data as the use of noisy variables may disturb separation between samples (Boccard et al. 2010).

A relatively new method for data analysis is mass spectral molecular networking developed by Dorrestein and coworkers (Watrous et al. 2012). It builds on an algorithm (Liu et al. 2009; Ng et al. 2009) capable of comparing characteristic fragmentation patterns and thus highlighting molecular families with the same structural features and thus potentially same biosynthetic origin. This enables the study and comparison of a high number of samples, at the same time aiding in dereplication and tentative structural identification (J. Y. Yang et al. 2013). Mass spectral networking was used for two of the projects I worked on as part of this thesis. In one project, it was combined with isotopic labeling in a novel procedure for detection of biosynthetic analogs and subsequent identification of these, as described in section 2.2.4 (**Paper 6**). In another study it was used for detection of biosynthetic analogs of marine bacterial metabolites, as described further in section 2.3 (**Paper 7**).

1.5 Metabolomics

It is no easy task to define the term metabolomics. Jeremy Nicholson, Chair in Biological Chemistry at Imperial College, London, UK has said that: “Metabolomics has about 20 published definitions, conflicting but all analytical, all about measuring some stuff in some other stuff” (Hunter 2009). The term is mostly used to refer to the experimental designs based on the detection and quantification of global metabolite levels without prior identification of the metabolites. As such, metabolomics is focused on the study of the metabolism of both endogenous and exogenous metabolites in biological systems (Dunn 2008). Metabolites also serve as direct signatures or markers of biochemical activity. Genes and proteins can on the other hand be subject to epigenetic regulation and post-translation modifications, respectively. Metabolites are therefore easier to correlate with phenotypes (Patti et al. 2012). Metabolomics therefore allow for study of organisms for a wide variety of experiments, such as finding new compounds and optimizing industrial biotechnology process, helping to further our understanding of biology (Hendriks et al. 2011).

1.5.1 Feature extraction and alignment

One of the main challenges in metabolomics is the complexity of the samples being analyzed. As the samples contain many different compounds, with different physical-chemical properties, we need a very versatile method for extraction and analysis. One such method is LC-MS. Again, the TOF-based instruments are well suited because of their high dynamic range, allowing for analysis of extracts containing compounds in very different concentrations, or for analysis of compounds with very different ionization efficiencies. The workflow used in metabolomics is often divided into several stages, including filtering, feature detection, alignment, and normalization (Hendriks et al. 2011; Katajamaa and Oresic 2007). I will only describe the feature extraction and alignment in detail, since these are the areas I focused on in my studies.

For metabolomics analysis, all compounds present in a sample first need to be extracted from the data file. Each compound is referred to as a chemical feature. To be able to compare chemical features extracted from different samples, all chemical features need to be matched across all samples so that the same compound, found in two different samples, is recognized as the same chemical feature. This can be done in different ways depending on the algorithm used, but a simplified view is that extracted ion chromatograms (EICs) are extracted at a fixed interval across the analyzed mass range. Many feature extraction algorithms now allow for concatenation of ions into a single chemical feature. In this way pseudo molecular ions corresponding to the same compounds are combined into one chemical feature, which is a great advantage, as it reduces the complexity of the data without any loss of information, as seen in Figure 6.

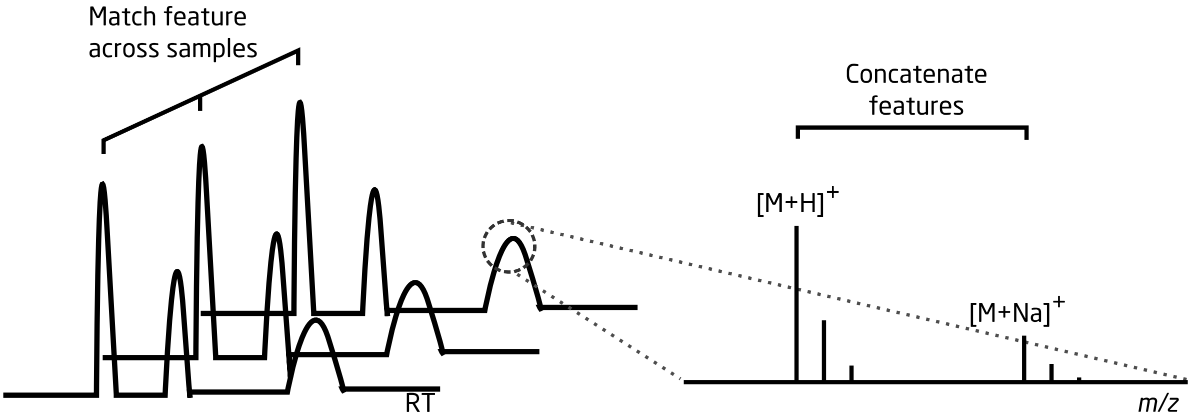


Figure 6 - Feature detection and matching across samples.

Each extracted chemical feature will therefore be a unique combination of ion *m/*z value and RT, as illustrated in Figure 3. In practice, many more factors are used for determination of a chemical feature. By taking into account the isotopic pattern, it can be assessed whether an ion corresponds to a compound or if it is merely noise. The chromatographic behavior of a compound can also be taking into consideration by examining if the intensity of the EIC displays a clear maximum and peak shape like a true compound would. Further complication can be caused by concentration dependent adduct-formation, as described in **Paper 1** (Figure S3). Analysis showed that different concentrations of the metabolite roridin A, lead to very different adduct patterns, and compounds exhibiting this behavior might cause problems when extracted as a chemical feature.

Extracted chemical features then have to be matched across all analyzed samples. Although this sounds very simple, in practice this can be very difficult, as long sample sequences can lead to changes in the LC-MS system through the sequence e.g. build-up of impurities in the column leading to degraded chromatographic separation, or deposition of impurities in the LC-MS interface leading to lower ionization efficiency and thus lower detected intensities. The complex nature of samples often means that the compounds present in the samples are impacted differently by this, leading to non-linear shifts in RT and intensity. In general LC-MS exhibits poorer reproducibility of retention time (RT) and mass spectra compared to gas chromatography (GC)-MS (Lee et al. 2013). Because of this many different algorithms for alignment of data exist for GC-MS data. However, because LC-MS can be used to analyze such a wide variety of analytes, a lot of research has been performed to develop methods for feature extraction, alignment etc. allowing for LC-MS based metabolomics to become a very widely used analysis method (Moco et al. 2006).

As mentioned above, the shifts in RT and loss in intensity throughout an analysis sequence can lead to various undesirable situations, as illustrated in Figure 7. The schematics illustrate situations requiring alignment and use of quality control samples for the analysis for untargeted analysis. A) Reference sample. B) In this situation the RT for all compounds has been shifted up. This can be alleviated by a linear alignment of RT across samples. C) In this situation the RT has been shifted only for certain compounds. The data can be treated with a non-linear warping function to align compounds across samples. D) The detected intensity one compound is lower than expected. This can be corrected by using quality control samples with known concentrations of compounds. E) The detected intensity of compounds in the sample is lower than expected. To correct for this, the quality control sample must contain compounds exhibiting the same behavior as the compound in question, for instance a sample containing a mix of fractions from several different samples (Hendriks et al. 2011). F) Finally, peak broadening leading to overlapping peaks. This is one of the most difficult situations to correct for. The problem can be alleviated by using a detector that can be used to deconvolute signals, that is extract spectra for a specific compound from a spectrum of a mixture of several compounds signals, or by using a mass analyzer such as a TOF (Katajamaa and Oresic 2007; Patti 2011; W. Zhang et al. 2014).

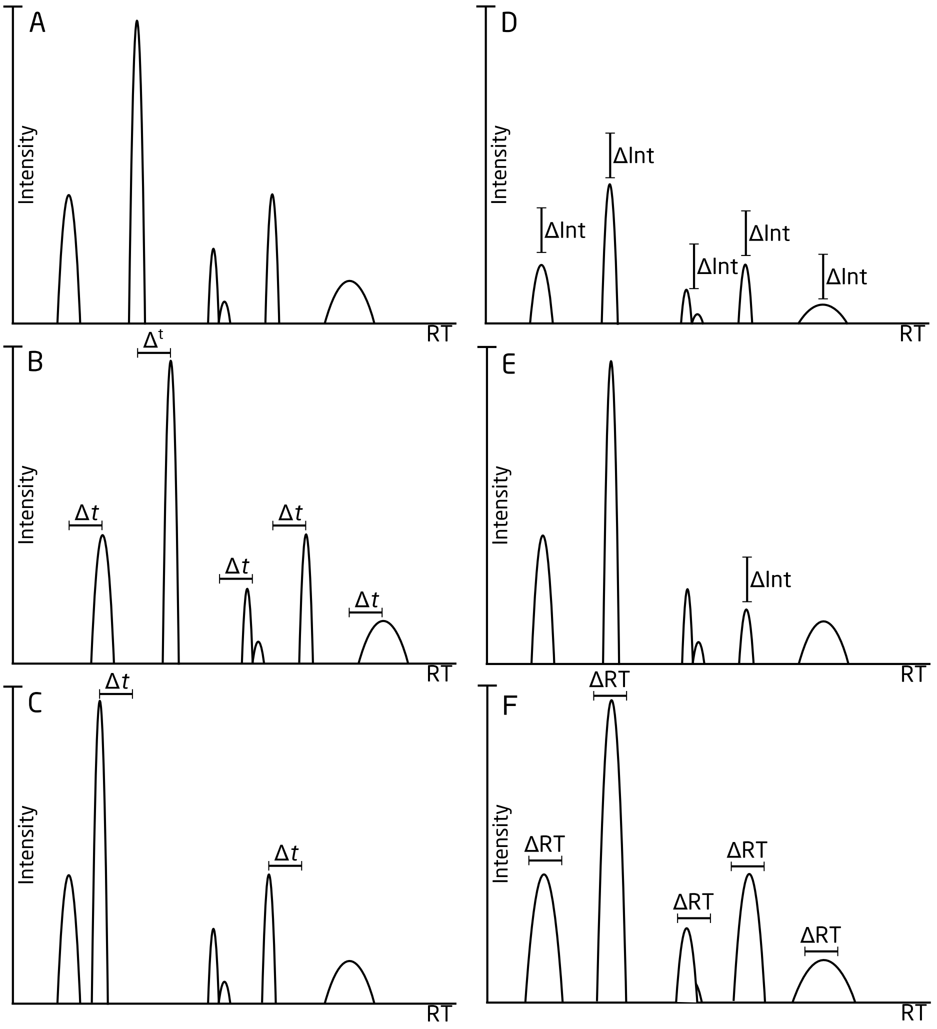


Figure 7 - Schematic illustration of a chromatogram obtained from analysis of a sample using an LC-MS system. Retention time (RT), intensity (Int). The schematics illustrate situations requiring alignment and use of quality control samples for the analysis for untargeted analysis. A) Reference sample. B) In this situation the RT for all compounds has been shifted. C) Shift in RT for some compounds. D) Detected intensity is lower in the sample. E) The detected intensity of some compounds is lower than expected. F) Peak broadening leading to overlapping peaks.

One way to reduce the problem of data alignment is to use binning: by summing *m/z* data across preset time windows, the alignment error will be confined to the edges of the bins. Subsequent analysis can then reveal the data points responsible for deviation in the alignment (Nordström et al. 2006).

Many different software packages have been developed for feature extraction and subsequent feature alignment (Sugimoto et al. 2012). Some of the most well-known are: Metalign (Lommen 2009), MZmine (Katajamaa et al. 2006; Pluskal et al. 2010), and XCMS (Gowda et al. 2014; Huang et al. 2014; Tautenhahn, Patti, et al. 2012). Most instrument vendors have developed their own proprietary analysis software that utilize their own feature extraction algorithms, such as Agilent Technologies' Molecular feature extractor, and Bruker Daltonics' Find molecular feature algorithms.

Because of the complexity of the task of extracting chemical features and then aligning them, several methods and protocols for optimization of the data processing step in LC-MS based metabolomics have been published (Eliasson et al. 2012; Zheng et al. 2013). In spite of this, some prior knowledge about the dataset and the compounds present in the samples can be almost mandatory for successful design of metabolomics experiments. This is in spite of the fact that metabolomics is often referred to as an "unbiased" method of analysis, while in reality one could argue that even the choice of a specific feature extraction algorithm imposes a bias on the analysis (Fiehn 2002; Kluger et al. 2014). A study by Lange *et al.* comparing the most widely used feature extraction algorithms, showed that significantly different results were obtained from analysis of the same dataset when using different feature extraction algorithms (Lange et al. 2008). This demonstrates the complexity of the feature extraction step and highlights the need for more standardized operations and benchmarks for evaluation of metabolomics data analysis.

The type of metabolomics workflow described here was used for the study of metabolites from marine bacteria as described in chapter 2.3 (**Paper 7**). In this study, many of the subjects discussed here, such as feature extraction, alignment and data analysis are discussed from a practical point of view.

1.6 Targeted analysis contra metabolomics analysis

As outlined in the section 1.4 and 1.5, targeted and untargeted metabolomics analysis are distinctly different methods of analysis. The methods require different experimental setups, different methods of data analysis, and are often used in the examination of very different hypothesizes.

One of the main advantages of a targeted analysis is the possibility of using samples acquired at different time points. As described in section 1.5, proper metabolomics analysis requires the alignment of chemical features for successful analysis. By combining samples analyzed in different sample batches, alignment becomes almost impossible, even with the use of high quality control samples. The type of targeted analysis methods described in this thesis allows for comparison of data obtained from different analytical runs, allowing one to compare samples that have been run months apart. This makes the method very well suited for biosynthesis studies, where sample can be retroactively screened for a compound of interest. Because of this, the two methods are complementary and can be used for finding answers to different hypotheses. A comparison of targeted and untargeted analysis methodologies is given in Table 2.

Table 2 - Comparison of targeted and untargeted analysis methodologies.

Targeted analysis	Untargeted metabolomics analysis
Analysis of specific compounds	Analysis of "everything" – broad range of compounds analyzed in each sample
Samples used for analysis can be from different analytical runs	Samples used for analysis must be from the same analytical batch for proper alignment
	Requires significant experimental planning and quality control
LC-MS method can be optimized for the target compound	LC-MS method will be a compromise to enable analysis of a broad range of metabolites
Limited information about quantities if no standards are used Better quantitation through use of internal standards	Offers comparison of relative abundances across of compounds in the samples
Reductive analysis	Exploratory analysis

1.7 Dereplication

In natural product chemistry, the main focus is on discovery and identification of new compounds. Samples extracted from microorganisms contain a wealth of compounds, but some of these compounds could have been identified previously. Because of this, one of the most important steps in the analysis of samples from natural extracts is “dereplication”, or tentative identification of compounds in the samples. The term dereplication was first used in the CRC Handbook of antibiotic Compounds that was published in 1980, and was used to describe the process of recognizing and eliminating the active substances already studied in the early stage of the screening process (Ito and Masubuchi 2014). By determining which compounds that are potentially novel as quick and as early as possible, resources can be focused on identification and profiling of the possible new compounds rather than squandering resources on already known compounds.

Several methods and protocols for dereplication have been developed throughout the years utilizing different types of instruments and detectors. Several reviews on the topic of dereplication of microbial compounds have been published, thoroughly describing commonly used protocols and instrumental setups (Callahan and Elliott 2013; Eugster et al. 2011; Ito and Masubuchi 2014; Wolfender et al. 2003, 2010, 2014). I have therefore chosen only to briefly introduce the most common methods, and to present some of the most recently developed methods for dereplication, focusing on automated methods.

One of the most commonly employed methods of dereplication is by analysis using liquid chromatography – diode array detector – mass spectrometry (LC-DAD-MS) systems. Using this hyphenated analysis method, analytes can be evaluated on several different parameters: the RT, the nature of UV/Vis absorption, and the mass spectrum.

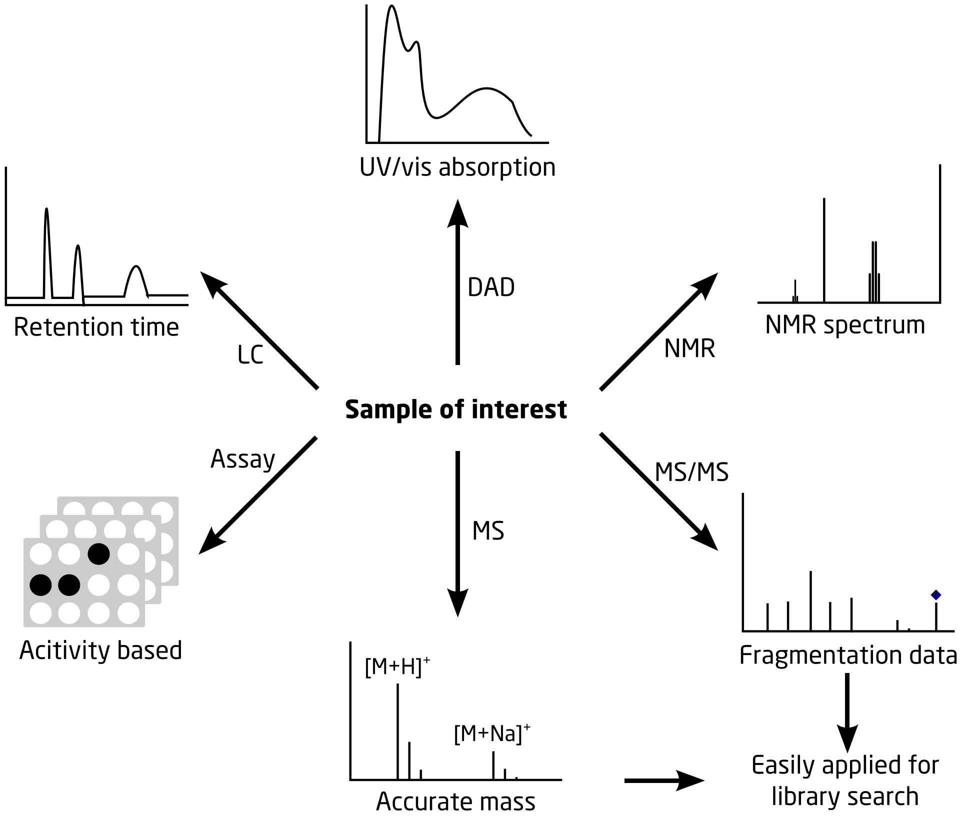


Figure 8 - Commonly used dereplication techniques. Dereplication using DAD, MS/MS, and NMR can be used to directly determine structural characteristics of the compounds being dereplicated, whist MS can provide the elemental composition of a compound. MS and MS/MS data can readily be used for library searches, and are often used in conjunction with RT data. LC can be used to infer structural characteristics by using the RT to estimate the logD of a compound, but is otherwise used as part of a hyphenated system to separate compounds in a sample. The activity based screening offers no direct information about the investigated compounds, but is used to prioritize samples or fractions for further analysis.

LC-DAD based dereplication using UV-VIS, is very powerful for identification of compounds with distinct chromophores, but can only be used to deconvolute spectra if compounds are chromatographically resolved, and can of course only be used for analysis of compounds containing chromophores. Currently, UV-Vis data is used for dereplication by manual extracting the absorption spectrum for a compound of interest and then comparing the spectrum to a reference. Several methods for automation of this workflow have been suggested by development of algorithms that allow for automatic comparison of spectra to databases(Larsen and Hansen 2007), but currently LC-DAD is mostly applied in a hyphenated manner along with MS.

Recently, a new data analysis package has been developed for the open-source statistical computational environment R (R Core Team 2014) for analysis of LC-DAD data, called Alsace (Wehrens et al. 2014). The software allows for automated extraction and analysis of LC-DAD allowing for faster analysis of data. Data obtained from the LC-DAD analysis may also be combined with LC-MS data, and could be used to more easily combine data from the two detector types, and for alignment of data, which was discussed in section 1.5.1.

LC-MS based dereplication relies on ionization of the compounds of interest followed by measurement of the accurate mass and isotopic pattern of the formed ions (Forner et al. 2013; K. F. Nielsen and Smedsgaard 2003; K. F. Nielsen et al. 2011; Z.-J. Zhu et al. 2013) (**Paper 1 and 2**). The accurate mass of these ions can be used to determine the elemental composition of the compounds, but can result in ambiguous determination even at less than 1 ppm error (Kind and Fiehn 2006). To achieve unambiguous determination of the elemental composition, accurate detection of the isotope pattern of the compounds is required as well (Kind and Fiehn 2007; K. F. Nielsen et al. 2011). Using MS, detected signals can also be deconvoluted, making the method very well suited for extracts that contain many different compounds. This method is well suited for use in database searches, because the accurate mass or calculated elemental composition is easy to use a search queries. This is vital, as analysis using screening libraries allow for much faster analysis of data. The application of LC-MS dereplication and the use compound libraries is described in further detail in **Papers 1 and 2**.

Recently, several methods for MS/MS based dereplication have been developed. The aim of these methods has been to offer increased confidence in matches against libraries, as well as to allow for different methods of data analysis. One way of utilizing the MS/MS data is to match the acquired data against a database containing recorded spectra (El-Elimat et al. 2013; Horai et al. 2010; Smith et al. 2005). The application of MS/MS based dereplication using compound libraries is described in further detail in **Paper 2**.

During my studies I have worked on the development of two methods for dereplication of extracts from microbial samples, described in **Papers 1 and 2**. The two methods are both based on matching libraries of known compounds against those detected in samples utilizing MS and MS/MS data, respectively, and are further discussed and compared in chapter 2.1. Both of the developed methods rely on libraries for searching of spectral data, and as such, the libraries are essential for the success of dereplication, as further explored in section 1.7.1.

As mentioned in section 1.4.2, molecular networking using MS/MS data can also be used for dereplication by grouping compounds that exhibit similar fragmentation spectra. In this way compounds that share structural similarities may be grouped together with analogs with e.g. different substitution patterns. By using spectra obtained from standards or other already identified compounds, analogs of these can thus be detected (J. Y. Yang et al. 2013).

As neither NMR based or activity based dereplication were used in my studies, the reader is encouraged to consult either the before mentioned reviews or Halabalaki (Halabalaki et al. 2014) for more information on NMR consult Lang and coworkers (Lang et al. 2006), and López-Pérez (López-Pérez et al. 2007) for information of activity based dereplication.

1.7.1 Importance of databases

Databases are essential in biological sciences, as they allow for collection of information and knowledge that can then be leveraged for different types of analyses. In fact comprehensive databases are essential for successful dereplication, as described in chapter 1.7.

In my studies, I have mostly worked with compound databases, which contain information such as name, structure, elemental composition and MS^n data. The databases that I have primarily worked with are listed in Table 3.

Table 3 - List of databases used in my study.

Database	Area of focus	Contents	Availability	Open for contributions
Antibase (Laatsch 2012)	Metabolites from fungi and bacteria	Contains the name, structure, chemical formula as well as other physical properties for >40,000 microbial product entries Contains experimentally obtained and calculated NMR spectra	Commercially available	Compounds are added from literature
Marinlit ("MarinLit" n.d.)	Marine natural products	Contains the name, structure, chemical formula as well as other physical properties for >40,000 marine natural products Contains experimentally obtained and calculated NMR spectra	Commercially available	Compounds are added from literature
Dictionary of natural products (Press n.d.)	Natural products mainly from plants	Contains the name, structure, and chemical formula as well as other physical properties for >260,000 natural product entries	Free access to a subset of compounds. Full access as well as batch look-up requires paid access	Compounds are added from literature
METLIN (Smith et al. 2005)	Primary metabolites and peptides primarily from *Homo sapiens*	Accurate mass data for >240,000 metabolites >12,000 metabolites with HRMS/MS	Free access	No
LIPID MAPS (Sud et al. 2007)	Lipids	Contains the name, structure, and chemical formula, LIPID MAPS ID, category, main class, and subclass for >37,000 unique lipids	Free access	Compounds are added from literature and partners

Global natural products social molecular networking (GNPS) ("GnPS: Global Natural Products Social Molecular Networking" n.d.)	No particular focus	Contains the name, structure, chemical formula, and MS/MS spectra of>1,600 standards and identified compounds Contains >250,000 MS/MS spectra of unknown compounds	Free access	Yes
Massbank (Horai et al. 2010)	Small chemical compounds for life sciences including natural and synthetic compounds	Accurate mass data for >20,000 compounds >20,000 MS/MS spectra	Free access	Yes
DTU Mycotoxin-Fungal Secondary Metabolite MS/HRMS library (**Paper 2**)	Mycotoxins and fungal secondary metabolites	Contains the name, structure, chemical formula, and MS/MS spectra of 277 different compounds at different collision energies	Free access Requires Agilent PCDL library management software	No

The Global natural products social molecular networking (GnPS) database is special case, as it also acts as a data repository (Bouslimani et al. 2014). This means that it contains both spectra from known standards, as well as spectra from unknown compounds. Care must therefore be taken if the database is used for dereplication purposes.

As part of the development of the high-resolution MS/MS (HRMS/MS) library (**Paper 2**), a database containing MS/MS data for 277 mycotoxins and fungal SMs metabolites was made publically available.

Although I have mainly used Antibase for my studies, I frequently used other databases from Table 3 to investigate signals from unknown compounds. However, choosing the right database to search can be difficult. This is because the amount of data generated in biology is ever increasing, and with this increase in data, the number of databases containing information has also increased dramatically. In 2010 the number of database publications indexed in PubMed reached more than 1100, and it was estimated by Bolser *et al.* that this number might top 2000 publications in 2015 (Bolser et al. 2012). This number covers databases in the whole field of biology including databases containing genome data such as GenBank (Benson et al. 2011), metabolic pathways (Frolkis et al. 2010), and compounds (Laatsch 2012). Whilst it is an unmitigated success that so much information is being made available, the sheer number of segregated databases presents new challenges. With so many new databases being published, it is an daunting task to keep track of which databases are available and which areas of research they cover, and the segregated nature complicates the integration of available data (Searls 2005). Because so many different databases exist, it can be quite challenging to determine which ones are most relevant for a given project, and as well as to assess the quality of data in the database. To alleviate this, several meta-databases, or databases containing information about other databases, have been launched, including MetaBase (Bolser et al. 2012;

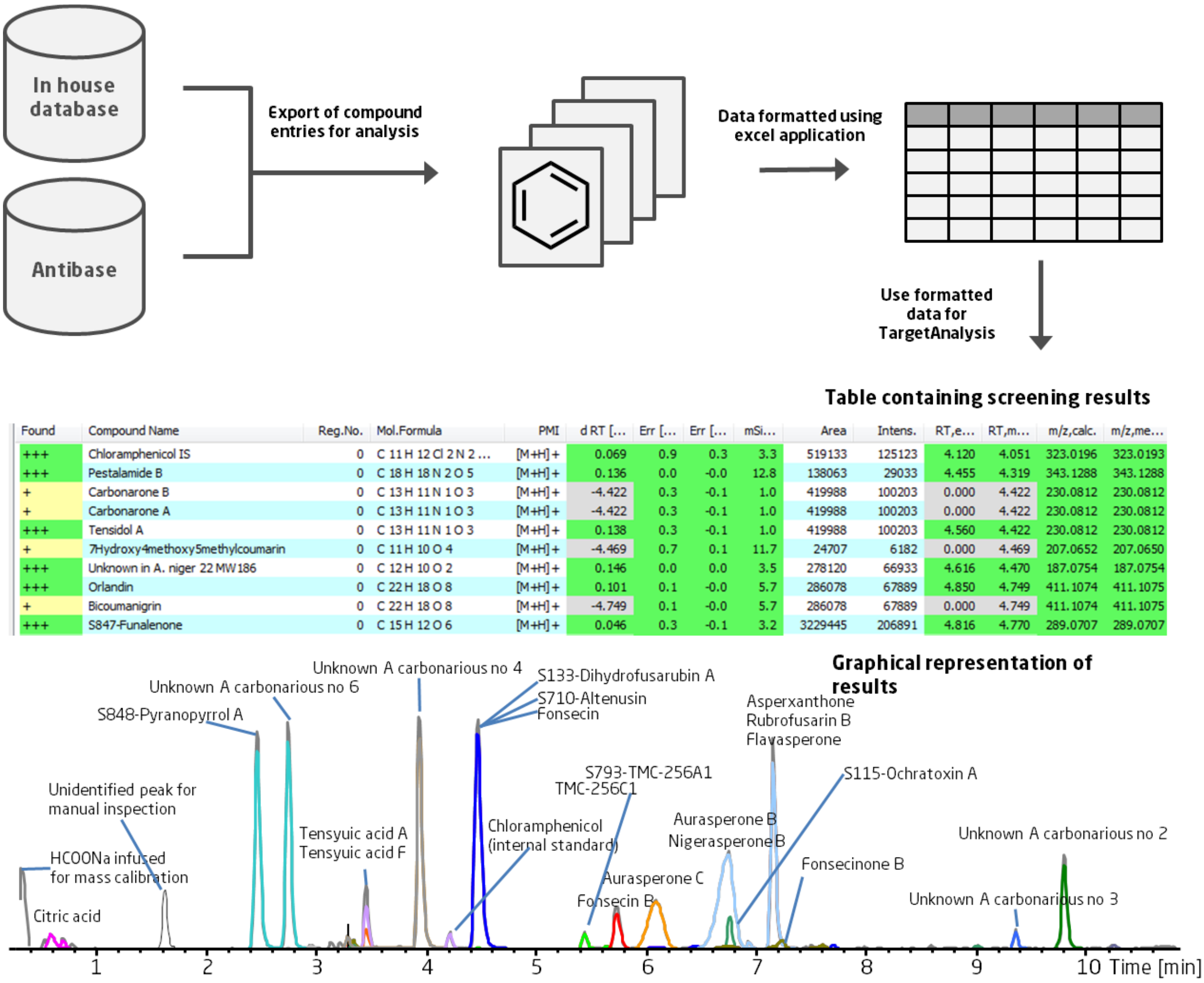
“MetaBase” 2014) and The Bioinformatics Links Directory (Chen et al. 2007). These meta-databases allow for the discovery of relevant databases for a given project.

Unfortunately, not many databases containing microbial products exist, and those that do exist contain no MS/MS data. A possible solution to this problem could be to encourage more sharing of data between research groups, and to agree on standards of reporting in the field. This will be further discussed in section 3.2.

2 Results and discussion

2.1 Targeted analysis for dereplication

During my thesis I have worked on the development of two different targeted screening methods: aggressive dereplication and HRMS/MS dereplication (**Papers 1 and 2**). Both methods were developed as means to speed up the traditional manually performed dereplication process, by quickly determining which of the detected compounds in a sample that were already known, and instead allowing researchers to focus their attention on the tentatively unknown compounds. The principle behind the two methods is the same: first, an extract of an organism of interest is analyzed using an LC-MS system. Compounds are then matched to entries in the library. If any compound from the library is detected in the sample, the peak in the chromatogram that corresponds to the compound is colored. By simply looking at the chromatogram, it is then possible to see which peaks correspond to known compounds, and which peaks might correspond to unknown compounds. The tentatively unknown compounds may then be further investigated manually or by other means.


In house database
Antibase
Export of compound entries for analysis
Data formatted using excel application
Use formatted data for TargetAnalysis
Table containing screening results
Graphical representation of results
S848-Pyranopyrrol A
Unknown A carbonarious no 6
Unknown A carbonarious no 4
S133-Dihydrofusarubin A
S710-Altenusin
Fonsecin
Asperxanthone
Rubrofusarin B
Flavasperone
S793-TMC-256A1
TMC-256C1
S115-Ochratoxin A
Unidentified peak for manual inspection
HCOONa infused for mass calibration
Citric acid
Tensyuic acid A
Tensyuic acid F
Chloramphenicol (internal standard)
Aurasperone B
Nigerasperone B
Aurasperone C
Fonsecin B
Fonsecinone B
Unknown A carbonarious no 2
Unknown A carbonarious no 3
1
2
3
4
5
6
7
8
9
10 Time [min]

Found	Compound Name	Reg.No.	Mol.Formula	PMI	d RT [...	Err [...	Err [...	mSi...	Area	Intens.	RT,e...	RT,m...	m/z,calc.	m/z,me...
+++	Chloramphenicol IS	0	C 11 H 12 Cl 2 N 2 ...	[M+H]+	0.069	0.9	0.3	3.3	519133	125123	4.120	4.051	323.0196	323.0193
+++	Pestalamide B	0	C 18 H 18 N 2 O 5	[M+H]+	0.136	0.0	-0.0	12.8	138063	29033	4.455	4.319	343.1288	343.1288
+	Carbonarone B	0	C 13 H 11 N 1 O 3	[M+H]+	-4.422	0.3	-0.1	1.0	419988	100203	0.000	4.422	230.0812	230.0812
+	Carbonarone A	0	C 13 H 11 N 1 O 3	[M+H]+	-4.422	0.3	-0.1	1.0	419988	100203	0.000	4.422	230.0812	230.0812
+++	Tensidol A	0	C 13 H 11 N 1 O 3	[M+H]+	0.138	0.3	-0.1	1.0	419988	100203	4.560	4.422	230.0812	230.0812
+	7Hydroxy4methoxy5methylcoumarin	0	C 11 H 10 O 4	[M+H]+	-4.469	0.7	0.1	11.7	24707	6182	0.000	4.469	207.0652	207.0650
+++	Unknown in A. niger 22 MW186	0	C 12 H 10 O 2	[M+H]+	0.146	0.0	0.0	3.5	278120	66933	4.616	4.470	187.0754	187.0754
+++	Orlandin	0	C 22 H 18 O 8	[M+H]+	0.101	0.1	-0.0	5.7	286078	67889	4.850	4.749	411.1074	411.1075
+	Bicoumanigrin	0	C 22 H 18 O 8	[M+H]+	-4.749	0.1	-0.0	5.7	286078	67889	0.000	4.749	411.1074	411.1075
+++	S847-Funalenone	0	C 15 H 12 O 6	[M+H]+	0.046	0.3	-0.1	3.2	3229445	206891	4.816	4.770	289.0707	289.0707

Figure 9 - Example of workflow for screening of fungal extracts using the aggressive dereplication method performed using the software TargetAnalysis (TA). In this example an extract from the fungus *Aspergillus niger* has been analyzed using a library containing compounds from an in-house database as well as compounds from Antibase. Signals detected in the sample are matched against the assembled libraries, and results are presented both in graphical and tabulated form. The graphics shows the BPC overlaid with colored peaks corresponding to tentatively known compounds. The peaks with no overlay, such as the one at 4.0 min, have not been tentatively identified, and could thus correspond to a novel compound (Paper 1).

The method entitled *aggressive derepliction* (**Paper 1**) was developed first and was based on the creation of a HRMS library for screening samples based on UHPLC-DAD-QTOF data acquisition. The library used for the screening could be created using different sources, dependent on the organism that was to be analyzed. In the case of an extract from *Aspergillus nidulans*, a library consisting of all known metabolites from that fungus could thus be used. This library could be compiled using commercially available databases such as Antibase (Laatsch 2012), and could be supplemented by including other compounds of interest such as tentatively identified compounds, and even known impurities such as plasticizers. One of the advantages of the method was that it was very effective for quickly determining how many metabolites were known for given organism. Some of the well investigated species, such as *A. niger*, exhibited very few unidentified peaks, while an extract of *Penicillium melanoconidium* showed almost no identified peaks, thus allowing one to focus the dereplication efforts on the extract from *Penicillium.* A disadvantage of the method was that, unless the RT of a compound was known, it was not possible to distinguish between structural isomers with the same elemental composition. Because of this a more specific targeted analysis method was needed.

To address the need for specificity a new automated dereplication procedure was developed. The method entitled *HRMS/MS dereplication* (**Paper 2**) was based on creation of a HRMS/MS library for screening samples by UHPLC-DAD-QTOF based data analysis, but this time requiring data acquired in AutoMS/MS mode. The spectral library was prepared by analyzing compound standards at three different collision energies (10, 20, and 40 eV). By using different fragmentation energies, the chance of acquiring an MS/MS spectrum of sufficient quality for spectral matching increased. The confidence of a hit i.e. identification of an unknown compound, using this method, was much improved over the *aggressive dereplication* method, and the method could even distinguish between some structural isomers. However, as each standard must be analyzed using the LC-MS system, creation of the library itself was initially very labor intensive, while subsequent use of the method required no extra work.

The methods were compared (**Paper 2**) by applying both methods to data files obtained from analysis of a range of different marine fungi, and the advantages and disadvantages of the two methods have been summarized in in Table 4.

Table 4 - Comparison of the advantages and disadvantages of the two targeted analysis methods *aggressive dereplication* and *HRMS/MS dereplication*

Method	Advantages	Disadvantages
Aggressive dereplication (HRMS library)	High confidence of hit (if combined with RT) Can be used to quickly evaluate if a sample is well described Reference standards are not required	Requires a library of a certain size for effective analysis Low confidence if RT is unknown. RT is defendant on the LC-method used Cannot distinguish structural isomers (without RT) Requires curated libraries to reduce false positives
HRMS/MS dereplication (HRMS/MS library)	Very high confidence of hit Can distinguish between some structural isomers Does not require RT - can thus be used with different LC-methods Does not require a library curated for the specific sample	Requires library spectrum of compounds to be analyzed for No commercially HRMS/MS libraries for fungi available Lower sensitivity Analysis of standards for library creation is very work intensive

The two methods are currently used in a complementary manner. The aggressive dereplication method will be superior for well described organisms, where appropriate libraries can easily be assembled. This means that the method is most effective if some information about the sample or organism being analyzed is already known. For instance, if an extract of *A. niger* is to be analyzed, a library containing compounds previously detected form *A. niger* will be ideal. A library containing all compounds isolated from the *Aspergillus* genera could also be used. However, because of the inability to distinguish between isomers without RT, the libraries can reach a size where the number of false positives makes the method less effective.

The limiting factor of the *HRMS/MS dereplication* method is the small size of the library. As the size of the library increases by addition of new compound data, the effectiveness of the method will increase as well. Because of the increased confidence of hits over the *aggressive dereplication* method, the whole library could potentially be leveraged for every search, instead of having to use a curated library to reduce the number of false positives. Because of this, the method can be used with good effect when screening extracts from organisms of unknown taxonomy.

Both of the described methods have the potential of becoming more useful in the future. The development of more advanced instrumentation, better predictions models for compound RTs in LC, and better prediction of MS/MS spectra will allow for a higher degree of confidence in tentative identification of the dereplicated compounds. This will be further explored in section 3.1.

2.2 Investigation of biosynthesis

One main goal of this study was to link fungal SMs to genes, however, it can be hard to determine which genes are involved in the biosynthesis of fungal metabolites. As described in section 1.3, this is because it is still not possible to computationally predict the end products from iterative PK synthases, and thus easily link genes to the corresponding metabolite(s) (Hertweck 2009; Walsh and Fischbach 2010). For NRPs the situation is simpler as prediction tools can in some, but not all, cases be used to predict the product (Challis et al. 2000). In the case of nidulanin A, described in section 2.2.4 (**Papers 5 and 6**), it was not possible to predict the correct AA sequence.

The traditional workflow for establishing biosynthetic pathways has been to work backwards from the compound of interest, by proposing a possible biosynthetic route using the same principles as those used for retro synthesis in classical organic chemistry: relying on the knowledge of possible enzyme catalyzed reactions. To further investigate the biosynthetic route, and to determine if a suggested route is correct further investigation is needed. In most cases, the next step would be to perform targeted gene deletions in the organism, followed by chemical analysis and isolation of compounds of interest, to determine the effect of the gene deletion. By deleting a biosynthetic gene, the production of enzymes encoded by that gene is stopped, and the enzyme is no longer present to catalyze formation of the compound. If several genes are involved in biosynthesis of a specific compound, deletion of one gene can lead to accumulation of an intermediate towards the compound of interest. An example of this is shown in **Paper 3**, where investigation of yanuthone D was performed. In one case, a strain of *A. niger* was prepared where the *yanH* gene had been deleted, and the corresponding YanH enzyme was therefore not produced. A comparison of the chemical profiles of an reference strain *A. niger* and the *yanH*Δ deletion strain showed that several new peaks appeared, while other peaks disappeared, as illustrated in Figure 10, where the BPC of the reference strain *A. niger* is compared to that from the *yanH*Δ deletion strain.

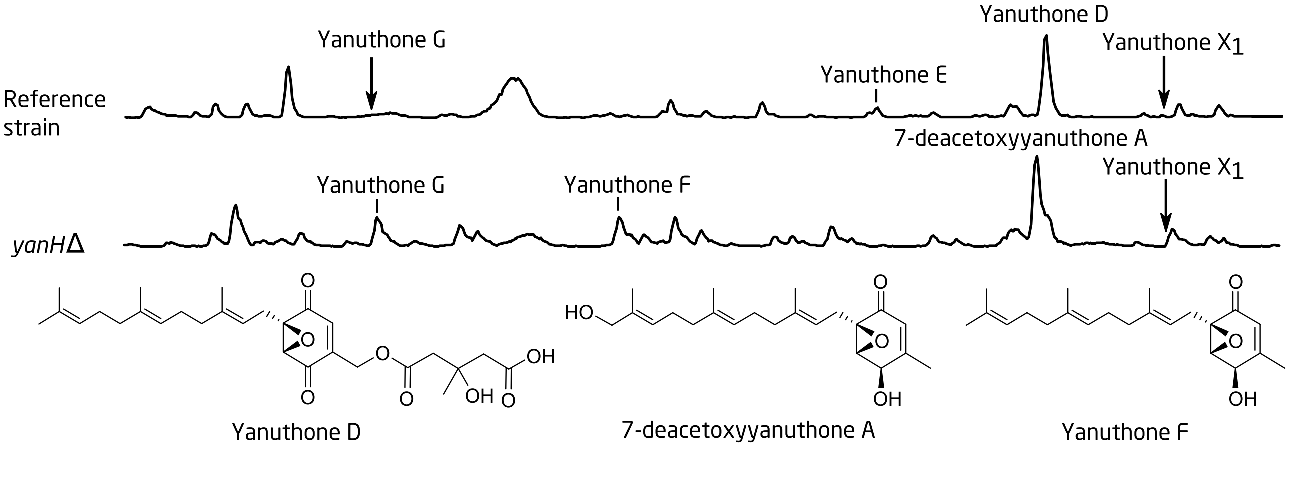


Figure 10 - Comparison of BPCs of reference strain *A. niger* strain (KB1001) (top) and deletion strain (*yanH*Δ). Gene deletions lead to altered chemical profiles of the strains, as the biosynthesis of compounds is interrupted. The disappearance of yanuthone D in the deletions mutants indicate that specific genes involvement in the biosynthesis (Paper 3).

Yanuthone D was produced by the reference strain *A. niger* (Figure 10), while the *yanH*Δ deletion strain did not produce it. Inspection of the BPCs did however show several new peaks compared to the intact strain, here identified as 7-deacetooxyyanuthone A and yanuthone F. These compounds were isolated, structure elucidated, and were found to be biosynthetic intermediates to yanuthone D, with YanH being responsible for conversion of 7-deacetoxyyanuthone to other intermediates, as described in **Paper 3**. By deleting the gene, higher amounts of 7-deacetoxyyanuthone accumulated instead of the end product, yanuthone D. To complete the elucidation of the biosynthesis more gene deletions were constructed by deleting the putative biosynthetic genes individually, until the whole biosynthesis of yanuthone D was characterized. Although very effective for elucidation of biosynthetic pathways, the individual deletion of genes is very labor-intensive.

2.2.1 Use of stable isotope labeled precursors

Another method for biosynthetic pathway elucidation is the use of stable isotope labeling (SIL). Biosynthesis studies by isotope labeling using radioactive labeled substrates is a well-known procedure that has been used since the 1950's (Hanahan and Al-Wakil 1952). The first experiments were carried out with radioactive isotopes using sensitive radiation detectors (Griffith 2004; Townsend and Christensen 1983). However, during the last 10 years advances in GC-MS and LC-MS instrumentation has made it possible to use stable isotope labelled nitrogen, carbon, and sulfur substrates for both kinetic, flux, and metabolite identification as the new mass analyzers are able to provide adequate sensitivity and resolution without the risks associated with working with radioactive material (Tang et al. 2012). One popular method is ^{13}C biosynthetic pathway elucidation, where a known precursor to a compound of interest is added to the cultivation media of an organism, and the resulting mass spectrum of the compound is then compared to the predicted ^{13}C labeling pattern (Simpson 1998; Steyn et al. 1984; Tang et al. 2012). In a study by Grunwald *et al.* the use of radioactive labeling and stable isotope labeling was compared for elucidation of metabolic products, showing that the two methods performed very similarly, though quantitative results were better for the radioactive labeling (Grunwald et al. 2013). However, stable isotope labeling has the great advantage of not requiring the use of potentially hazardous radioactive material.

SIL precursors are ideal for analysis by LC-MS. As the isotopes have the same chemical properties, they will have the same RT when analyzed using LC, but the compounds will have a different monoisotopic mass when analyzed using the MS, as shown in Figure 11.

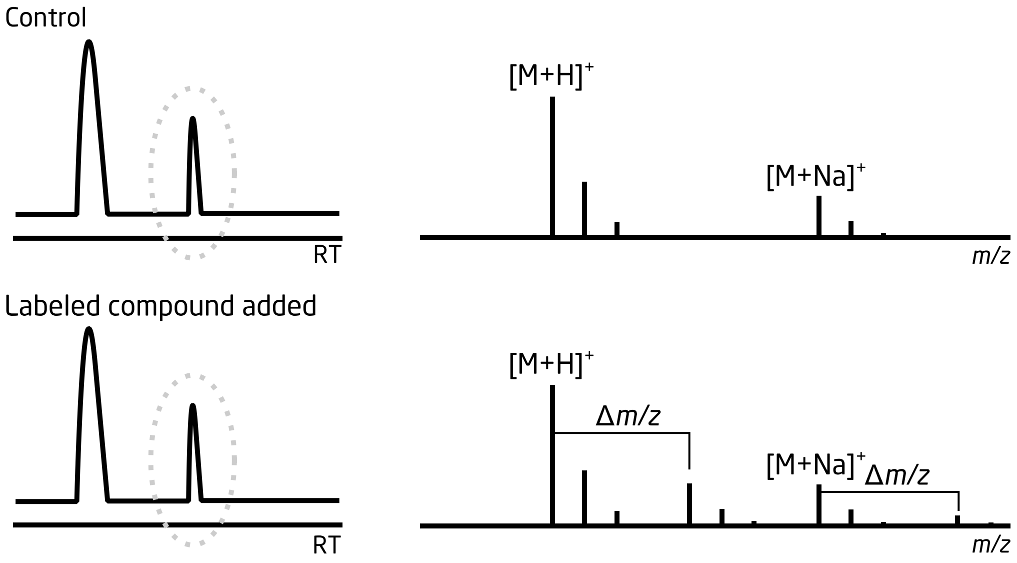


Figure 11 - Principle of addition of SIL precursor. Incorporation of the labeled precursor causes no change in chemical characteristics such as RT of the compound (dotted circle), but leads to a change in *m/z* of Δ, where Δ*m/z* is the mass difference between the labeled and unlabeled precursor.

SIL have been used in several studies of the aflatoxin pathway (Townsend and Christensen 1983), the asticolorin pathway (Steyn et al. 1984), and recently the yanuthone pathway (**Paper 3**) (Petersen et al. 2014).

The choice of using LC-MS for determination of labeling also influences the choice of SIL used for experiments. Some of the earliest ^{13}C labeling studies were carried out using doubly labeled acetate[1,2-$^{13}C_2$], which could then be used to trace the incorporation of intact acetate units into a wide range of metabolites. Samples would then be analyzed using NMR, as the two adjacent ^{13}C-atoms exhibit characteristic signals (Simpson 1998). For LC-MS, however, the use of precursors labeled with only ^{13}C is not optimal. To be able to determine that a compound has been labeled, one would have to observe an ion corresponding to the labeled compound. If one were to use a SIL containing only two labeled atoms, this ion might overlap with the A + 2 isotope of the unlabeled form of the compound complicating investigation of the labeled ion. Although this would lead to a change in intensity of that signal, it might not be possible to conclusively determine incorporation of a single acetate unit, if the degree of incorporation is very low.

In my studies I have worked on developing new protocols for the use of SIL precursors for investigation of the biosynthetic pathways using LC-MS. Different methods were developed for compounds of different biosynthetic origin, depending on whether they were PK or NRP derived.

2.2.2 Labeling of polyketide derived compounds - investigation of yanuthone D

Stable isotope labeling was used to investigate the biosynthesis of several different PKs and PK-like compounds. Initially, a method was developed for characterization of the yanuthone D biosynthesis. The compound yanuthone D was first isolated from *A. niger*, and described by Bugni and co-workers (Bugni et al. 2000). The yanuthone family of meroterpenoid derived compounds were described in detail in **Paper 3** and by Petersen and coworkers (Petersen et al. 2014). The complete study on the biosynthesis of yanuthone D and the use of stable isotope labeling is found in **Paper 3**.

As described in section 1.3.1, PKs are biosynthesized from a small number of different starter units. Because these are used for biosynthesis of a wide range of compounds, they can be unsuitable for investigation biosyntheses of specific compounds or for investigation of specific pathways. Instead one can use a more specific precursor, thereby targeting the biosynthesis of specific compound, ideally only leading to incorporation into compounds from the same biosynthetic pathway. By combining this with the developed targeted analysis methods, it was possible to quickly investigate compounds suspected of being biosynthesized from the SIL precursor, by creating a library containing all possible compounds of interest.

Based on initial genetic experiments, it was hypothesized that yanuthone D was biosynthesized from 6-MSA (Figure 13). Labeling experiments using ^{13}C-labeled 6-MSA were therefore performed to investigate if it was possible to add the labeled precursor to the growth medium of the fungus, and for the fungus to take up and incorporate the precursor into the biosynthesis of yanuthone D. As labeled 6-MSA was not commercially available, it was produced in-house by fermentation of a genetically modified heterologous producer strain. The labeling experiment was performed by inoculating *A. niger* on solid growth medium, and then adding the labeled precursor in solution. After cultivation plugs of the fungi were excised, extracted, and analyzed using LC-MS, as described in Figure 12.

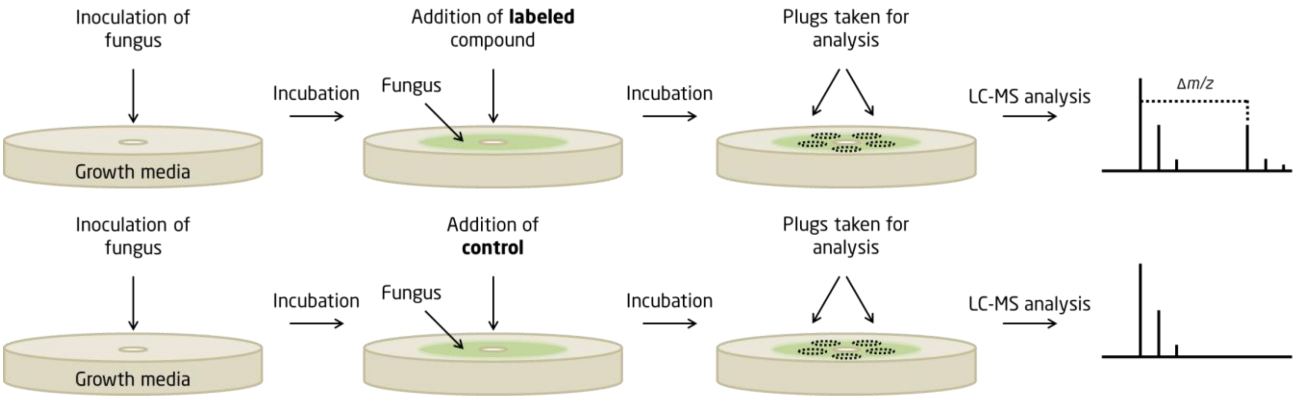


Figure 12 - Experimental procedure used for labeling of fungi. Incubation time before and after addition of labeling solution varied in different experiments. The labeling solution contains a SIL precursor that induces a Δ*m/z* shift when incorporated.

Analysis by LC-MS showed that the fungus successfully took up the ^{13}C-labeled 6-MSA, and by examining the mass spectra, it was possible to detect a shift in mass for compounds incorporating 6-MSA. Using a combination of gene deletions and labeling with a SIL precursor, it was possible to elucidate the biosynthesis of yanuthone D, as shown in Figure 13. In the figure, the labeled carbon atoms originating from 6-MSA are marked in red.

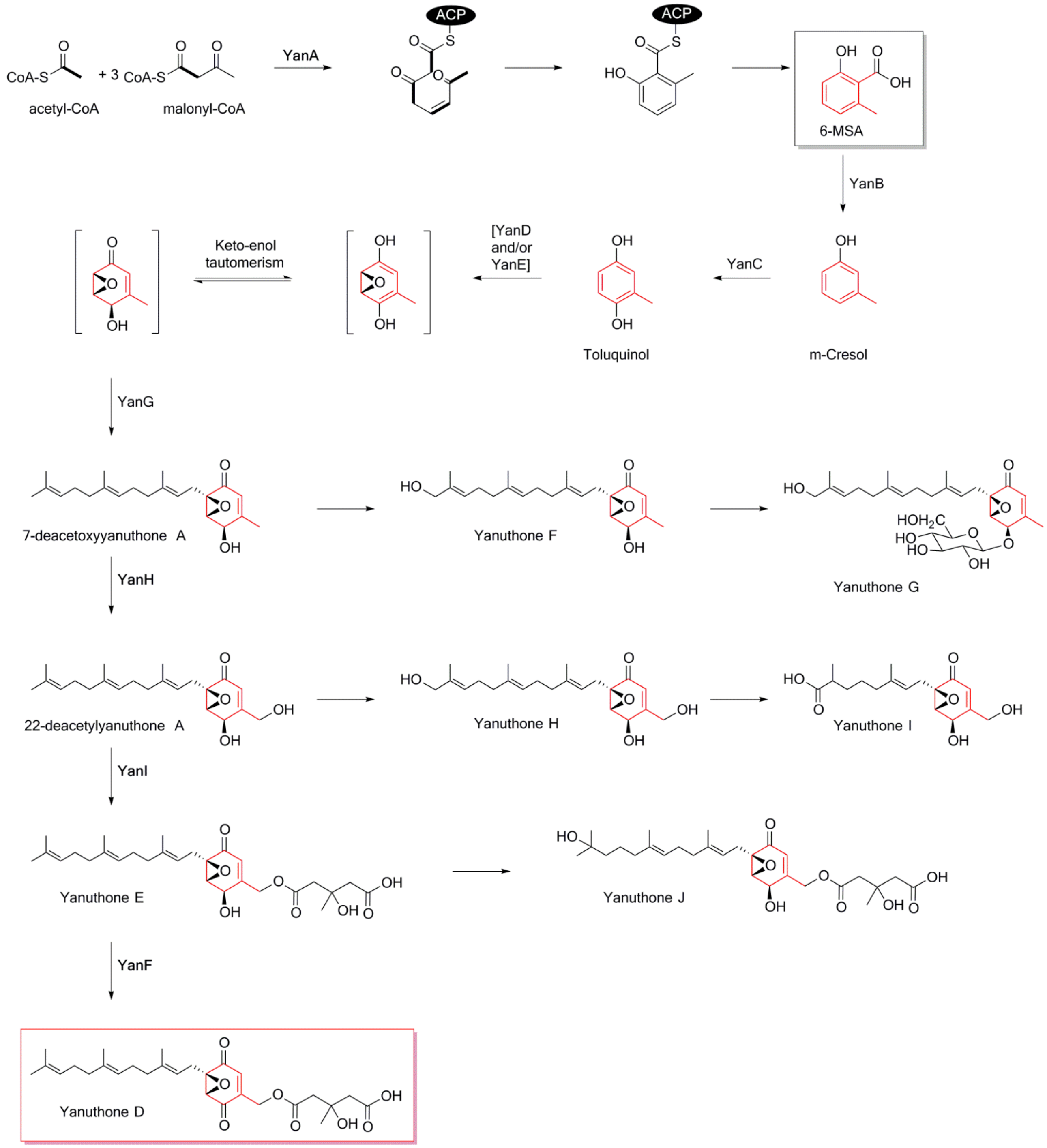


Figure 13 - Biosynthesis of yanuthone D. $^{13}C_8$-labeled 6-MSA (black box) was added to *A. niger* and 6-MSA was taken up and incorporated into the biosynthesis of yanuthone D (red box). As one ^{13}C-atom is lost due to decarboxylation, $^{13}C_7$ is incorporated into yanuthone D. The labeled carbon atoms originating from 6-MSA are shown in red (Paper 3).

Analysis using LC-MS showed that the incorporation degree of the labeled precursor into yanuthone D was around 18 %. Incorporation of 6-MSA was also high enough for labeled compounds to be spotted by a cursory look at the mass spectra, which accelerated the determination of which compounds were biosynthesized from 6-MSA.

Interestingly, analysis of samples fed with SIL 6-MSA showed that one yanuthone, yanuthone X_1 (Figure 14), did not exhibit any sign of incorporating the precursor, indicating that this compound is not biosynthesized from 6-MSA.

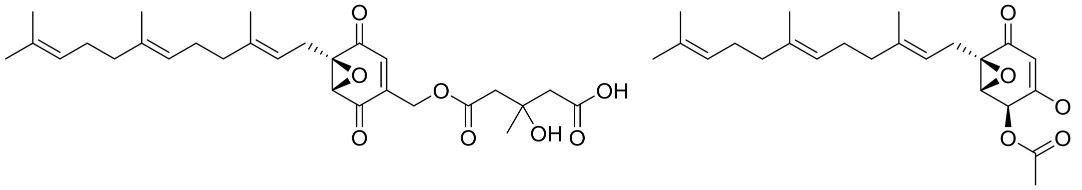


Figure 14 - Left) yanuthone D. Right) yanuthone X_1

Further experiments proved that, although structurally very similar to the other yanuthones, yanuthone X_1 was not biosynthesized from 6-MSA but instead from a still unknown precursor, highlighting the strength of the labeling method for quickly investigating biosynthesis. When the yanuthones were first discovered, Bugni and co-workers speculated that the yanuthones were biosynthesized from shikimate, a product of the shikimic acid pathway (Bugni et al. 2000). Further studies into the yanuthones have revealed a second yanuthone, yanuthone X_2, that is not biosynthesized from 6-MSA (Petersen et al. 2014). It would therefore be very interesting to conduct further labeling studies, this time using predicted precursors from the shikimic acid pathway to investigate these class II yanuthones of unknown biosynthetic origin.

2.2.3 Labeling of other polyketides

Based on the successful labeling of yanuthone D, I decided to further explore the applications of SIL precursors for the investigation of PK biosynthetic pathways, and to further developed methods for its use. This study is described in detail in **Paper 4**.

Initial feeding was carried out in seven different fungi: *P. griseofulvum*, *P. paneum*, *P. carneum*, *A. clavatus, B. nivea*; or terreic acid: *A. hortai*, and *A. floccosus.* These were attempted labeled using both SIL $^{13}C_8$-6-MSA, for labeling of patulin and terreic acid, respectively, using the same experimental setup as described in Figure 12. Again, the data analysis was performed using a library of compounds for targeted analysis. A library was created containing all compounds, and predicted precursors of these, believed to be biosynthesized from 6-MSA. In theory, screening of samples should therefore quickly reveal any compounds showing any signs of incorporation. Structures of the compounds described in the text are shown in Figure 15.

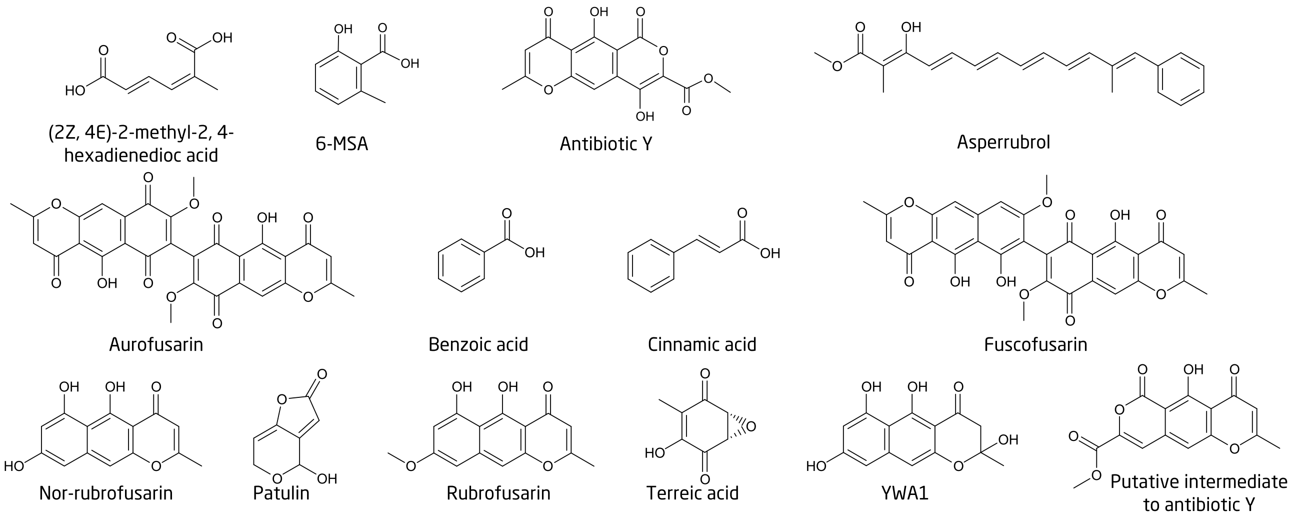


Figure 15 - Structures of PK compounds referenced in text.

Unfortunately, LC-MS analysis showed no signs of incorporating 6-MSA into patulin or terreic acid for any of the tested fungi. Results from the labeling experiments are summarized in Table 5. It was surprising that no incorporation was detected for patulin or terreic acid, as 6-MSA is a known precursor to both compounds (Guo et al. 2014; Tanenbaum and Bassett 1959), and thus we hypothesized that this could be caused by the fungus degrading the 6-MSA, as chemical analysis showed that 6-MSA was expended from the medium. Another explanation could be that the enzymatic activities involved in biosynthesis are linked in a manner that does not allow entry of an “external” precursor. A recent paper by Guo and coworkers (Guo et al. 2014) showed that (2Z,4E)-2-methyl-2,4-hexadienedioic was a shunt product in the terreic acid pathway, and we subsequently detected a peak corresponding to the correct accurate masse in an extract from *A. floccosus.* Investigation of the mass spectrum also revealed the presence of an ion corresponding to one with $^{13}C_7$ incorporated. For the work performed here, the degree of labeling was defined as:

$$\text{Degree of labeling} = \frac{\text{Signal}_{\text{labeled form}}}{\text{Signal}_{\text{labeled form}} + \text{Signal}_{\text{unlabeled form}}}$$

For (2Z, 4E)-2-methyl-2,4-hexadienedioic the degree of labeling was thus 76 % in *A. floccosus* fed after 3 days. Interestingly (2Z,4E)-2-methyl-2,4-hexadienedioic was also found in the extracts from the patulin producers, where it was also labelled, indicating that it is also a shunt product in the patulin biosynthesis. This strongly indicates that it is a result of detoxification reaction in the cytoplasm, and that patulin and terreic acid are produced in compartments, as is the case for aflatoxin production in *A. parasiticus* (Chanda et al. 2009). This would make sense as patulin is an antifungal compound. The need for a detoxification process also seems to be important as (2Z,4E)-2-methyl-2,4-hexadienedioic was detected in amounts corresponding to 10-20 % of the amount of patulin produced, as determined using by analysis UV/Vis peak areas, measured using the DAD at 280 nm. In order to investigate the hypothesis that production was taking place in a compartmentalized fashion, the genome sequence from the terreic acid gene cluster (Guo et al. 2014) was analyzed in order to predict any membrane bound proteins, using a range of different prediction tools. However, no conclusive results were obtained.

Table 5 - Results obtained from PK labeling experiments, where only the highest determined degree of incorporation is listed along with the producer organism. No incorporation detected (ND).

Target compound	Producer organism	Precursor	Time of precursor addition (d)	Degree of incorporation (average of duplicates)
Patulin	*P. griseofulvum,* *P. paneum,* *P. carneum,* *A. clavatus,* *B. nivea*	6-MSA	3	ND
(2Z,4E)-2-methyl, 4-hexadienoic acid				45 % *(A. clavatus)*
Terreic acid	*A. hortai,* *A. floccosus*	6-MSA	3	ND
			6	ND
(2Z,4E)-2-methyl, 4-hexadienoic acid			3	76 % *(A. floccosus)*
			6	58 % *(A. floccosus)*
Asperrubrol	*A. niger*	Cinnamic acid	3	ND
			6	ND
		Benzoic acid	3	1.3 %
			6	ND
Aurofusarin	*F. avanaceum,* *F. graminearum*	YWA1	3	1.2 % (*F. graminearum* DFM)
			7	0.3 % (*F. graminearum* Bell's)
			10	0.4 % (*F. graminearum* Bell's)
Antibiotic Y			3	ND
			7	0.7 % (*F. avanaceum* Bell's)
			10	0.4 % (*F. avanaceum* Bell's)
Rubrofusarin			3	0.4 %(*F. graminearum* Bell's)
			7	10 % (*F. graminearum* Bell's)
			10	17 % (*F. graminearum* Bell's)
Putative intermediate to antibiotic Y			3	ND
			7	2.2 % (*F. avanaceum* Bell's)
			10	2.2 % (*F. avanaceum* Bell's)

In another labeling experiment fully ^{13}C-labled YWA1, produced in-house by fermentation of a genetically modified producer strain was used. This precursor was tested for labeling of compounds in four different strains of *Fusarium* using the same experimental setup as described in Figure 12. Addition of the labeling solution resulted in the labeling of several different compounds as seen in Table 5. However, for many of the compounds, the degree of incorporation was lower than what was observed in the yanuthone study. Because of this, it was harder to determine if a compound had been labeled purely by visual inspection of the data.

Compared to other published studies, the incorporation degrees obtained in this study range from typical to very high. In a study of the mycotoxin terretonin by McIntyre *et al.* incorporation of several different differentially labeled precursors was investigated (McIntyre et al. 1989). Incorporation was reported to range from 0.3-2.5 % depending on the precursor or cultivation conditions usedA study by Yoshizaws *et al.* investigated the incorporation of acetate in the biosynthesis of dehydrocurvalarin and found that these were incorporated at around 2 % (Yoshizawa et al. 1990). Finally, Yue et al. reported incorporation of 6 % for an investigation of macrolide biosynthesis (Yue et al. 1987).

As a consequence of the detected determined incorporation rates, a targeted approach was used to screen for compounds that were predicted to be labeled, based on structures and theoretical biosynthetic intermediates. Using this approach four different compounds were found to be labeled, and thus biosynthetically derived from YWA1, see Figure 16.

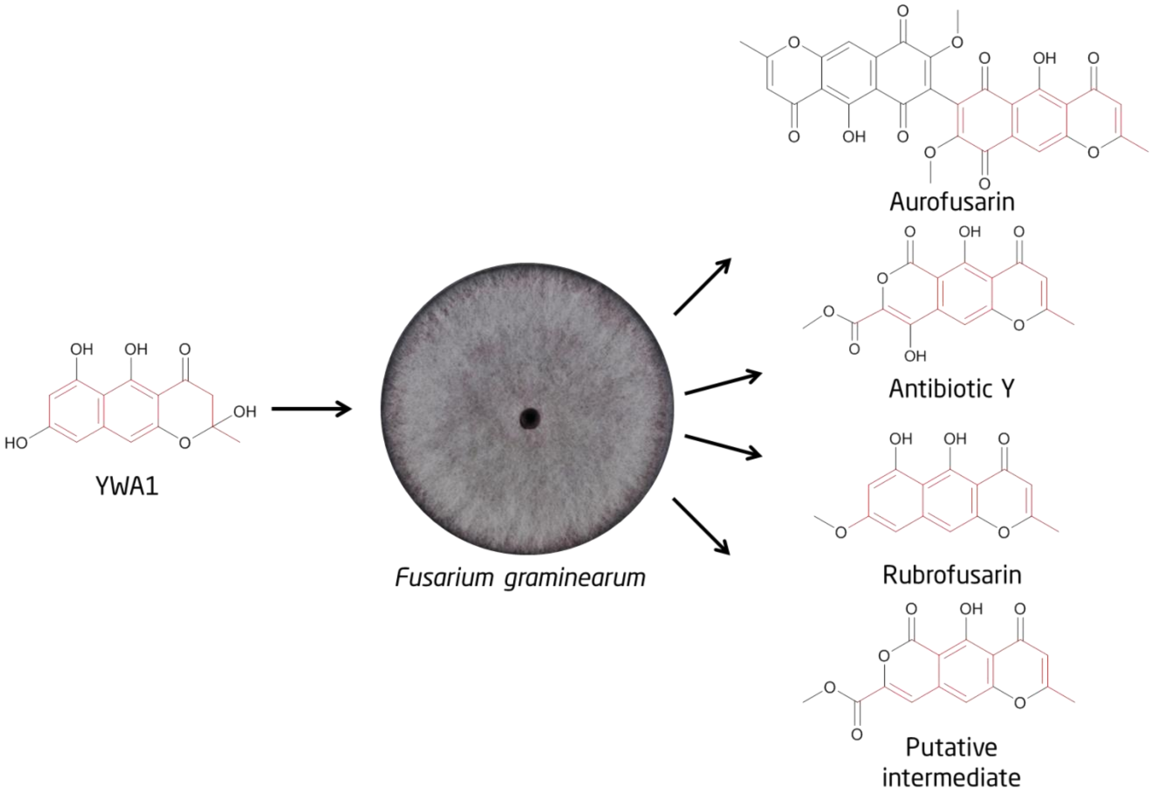


Figure 16 - Addition of ^{13}C-labeled YWA1 to *F. graminearum* lead to the production of several different ^{13}C-labeled products demonstrating that these were biosynthesized from YWA1. The atoms marked in red in the YWA1-molecule are ^{13}C-atoms, while the red atoms in the products are thought to be directly incorporated from YWA1.

One of these compounds was antibiotic Y (avenacein Y). This was first isolated form *F. avenaceum* in 1986 and its biosynthetic origin is unknown (Goliński et al. 1986), however, it displays several structural features in common with YWA1 and rubrofusarin. Based on the mass spectrum obtained for antibiotic Y, shown in Figure 17, it is indeed biosynthesized from YWA1 with incorporation of around 2.2 %.

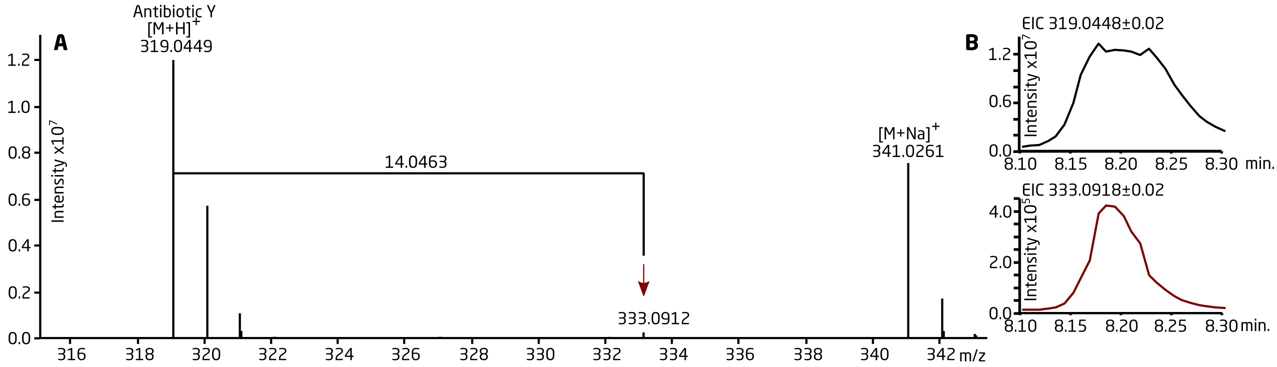


Figure 17 - A) A) Mass spectrum extracted at RT 8.2 min with [M+H]$^+$ (*m/z* 319.0449) and [M+Na]$^+$ (*m/z* 341.0261) pseudomolecular ions corresponding to antibiotic Y. Mass shift of $^{13}C_{14}$ suggest incorporation of labeled YWA1 (red arrow). B) EICs corresponding to antibiotic Y (upper) and antibiotic Y with $^{13}C_{14}$ (lower) (Paper 4).

EICs corresponding to unlabeled antibiotic Y and antibiotic Y with $^{13}C_{14}$ incorporated (Figure 17B) exhibited similar peak shapes and RT, confirming that the labeled YWA1 precursor is incorporated into Antibiotic Y. The unlabeled form was present in high enough amounts to saturate the detector, leading to a non-linear response curve. To calculate the degree of incorporation, the intensity of the [M+H]$^+$ + 1 ion, which was not saturated, was used. Using the predicted abundance of the isotopes, the degree of incorporation of the labeled Antibiotic Y was calculated to be 0.4 %.

Results from the labeling experiments demonstrate that SIL precursors can be very effective for investigation of biosynthetic pathways. A comparison of incorporation for the different PK precursors showed that the incorporation degree varied widely between organisms and compounds. Based on this, it would be interesting to further investigate the uptake of precursors by fungi. One explanation for the variation in incorporation degrees could be that the enzymatic activities involved in biosynthesis are linked in a manner that does not allow entry of an "external" precursor. One such linkage could be formation of a protein complex consisting of several discrete enzymes, which are dependent on each other for proper conformation, like the so-called metabolon model, which has been proposed for the tricarboxylic acid cycle (Meyer et al. 2011; Vélot et al. 1997).

Alternatively, biosynthesis of the toxic compounds is compartmentalized in specialized organelles, into which an external precursor is not transported, as suggested for patulin or terreic acid. As the feeding experiments with patulin and terreic acid showed, formation of shunt products acting as sinks for the SIL precursor could also explain the missing labeling of the desired end products. A reason for the low degrees of labeling observed in both the experiments performed in this study, as well as studies performed by others, could be due to unknown shunt products being labeled instead of the investigated one, as was the case for patulin and terreic acid.

As a next step it would be interesting to perform quantitative analysis to accurately determine how much of the added SIL precursor is taken up by the organism, and further, how much is incorporated into any other compounds.

2.2.4 Labeling of nonribosomal peptide derived compounds - investigation of nidulanin A

One of the first projects I worked on during my studies was investigation of metabolites from *A. nidulans.* This work resulted in discovery and identification of the metabolite nidulanin A, see Figure 18. Nidulanin A is a cyclic tetrapeptide consisting of one L-phenylalanine (Phe) residue, one L-valine (Val) and one D-Val residues, one L-kynurenine residue, and one isoprene unit. In the original study (**Paper 5**), nidulanin A proved difficult to isolate in the quantities needed for structure elucidation by NMR, and thus two putative analogs containing one and two additional oxygen atoms, respectively, produced in lower quantities were not isolated and fully characterized. Because of this, it was investigated whether the structure of any of the new analogs could be determined only using LC-MS.

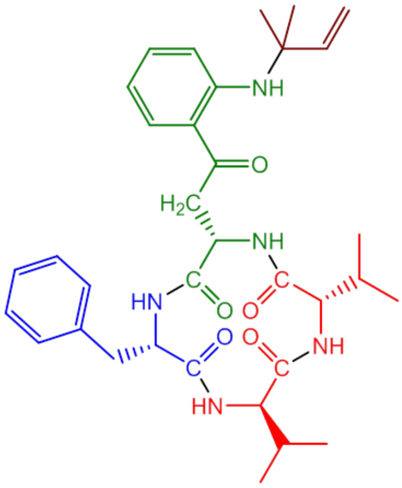


Figure 18 - Structure of nidulanin A. The coloring illustrates the different biosynthetic units that make up the metabolite: Blue -Phe, green - kynurenine, Red - Val, and brown - isoprene(Paper 6).

To do this it was decided to use SIL amino acids (SILAAs), as fungi are known to be able to take up AAs from their environment (Helmstaedt et al. 2007). This property has previously been exploited to study incorporation of labelled AAs into proteins from filamentous fungi using LC-MS (Collier et al. 2008; Georgianna et al. 2008). SILAAs might therefore be a suitable route for introducing NRP precursors into fungi to probe the NRP pathway like nidulanin A. This study is described in detail in **Paper 6**.

By utilizing information about the structure of nidulanin A, feeding studies were performed using SILAAs. In the experiment, *A. nidulans* was cultivated in liquid media, and several different concentrations of AAs were tested to determine the optimum for incorporation using LC-MS (Figure 19A). Samples were then analyzed using LC-MS/MS to provide structural information, as well as to perform molecular network analysis (Figure 19B).

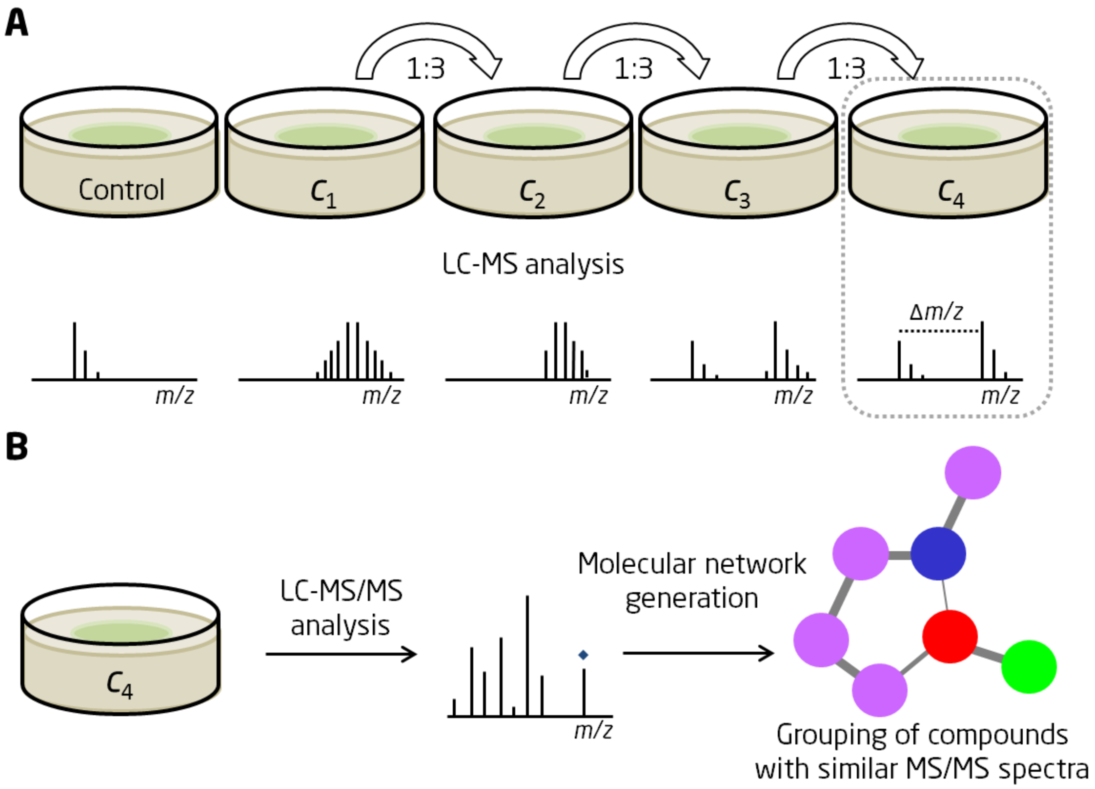


Figure 19 - Experimental setup for labeling of NRPs. A) Samples were extracted, and then analyzed using LC-MS to determine the optimal concentration of AAs for incorporation from a dilution series. B) Samples selected from A were analyzed via LC-MS/MS to provide structural information and for subsequent molecular network generation.

For the experiments, five different AAs (Table 6) were used and a majority of those used were fully ^{13}C-labled. It was not possible procure kynurenine. Instead, anthranilic acid was used, as it is a precursor to kynurenine. Addition of SILAAs to *A. nidulans* resulted in the incorporation into nidulanin A, as seen in Figure 20.

Table 6 - This table contains information about the SIL AAs used in the experiment. The chemical formula denotes the formula of the AAs and indicate the labeled atoms present. The mass difference denotes the mass difference between the SIL labeled AA and the natural occurring isotype. *Mass difference due only to ^{13}C-labeling.

AA	Elemental composition	Monoisotopic mass [Da]	Mass difference [Da]
Phe	$^{13}C_9H_{11}{}^{15}NO_2$	175.1062	10.0272 (9.0302)*
Val	$^{13}C_5H_{11}NO_2$	122.0958	5.0168
Anthranilic acid	$^{13}C_6{}^{12}CH_7NO_2$	143.0678	6.0201
Trp	$C_{11}D_8H_4N_2O_2$	212.1401	8.0502
Tyr	$^{13}C_9H_{11}{}^{15}NO_3$	191.1011	10.0272 (9.0302)*

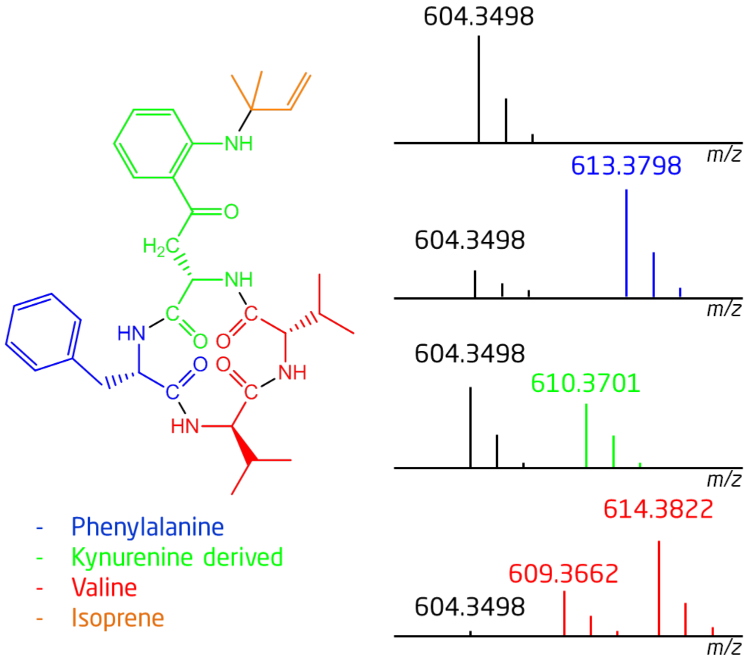


Figure 20 - Structure of nidulanin A (left) and mass spectra (right) obtained from LC-MS analysis after feeding with ^{13}C-labeled AAs. The black signals denote the unlabeled form, and the colored signals denote the signal arising from feeding with labeled AAs. Mass spectra are for illustrative purposes only and do not reflect the precise incorporation degree of the AAs.

Results from the feeding experiments could be used to determine which AAs nidulanin A was composed of, as well as provide information about the reported oxygenated analogs first described in **Paper 5**. By using labeled tyrosine (Tyr) it was possible to detect incorporation of the oxygenated analogue, confirming that the oxygenated form did indeed contain a Tyr residue. However, it was not possible to determine the structure of the analog containing two extra oxygen atoms.

Analysis of the MS/MS spectra obtained from nidulanin A could be used to determine the sequence of AAs present in the cyclic tetrapeptide, by utilizing the information provided by the labeling. Using this, the MS/MS spectrum of nidulanin A could be assigned. Fragmentation spectra of the labeled forms of nidulanin A, as well as a list of assigned fragments are shown in **Paper 6**. The fragmentation spectrum of unlabeled nidulanin A, as well as the most characteristic fragments allowing for determination of the AA sequence is shown in Figure 21.

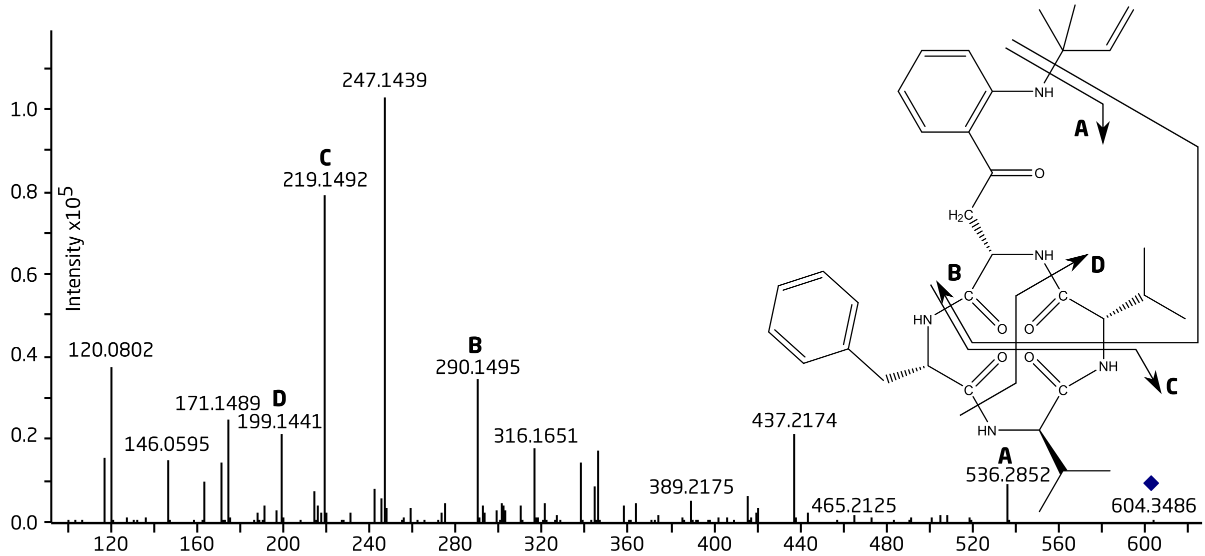


Figure 21 - MS/MS spectrum from nidulanin A. Indicated fragments can be used to determine the sequence of the tetrapeptide.

Samples from the labeling experiments were then analyzed using LC-MS/MS to obtain fragmentation data of the metabolites, and the data could be used to perform the molecular network generation. MS/MS spectra that exhibit the same fragment ions or the same neutral losses will be connected in the network with the thickness of the line indicates a better match or higher similarity of spectra. The mass spectrum from a given compound, a node, will then be clustered together with compounds having similar MS/MS spectra. Biosynthetically similar compounds might therefore be grouped using the generated molecular networks, aiding in characterization of the biosynthesis. A molecular network was generated using the samples labeled with AAs, and the sub-network containing the node corresponding to nidulanin A is depicted in Figure 22.

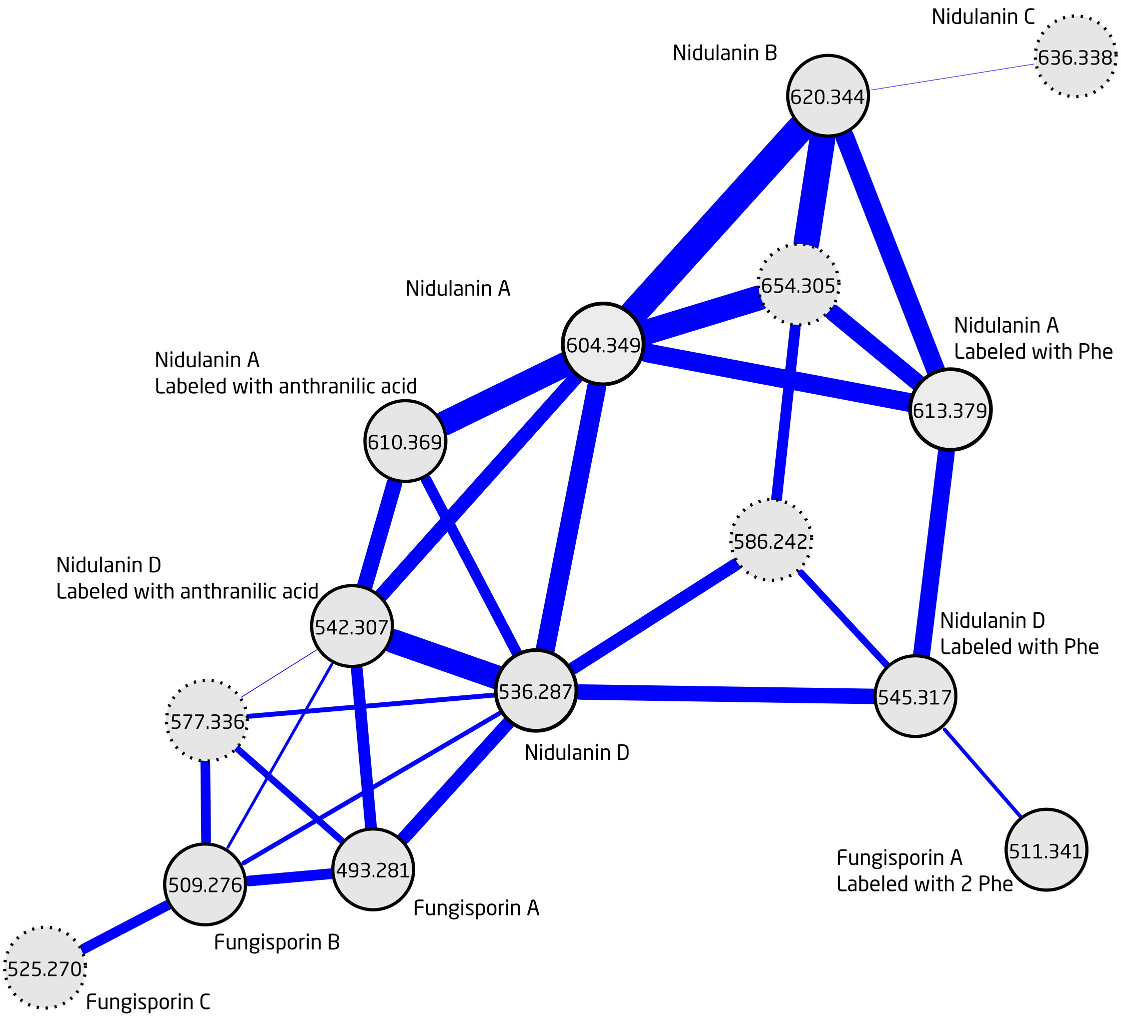


Figure 22 - Molecular network containing nidulanin A, and structurally related compounds. The thickness of the blue connecting line indicates a higher correlation of similarity for MS/MS spectra. Previously undescribed compounds are marked with a dashed outline.

Investigation of the sub-network containing nidulanin A revealed several nodes corresponding to labeled forms of nidulanin A, but it also identified nodes corresponding to unknown compounds. Utilizing both the LC-MS data as well as the LC-MS/MS data, it was possible to tentatively identify several of the compounds corresponding to nodes in the sub-network, as seen in Table 7.

Table 7 - Investigated compounds. The column labeling information lists the number of specific labeled AA residues detected for each compound (-) - no detection of incorporation, *- Compound also described by Ali and coworkers (Ali et al. 2014)**.**

Name	RT [min]	Molecular formula	Amino acid composition	Modification	*m/z* $[M+H]^+$	Labeling information				
						Phe	Val	Ant	Tyr	Trp
Nidulanin A	8.7	$C_{34}H_{45}N_5O_5$	Phe-Kyn-Val-Val	Prenylated	604.3493	1	2	1	-	-
Nidulanin B	7.7	$C_{34}H_{45}N_5O_6$	Tyr-Kyn-Val-Val	Prenylated	620.3443	1	2	1	1	-
Nidulanin C	7.3	$C_{34}H_{45}N_5O_7$	Not determined	Prenylated	636.3392	1	2	1	1	-
Nidulanin D	7.0	$C_{29}H_{37}N_5O_5$	Phe-Kyn-Val-Val		536.2867	1	2	1	-	-
Fungisporin A	7.3	$C_{28}H_{36}N_4O_4$	Phe-Phe-Val-Val*		493.2809	2	2	-	-	-
Fungisporin B	6.3	$C_{28}H_{36}N_4O_5$	Tyr-Phe-Val-Val*		509.2758	2	2	-	1	-
Fungisporin C	5.4	$C_{28}H_{36}N_4O_6$	Tyr-Tyr-Val-Val		525.2708	1	1	-	1	-
	7.3	$C_{33}H_{44}N_4O_5$	Not determined		577.3384	2	2	-	1	-
	7.9	$C_{35}H_{39}N_7O_6$	Not determined	Prenylated	654.3035	1	-	1	1	-
	6.7	$C_{30}H_{31}N_7O_6$	Not determined		586.2401	1	-	1	1	-
Fungisporin D	7.2	$C_{30}H_{37}N_5O_4$	Phe-Trp-Val-Val*		532.2924	1	2	1	-	-

Compared to the PK labeling study, the incorporation degrees were much higher when using SILAAs as precursors for labeling compounds in fungi. Because of this, it was possible to use the labeling in a much more exploratory manner. By combining the labeling procedure with the molecular network, it was possible to find new compounds, while the MS/MS data obtained could be used to determine the order in which the AAs were coupled in the cyclic tetrapeptide nidulanin A, and could be used to tentatively determine the structures of the metabolites. This proved instrumental in the analysis of the compounds, as they were produced in minute amounts precluding structure elucidation by NMR.

The molecular network revealed the presence of the compound fungisporin (Miyao 1955) as well as two analogs of this. Using information from the labeling experiments, it was hypothesized that these two analogs (fungisporin B) and (fungisporin C), corresponded to the exchange of one and two Phe residues for Tyr, respectively, which was confirmed using the labeling studies. The production of fungisporin has recently been linked to a specific NRPS, HcpA, in *P. chrysogenum* by Ali and coworkers (Ali et al. 2014). In that study 10 different cyclic tetrapeptides were found to be produced by the NRPS, including fungisporin and an analogue containing a Tyr instead of a Phe residue. By utilizing the labeling information it was possible to determine the peptide sequence of fungisporin C to be cyclo-(Phe-Phe-Tyr-Tyr).

Analysis of an *A. nidulans* deletion strain demonstrated that nidulanin A and fungisporin, as well as their respective analogs, were encoded by the same NRPS, thus highlighting the strength of the molecular networking method in correlating compounds with structural similarities.

Interestingly, an entry for fungisporin exists in Antibase (Laatsch 2012), however both the structure and molecular formula are wrong. Fungisporin's entry in Antibase references the Dictionary of antibiotics and related substances (Bycroft 1988), which contains a different structure than the one published by Miyao (Miyao 1955). The reason for this seems to be that the structure corresponds to a formulation prepared as a salt containing several fungisporin units. This highlights a very important point about these databases: that the curation procedures and quality controls are unknown. The fact that fungisporin has not previously been reported from *A. nidulans*, despite all the research in the organism, maybe also indicates that a lot of research groups use the same standard libraries for dereplication. This topic will be further discussed in chapter 3.2.

To put the results of the labeling study into perspective, an estimate of the total amount of AAs present in the fungus compared to the amount of added SILAA was made. Based on the parameters reported by Stephanopoulos *et al.* (Stephanopoulos et al. 1998) for *P. chrysogenum*, it was possible to give a rough estimate on the amount of the specific AAs produced by *A. nidulans* in the performed labeling experiments, summarized in Table 8. Unfortunately, the total dry weight of the fungus cultivated in each well was not measured. For the following calculations it was therefore estimated to be 0.10 g.

Table 8 - Amount of AA estimated to be produced by *A. nidulans* based on a dry weight of 0.10 g. All values for typical compositions from Stephanopoulos and coworkers (Stephanopoulos et al. 1998)**.**

		Typical composition	Estimated amount in well
Protein		0.45 [g (g dry weight)$^{-1}$)]	0.45 g
AA	Phe	3.4 mole % of protein	1.5×10^{-3} g
	Val	6.4 mole % of protein	2.9×10^{-3} g
	Tyr	2.6 mole % of protein	1.2×10^{-3} g

Based on the concentration of the AAs used in the labeling experiment, the amount of AA added to each well, containing 1.6 ml medium, could then be calculated, as seen in Table 9.

Table 9 - Concentration and added amounts of AAs used in labeling experiment. Blue shading corresponds to a higher amount of SILAA added than AA produced by the fungus and green a lower amount.

AA	c_1 [M]	m_1 [g]	c_2 [M]	m_2 [g]	c_3 [M]	m_3 [g]	c_4 [M]	m_4 [g]
Phe	1.7×10^{-2}	4.8×10^{-3}	5.7×10^{-3}	1.6×10^{-3}	1.9×10^{-3}	5.3×10^{-4}	6.4×10^{-4}	1.8×10^{-4}
Val	4.7×10^{-3}	9.2×10^{-4}	1.6×10^{-4}	3.1×10^{-5}	5.2×10^{-5}	1.0×10^{-5}	1.7×10^{-6}	3.3×10^{-6}
Tyr	2.9×10^{-2}	8.9×10^{-3}	9.7×10^{-3}	3.0×10^{-3}	3.2×10^{-3}	9.8×10^{-3}	1.1×10^{-4}	3.4×10^{-4}

Several caveats apply to the proposed estimates. Firstly, the dry weight of the fungus has not been experimentally determined but rather estimated. Secondly, the typical compositions have been determined from *P. chrysogenum* and not *A. nidulans* which was used in the experiment. Finally, the specific AA composition of the proteins has been determined in mole %, and not mass % as assumed for these calculations. Based on these calculations, it is observed that the amount of SILAA added to the organism at the highest concentrations, at least for Phe and Tyr, are 100 times higher than the amount of AA produced by the fungus. As shown in Figure 23, addition of high levels of SILAA caused distorted mass spectra, as the SILAAs enter the central carbon metabolism and are catabolized instead of being incorporated directly into any metabolites. At lower concentrations, the AA appeared to be preferentially incorporated into nidulanin A and not catabolized in the same degree. Labeling using Val resulted in incorporation to such a high degree that no unlabeled nidulanin was detected at (c_1), in spite of nidulanin containing two Val residues.

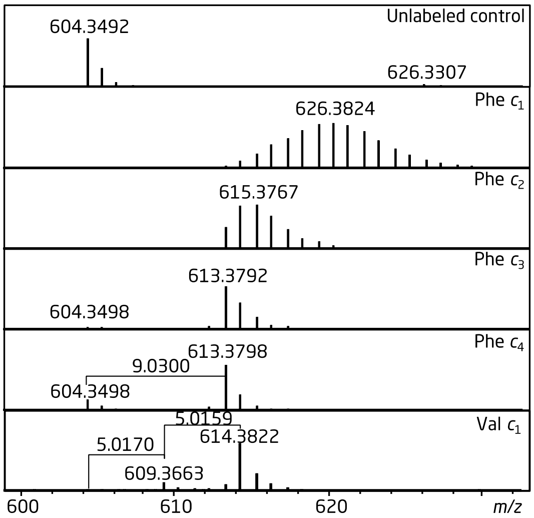


Figure 23 - Incorporation of SILAAs into nidulanin A.

Based on the labeling results obtained in this experiment, it would be interesting to further investigate the labeling patterns as a function of concentration of the different labeled SILAAs. Nidulanin A exhibited a higher degree of labeling with two Val residues than one residue, something also observed for the fungisporins as described in **Paper 6**, further suggesting that the amounts of SILAAs used were very high. Potentially, it would be possible to use lower concentrations of SILAAs, while still obtaining the same results.

The combination of stable isotope labeling and molecular network generation was shown to be very effective for detection of structurally related NRPs, while labeling was effective for determination of the peptide sequence, and could be used to provide information on biosynthesis of compounds. The fact that these compounds have not been reported before, also highlight the ability of the combined approach to extract spectral features from compounds that might otherwise be overlooked. This was the case for fungisporin and its two different analogs that had not previously been reported from *A. nidulans*. This illustrated the strength of the untargeted molecular networking generation in extracting structurally related but unknown compounds, and coupling these to known compounds and aiding in dereplication.

2.3 Untargeted analysis for profiling of the biosynthetic potential of *Pseudoalteromonas luteoviolacea*

Dereplication based methods, as the ones presented in section 2.1, were methodologies employing information from previously assembled libraries for analysis of samples. In cases where little or no information about an organism is available, other methods of data analysis must therefore be used. This was also the case for the study of biosynthetic potential of the marine bacterium *Pseudoalteromonas luteoviolacea*. For the study 13 different strains were isolated from around the globe, and the goal was to examine the biosynthetic potential of all these strains. Some information about produced metabolites was available, however, besides determining whether any of these metabolites were produced, the goal was to determine how *all* produced metabolites varied between the 13 strains.

In order to accurately assess the functional biosynthetic potential of the organisms, a method for combining both LC-MS based metabolomics, machine learning algorithms for data mining, mass spectral molecular networking, and genomics was developed and used to evaluate the biosynthetic richness of these marine bacteria. The study is described in detail in **Paper 7**. The combination of machine learning principles for analysis of chemical data, and the integration between LC-MS based metabolomics and genomics have not previously been used, and thus the developed combined method represents a whole new approach for the profiling the biosynthetic potential of a group of organism.

In this section, I will briefly present and discuss the results stemming directly from the untargeted analysis performed. In this study 13 different strains related to *P. luteoviolacea* were analyzed for their genomic potential and ability to produce SMs. Results from this analysis could then be used to determine which strains should be further investigated, effectively prioritizing the most chemically prolific species. An overview of the experimental work performed, as well as the data analysis, is shown in Figure 24.

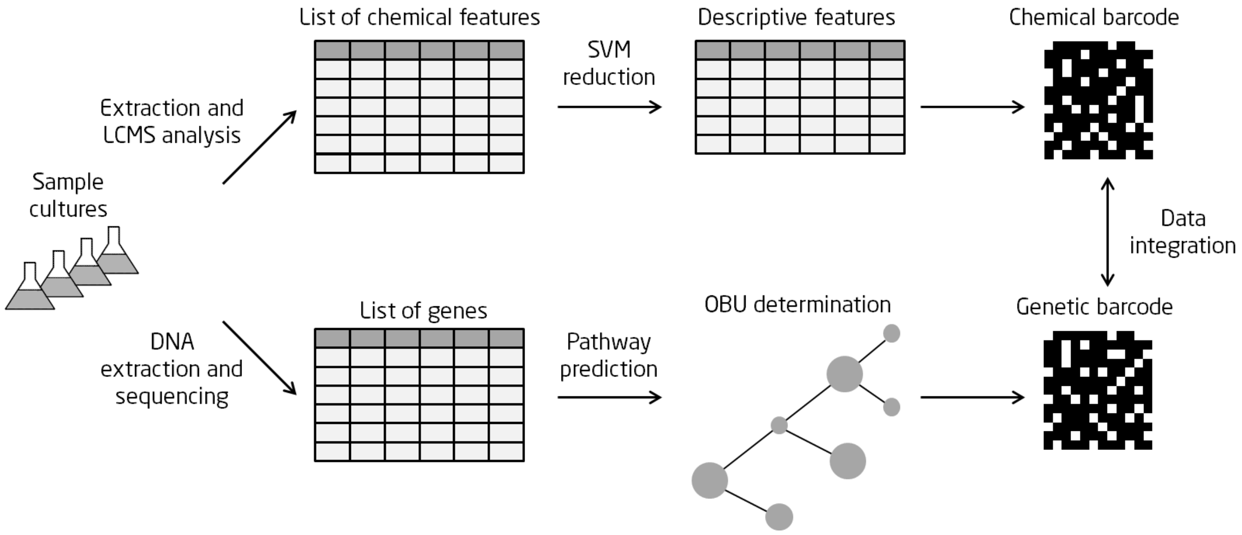


Figure 24 - Overview of experiments and data analysis performed in for the *P. luteoviolacea* projectSupport vector machine (SVM). Operational biosynthetic unit (OBU).

For my part the focus was on the chemical analysis of the extracts, which was performed using UHPLC-DAD-QTOF analysis. To obtain a global, unbiased view of the metabolites produced, molecular features were detected using the LC-MS in an untargeted metabolomics experiment, as described in section 1.5. As the workflow developed was intended as an “exploratory” tool, only two replicates of the strains were analyzed. Feature detection and extraction was performed using the Agilent Technologies’ MassHunter with the MFE algorithm.

Molecular features were detected and extracted in positive and negative ionization modes, and the feature lists were then merged to obtain a list of all chemical features detected across all samples. Features obtained from the positive and negative analysis were merged in a separate experiment followed by normalization of the data. However, as the intensity of the signals detected in negative ionization mode are generally lower, this means that features only detected in ESI$^-$ will have lower influence on the model. This problem was alleviated by normalization of the data before analysis. This is in contrast to other studies (Dai et al. 2014; Honoré et al. 2013), where feature extraction was performed in both positive and negative ionization modes, but without merging, requiring the work on two different features sets for further analysis.

This resulted in a table of chemical features detected from across all 13 strains, resulting in a feature table containing all detected compounds and their intensities for all strains. The whole dataset contained 7,190 extracted features from all strains, which is of course, too many features to investigate manually. Instead, the list of chemical features was investigated using a genetic algorithm (GA) combined with support vector machine (SVM). In the hybrid GA/SVM method applied in this study, GA works as a wrapper to select features to be evaluated in the SVM classifier, in that way reducing dimensionality and further improving the SVM performance (Li et al. 2014). The feature selection process is illustrated in simplified form in Figure 25.

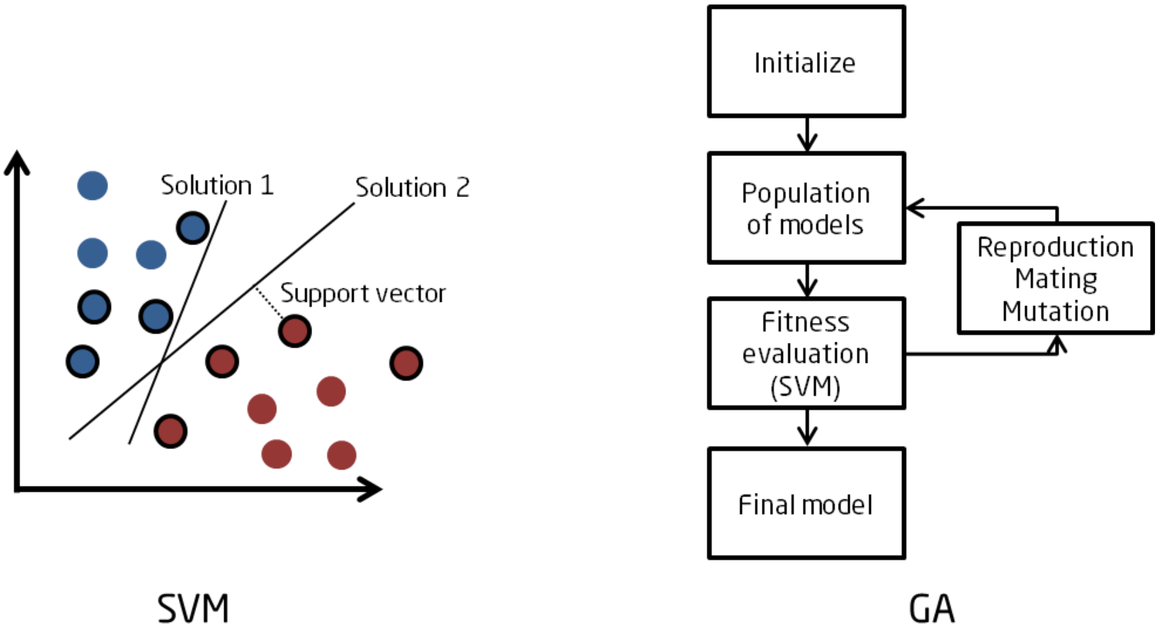


Figure 25 - GA/SVM procedure for feature selection. GA - genetic algorithm, SVM - Support vector machine.

A simplified explanation of the process is that case two different samples, red and blue, are analyzed using LC-MS, and the resulting features are extracted, illustrated by the colored circles. To find the differences between the groups, the most descriptive features need to be found. By using the GA, a number of features are evaluated, those with the black outline, to determine their descriptive power. This results in solution 1, which is successful in separating the two samples, that is, separate the red and blue circles. Solution 1 is then used to model a new solution by a process called cross-over, mimicking genetics, producing solution 2. This is done by using some of the same features, or circles, selected in solution 1, but randomly exchanging some of the features for others. Solution 2 is even better as the distance between the two samples, as determined via the support vectors, is larger. By repeating this procedure, the features that separate the different samples the best can then be determined. These features are thus the ones with the highest descriptive power.

The intrinsic nature of the GA makes it highly suitable for discovery purposes as it favors diversity in how the subset of features is selected , while SVM reduces the dimensionality of data in focusing on the minimal number of features that maximize the difference between the samples (Lin et al. 2011, 2012). There are only few examples on the use of support vector machine as classifier in untargeted SM profiling (Boccard et al. 2010; Mahadevan et al. 2008). In these cases, SVM was found to be superior to other multivariate analysis tools, because of its efficiency in reducing the dimensionality of data, resulting in its ability to reduce the dataset the most without leading to errors in prediction of groupings. A classifier based on this GA/SVM combination was used as a feature selection method in order to filter the most important features from the complex data set, starting with the 500 most intense ions and reducing it to the 50 most significant features to distinguish all 13 strains.

Features were dereplicated using molecular networking as well as database searches. Of the 50 descriptive features, only 15 could be tentatively assigned to known compound classes, including the four antibiotic classes identified in this species, underlining the utility of GA/SVM to prioritize not only strains but also compounds before the rate-limiting step of structural identification. Based on the list of descriptive chemical features, a matrix, or chemical barcode, could be created for each of the analyzed strains. For each of the 50 descriptive features, the strain would be assigned a 1, if it produced the compound, or a 0 if it did not. This could also be represented as a black and white line resembling a barcode.

For each strain, genomic DNA had also been extracted and sequenced. Biosynthetic pathways were predicted using the Antibiotics & secondary metabolite analysis shell (antiSMASH) (Medema et al. 2011) prediction tool, and these were then grouped into operational biosynthetic units (OBUs). This experiment was carried out using bacteria, which employ a different mechanism of PK biosynthesis. As discussed in section 1.3.1, fungi synthesize PKs using type 1 iterative PKSs, while bacteria use PKSs type 1 modular (non-iterative) PKSs, which allow for better computational prediction of products (Hertweck 2009). For each strain a genetic barcode was created, analogs to the chemical barcode, but in this case indicating whether the predicted pathway was present in the strain’s genome. Then by integrating data from the metabolomics and genomics experiments, it was possible to use data from one experiment to “interrogate” the other data. In that way information about a unique pathway in one strain could be used to search the chemical data for compounds unique to that strain.

Using this approach, we found that 30 % of all chemical features and 24 % of the biosynthetic genes were unique to a single strain, while only 2 % of the features and 7 % of the biosynthetic genes were shared between all. Features were dereplicated by MS/MS networking to identify molecular families of the same biosynthetic origin, and the associated pathways were probed by their pattern of conservation. Interestingly, most of the discriminating features were related to antibacterial compounds, including the thiomarinols that were reported from *P. luteoviolacea* here for the first time. Also, we could identify the biosynthetic cluster responsible for production of the antibiotic indolmycin based on the pattern of conservation, a cluster that could not be predicted by antiSMASH.

In conclusion, the workflow illustrates the strength of the untargeted approach, as the chemical potential of all strains could be investigated via comparison of detected chemical features. By comparing the distribution of these, it was possible to both reduce the list of chemical features dramatically, and to select the most descriptive features. The reduced dataset was then manually investigated and dereplicated leading to the tentative identification of several antibiotics, several of which had not previously been identified from the organism.

The combination of metabolomics and genomic data identifies obvious hotspots for chemical diversity among the 13 strains, which permit intelligent strain selection for more detailed chemical analyses. By randomly picking a single strain, worst case, only 38 % of the 500 most intense chemical features, and thus most relevant from a drug discovery perspective, are covered. However, if maximizing strain orthogonality by using the data generated to select the two strains with the highest number of unique genes, pathways, and chemical features, 82 % of the diversity can be covered, dramatically reducing the amount of data to analyze further.

Although the methodology developed here, and the results obtained from the analysis, were very encouraging, the study also served to highlight several complications regarding the experimental setup and analysis of data from this metabolomics based experiment. As a supposed “unbiased” form of analysis, there seem to be many sources of potential bias in metabolomics type studies. In section 11, it was described how the use of different feature extraction algorithms could significantly influence the results obtained from analyses (Lange et al. 2008). Taking a step back, the experimental procedures and generation of LC-MS data used for the analysis, will have a large impact on which compounds can be analyzed. While parameters such as extraction method and the stationary and mobile phases used in the LC clearly will have a huge influence – the impact of other settings might not be so clear. In **Paper 1**, a Bruker maXis QTOF system was used for screening of fungal extracts. This MS system contains a so-called ion-cooler which can be used to focus the ion beam. In this paper it was described how the ion-cooler settings influence the transfer efficiency of ions, favoring the transfer of ions in a specific *m/z*-range, while adversely affecting the transfer of ions in all other ranges. Thankfully, many researchers in the field has advocated for the standardization of reporting standards in the field, something that can help to identify these sources of bias (Sansone et al. 2007; Sumner et al. 2007).

As previously mentioned, this experiment was carried out using bacteria, which biosynthesize PKs using modular (non-iterative) PKSs, making it somewhat possible to predict the biosynthetic units and their products. To enable this in fungi, there is still a need to develop a better understanding of fungal biosynthesis to enable utilization of the tools that have been developed, as well as the new opportunities that developments in chemical analysis and metabolomics have provided. Although the field of metabolomics has evolved tremendously over the last decade, there are still many challenges regarding treatment and interpretation of obtained data. In spite of this, the results of this study show the importance and applicability of combining genomics and metabolomics, as well as the potential of its use.

3 Perspectives

3.1 Development of new methods and techniques

Improved instrumentation naturally leads to development of more advanced experimental techniques that can be used to gain even greater insights into the field of microbial SMs. However, as I have realized over the course of my study, advances in experimental procedures alone are not enough. Because of the important role of analysis of the acquired data, new methods for data analysis are just as important for the continued advancement of the field.

3.1.1 Metabolomics analysis based on LC-MS/MS data

In the field of metabolomics, databases of metabolites such as METLIN (Cho et al. 2014; Smith et al. 2005; Tautenhahn, Cho, et al. 2012) are continuously expanded upon, as more and more metabolites are being analyzed. This leads to a wealth of available MS/MS spectra that can be used for tentative identification of metabolites from other experiments. This information would be very valuable in standard metabolomics where full-scan instruments are used for untargeted analysis. Ideally, this means that the full-scan instrument operates in a way that allows for recording of both MS and MS/MS spectra of the metabolites. The MS/MS spectra could either be recorded at defined fragmentation voltages, or by modulating the fragmentation voltage based on other parameters such as the *m/z*-ratio of the precursor ion or the RT.

Acquisition of both MS and MS/MS metabolomics data would allow for directly matching against MS/MS libraries, distinguishing isomers and for determination of characteristic neutral loss fragments such as acylations, sulfanation, and prenylations. Recently, a method attempting to achieve this was published by Dai and coworkers (Dai et al. 2014) using a linear ion trap quadrupole (LTQ)-orbitrap system. The object of their study was to investigate the metabolite profile from human urine using an untargeted analysis, where the metabolites are expected to be heavily modified by acylation, sulfation, glucurinidation, and glucosidation. In their experiment they performed 18 different analytical runs varying the in-source collision-induced dissociation (ISCID) fragmentation voltage from 5 to 45 V in 5 V increments in both positive and negative ionization mode. Data from the different analytical runs were converted to peak tables and aligned. Using an in-house built data program, ions exhibiting the same RT and neutral ions were annotated as ion pairs of parent ions and fragment ions of modified metabolites, combined, and matched to produce a list containing specific metabolites with neutral losses. At present, the method developed by Dai *et al.* is an important first step and proof of concept for the general idea of performing metabolomics using MS/MS signals. However, there are a number of steps that need to be improved upon for more widespread adoption. One of the main disadvantages of the method is that it currently requires 18 analytical runs per sample, which is unfeasible in most cases. This could potentially be alleviated by development of new and improved instruments as well as other advances.

- One of the limiting factors in this procedure is the electronics of the mass spectrometer. With better digitizers the scan speed can be improved without a loss in mass accuracy and resolution. As famously predicted by Moore's law (Moore 1965), the rate of transistors in an integrated electronic circuit doubles approximately every second year, something that will greatly benefit the development of mass spectrometers.

- Improved electronics for the collision cell can make it possible to cycle through different fragmentation energies more rapidly, allowing for acquisition of data at multiple fragmentation energies during a single analytical run without losing sensitivity.

By reducing the number of different fragmentation energies used, the number of analytical runs needed would be reduced. This could also be achieved by employing a method where the fragmentation energy is varied as described in section 1.4.1. These methods work by modulating the fragmentation energy based on a parameter such as the *m/z* of the ions of interest. As mentioned earlier, the development of better methods for determination of fragmentation energies for the compounds and better algorithms for matching of MS/MS spectra to databases, the performance of these might be improved to a point where it will be feasible to use them for this type of analysis. Compared to normal full-scan MS analysis in metabolomics, this combined metabolomics MS/MS analysis approach has a number of advantages:

- As already mentioned, ordinary metabolomics analysis can be performed, and features of interest can be directly identified using MS/MS libraries.
- The obtained MS/MS data can be directly used in other forms of analysis e.g. molecular networking. In the case described in **Paper 7** this would have reduced the number of times the samples would have to be analyzed using the LC-MS system.
- By coupling data from the MS/MS experiments, information such as certain neutral losses, or information regarding the structural similarities of features could be directly coupled to the statistical analysis performed in the metabolomics part of the experiment. This means that up- or down-regulated features could quickly be examined to determine if they share structural similarities or modifications.

Further development on this type of analysis could be the use of MS^n, which would require instruments such as orbitraps, ions traps, or new hybrid instruments.

3.1.2 Development of improved prediction tools

To be able to better use the LC-MS data and to aid the interpretation of this better prediction tools need to be developed. For targeted analysis methods, such as the ones described in 2.1, the development of improved methods for prediction and modelling of compound RTs would allow for increased confidence in tentative identification of compounds (Miller et al. 2013; Moschet et al. 2013; Stanstrup et al. 2013). With increased use of MS/MS for identification of compounds, there is also a need for the development of more advanced methods for prediction of compound structures from MS/MS data and vice versa. This is especially important in the field of natural product discovery where standards of compounds of interest are likely not available. Several different methods and approaches have been developed, but such prediction remain far from trivial in many cases (Bandu et al. 2004; Bonn et al. 2010; Hufsky et al. 2012, 2014b; Ridder et al. 2012; Wang et al. 2014; Wolf et al. 2010).

3.1.3 Combining 'omics data

In the future, more and more projects will be based on systems biology approaches where the combination of heterogeneous data analysis will be imperative as the integration of genomics, transcriptomics, proteomics, and metabolomics will reveal information not attainable by analysis of any single type of data. A method for combining metabolomics and genomics data is described in **Paper 7**, demonstrating how metabolomics information could be coupled to genetic data revealing information about the biosynthetic genes responsible for the production of a specific metabolite. Other methods for combining metabolomics and transcriptomics data for the investigation of pathway enrichment have been published, demonstrating how correlation between the datasets can reveal new information (Eichner et al. 2014; Kaever et al. 2014). With the development of more advanced methods for analyzing combined data, the systems biology approach of holistic data analysis will become even more powerful, helping us to identify and explain changes and correlations in multi-'omics data.

3.2 Sharing of data

Sharing of data is more common in the other biological fields such as genomics, but we have seen a development towards more sharing of data in metabolomics as well. MetaboLights (Haug et al. 2013) allows for the sharing of data from metabolomics experiments. Libraries such as METLIN containing primarily human metabolites, as well as MassBank (Horai et al. 2010) are being made available online. In the GnPS project, described in section 1.7.1, all submitted data are, except under special circumstances, released publicly. This means that we are in the midst of a data revolution requiring the development of new analysis. The development in genetics will therefore probably be mirrored, possibly leading to the advent of new meta-metabolomics studies. But this also means that we have an opportunity to influence the best practices of these data repositories. This means that we should push for better documentation of data as well as standardized reporting formats (Griss et al. 2014). To be able to compare data obtained from different research groups all using their own methods and instruments, new methods for quality control also need to be established (X. Yang et al. 2014).

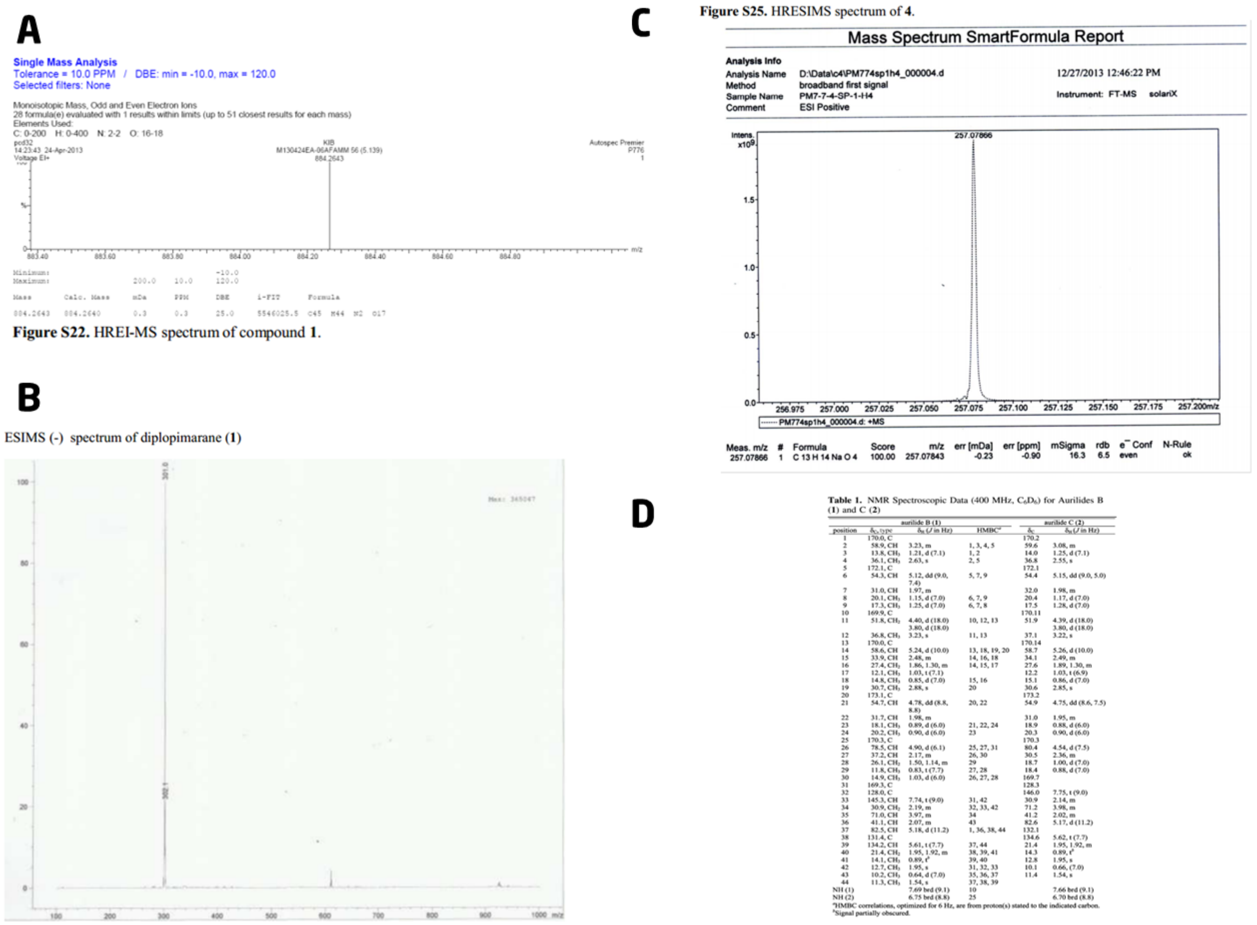
Biosynthetic pathways of compounds are often published as detailed figures with a wealth of annotations. However, these figures are hidden away as image files in different publications, making it hard to gain an overview of the available data and information. An example of this is the biosynthesis of emericellin in *A. nidulans*. The complete biosynthesis has been elucidated and published in steps by several different research groups, but is split out over multiple publications (Chiang et al. 2010; M. M. L. Nielsen et al. 2011). This is a common occurrence as the characterization and identification of biosynthetic routes is extremely work intensive, and is often a joint undertaking performed by several different research groups. However, with more and more pathways being described, it is becoming a Sisyphean task to keep track of published pathways and corrections in the current form. One solution could be to require that all biosynthetic be deposited in a publically accessible databank such as WikiPathways (Kelder et al. 2012; "WikiPathways" n.d.), which would facilitate faster data analysis as the pathways could be mined and used in data analysis software such as Agilent MPP (Agilent Techologies n.d.) or other integrated 'omics software workflows, and allow for easier dissemination of new information such as intermediates and enzyme reactions.

In the GnPS project large amounts of MS/MS data is being uploaded and released to the public. In its present state, not much metadata is provided for the sample, which of course is to encourage researchers to share their data, as it reduces the risk of other research groups "poaching" each other's results. However, this means that we are missing out on a wealth of data related to the samples. In the GNPS project we are able to find compounds that have similar MS/MS spectra. Unfortunately we do not have access to any of the instrumental parameters from the experiment. Imagine if we were to have access to this information or metadata. They would allow us to perform a large range of meta-experiments. By mining the meta-data parameters from the LC method such as RT of compounds, separation type (RP, HILIC, etc.), and mobile phase could be used to develop better methods for RT prediction. Likewise, MS parameters could be mined to better predict fragmentation spectra.

3.2.1 Reporting standards

The topic of required reporting standards is an often discussed topic in the field of chemistry and especially in the more specialized fields such as analytical chemistry, metabolomics, and natural products. Analytical chemistry is highly codified, stemming from its use in highly regulated industries such as pharmaceuticals and food and feed production. Because of this, reporting of experimental parameters such as limit of detection, limit of quantification, integration parameters, signal-to-noise, and detailed instrument settings are a prerequisite for publication of research.

In natural product chemistry the requirements for reportings on new compounds has naturally evolved over time with the development of new techniques and instruments. For newly described compounds, the accurate mass, UV-spectrum, optical rotation, and ^{1}H-NMR and ^{13}C-NMR spectra are often reported. The form in which these data are reported can, however, vary quite dramatically between publications. Examples of mass spectra reported from the Journal of Natural Products (Figure 26A-C), can be depicted in a way that does not allow the reader to identify the pseudomolecular ion, investigate the isotopic pattern of the compound, or observe any possible adducts.



Single Mass Analysis
Tolerance = 10.0 PPM / DBE: min = -10.0, max = 120.0
Selected filters: None

Monoisotopic Mass, Odd and Even Electron Ions
28 formula(e) evaluated with 1 results within limits (up to 51 closest results for each mass)
Elements Used:
C: 0-200 H: 0-400 N: 2-2 O: 16-18
pcd32
14:23:43 24-Apr-2013
Voltage EI+
KIB
M130424EA-06AFAMM 56 (5.139)
884.2643
Autospec Premier
P776
1

```
Minimum:                              -10.0
Maximum:              200.0    10.0   120.0

Mass       Calc. Mass   mDa    PPM    DBE     i-FIT       Formula

884.2643   884.2640     0.3    0.3    25.0    5546025.5   C45  H44  N2  O17
```

Figure S22. HREI-MS spectrum of compound **1**.

ESIMS (-) spectrum of diplopimarane (**1**)

Figure S25. HRESIMS spectrum of **4**.

Mass Spectrum SmartFormula Report

Analysis Info
Analysis Name D:\Data\c4\PM774sp1h4_000004.d
Method broadband first signal
Sample Name PM7-7-4-SP-1-H4
Comment ESI Positive
12/27/2013 12:46:22 PM
Instrument: FT-MS solariX

```
Meas. m/z   #  Formula           Score    m/z        err [mDa]  err [ppm]  mSigma  rdb  e⁻ Conf  N-Rule
257.07866   1  C 13 H 14 Na O 4  100.00   257.07843  -0.23      -0.90      16.3    6.5  even     ok
```

Table 1. NMR Spectroscopic Data (400 MHz, C_6D_6) for Aurilides B (**1**) and C (**2**)

	aurilide B (**1**)			aurilide C (**2**)	
position	δ_C, type	δ_H (J in Hz)	HMBC[a]	δ_C	δ_H (J in Hz)
1	170.0, C			170.2	
2	58.9, CH	3.23, m	1, 3, 4, 5	59.6	3.08, m
3	13.8, CH_3	1.21, d (7.1)	1, 2	14.0	1.25, d (7.1)
4	36.1, CH_3	2.63, s	2, 5	36.8	2.55, s
5	172.1, C			172.1	
6	54.3, CH	5.12, dd (9.0, 7.4)	5, 7, 9	54.4	5.15, dd (9.0, 5.0)
7	31.0, CH	1.97, m		32.0	1.98, m
8	20.1, CH_3	1.15, d (7.0)	6, 7, 9	20.4	1.17, d (7.0)
9	17.3, CH_3	1.25, d (7.0)	6, 7, 8	17.5	1.28, d (7.0)
10	169.9, C			170.11	
11	51.8, CH_2	4.40, d (18.0) 3.80, d (18.0)	10, 12, 13	51.9	4.39, d (18.0) 3.80, d (18.0)
12	36.8, CH_3	3.23, s	11, 13	37.1	3.22, s
13	170.0, C			170.14	
14	58.6, CH	5.24, d (10.0)	13, 18, 19, 20	58.7	5.26, d (10.0)
15	33.9, CH	2.48, m	14, 16, 18	34.1	2.49, m
16	27.4, CH_2	1.86, 1.30, m	14, 15, 17	27.6	1.89, 1.30, m
17	12.1, CH_3	1.03, t (7.1)		12.2	1.03, t (6.9)
18	14.8, CH_3	0.85, d (7.0)	15, 16	15.1	0.86, d (7.0)
19	30.7, CH_3	2.88, s	20	30.6	2.85, s
20	173.1, C			173.2	
21	54.7, CH	4.78, dd (8.8, 8.8)	20, 22	54.9	4.75, dd (8.6, 7.5)
22	31.7, CH	1.98, m		31.0	1.95, m
23	18.1, CH_3	0.89, d (6.0)	21, 22, 24	18.9	0.88, d (6.0)
24	20.2, CH_3	0.90, d (6.0)	23	20.3	0.90, d (6.0)
25	170.3, C			170.3	
26	78.5, CH	4.90, d (6.1)	25, 27, 31	80.4	4.54, d (7.5)
27	37.2, CH	2.17, m	26, 30	30.5	2.36, m
28	26.1, CH_2	1.50, 1.14, m	29	18.7	1.00, d (7.0)
29	11.8, CH_3	0.83, t (7.7)	27, 28	18.4	0.88, d (7.0)
30	14.9, CH_3	1.03, d (6.0)	26, 27, 28	169.7	
31	169.3, C			128.3	
32	128.0, C			146.0	7.75, t (9.0)
33	145.3, CH	7.74, t (9.0)	31, 42	30.9	2.14, m
34	30.9, CH_2	2.19, m	32, 33, 42	71.2	3.98, m
35	71.0, CH	3.97, m	34	41.2	2.02, m
36	41.1, CH	2.07, m	43	82.6	5.17, d (11.2)
37	82.5, CH	5.18, d (11.2)	1, 36, 38, 44	132.1	
38	131.4, C			134.6	5.62, t (7.7)
39	134.2, CH	5.61, t (7.7)	37, 44	21.4	1.95, 1.92, m
40	21.4, CH_2	1.95, 1.92, m	38, 39, 41	14.3	0.89, t[b]
41	14.1, CH_3	0.89, t[b]	39, 40	12.8	1.95, s
42	12.7, CH_3	1.95, s	31, 32, 33	10.1	0.66, (7.0)
43	10.2, CH_3	0.64, d (7.0)	35, 36, 37	11.4	1.54, s
44	11.3, CH_3	1.54, s	37, 38, 39		
NH (1)		7.69 brd (9.1)	10		7.66 brd (9.1)
NH (2)		6.75 brd (8.8)	25		6.70 brd (8.8)

[a] HMBC correlations, optimized for 6 Hz, are from proton(s) stated to the indicated carbon.
[b] Signal partially obscured.

Figure 26 - A-C) Reportings of the pseudomolecular ion a given compound. The spectra depicted in A(Gu et al. 2014) **and C**(Liou et al. 2014) **only depict a mass range of 1 *m/z*, precluding observation of isotope pattern or possible adducts, while the quality of the spectrum in B**(Andolfi et al. 2014) **is of insufficient quality to conclusively identify the pseudomolecular ion. D) Table of NMR spectroscopic data from the Journal of Natural Products Author Guidelines**("Author guidelines for submission to the Journal of Natural Products" n.d.)**. In the guidelines it is captioned "The correct presentation of NMR spectroscopic data is shown in the table below".**

In natural products chemistry, new compounds are routinely analyzed using MS to obtain the molecular formula of the compound, often only reporting the measured *m/z* value and the calculated mass error. Errors in assignment of the ions of interest caused by water loss or adduct formation are thus hard to address, both in review of the article and after is has been published, a problem further complicated by the fact that in natural product chemistry it is still not standard to publish MS/MS spectra of newly described compounds. With more advanced screening techniques such as the ones described in this thesis (**Papers 1 and 2**), this type of information is essential for initial dereplication efforts and for expanding the databases. More advanced experiments such as MS^n-type analysis are even more complex, but are still routinely reported in the form of a table of fragment ions. Instead, if the data was made available in standardized formats, it would be possible for other research groups to analyze the data and add it to a public database.

In the Journal of Natural Products' Author Guidelines("Author guidelines for submission to the Journal of Natural Products" n.d.), an example of the reporting of NMR spectroscopic data is shown (Figure 26D). This tabulated form is clear and concise for the reader, but by not including the NMR-spectra themselves it is not possible to actually review the data and analyze the data for oneself. Even in cases where spectra are published as well, it can be very difficult to actually interpret the data from a low-quality image, as it is not unusual to see scanned versions of printed spectra complete with hand drawn annotations. NMR spectra obtained from analysis of complex compounds can be very hard to assign without the use of specialized software, thereby necessitating access to the original data file obtained from the experiments. Access to the data file would again also allow for use of the data for training of structure elucidation purposes.

3.2.2 Open-Access - Opening of databases

Open-access journals have been defined as being available online “without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself” (Suber 2012). In practice this means that the individual scientific articles can be downloaded free-of-charge from the internet either at the same time as the article itself is published or after an embargo period. In the case where an embargo is imposed, so called green open access, the publishers are able to recoup any costs associated with the article by charging for access during the embargo period or for subscriptions to the journal. Scientific articles that are made open-access immediately are referred to as gold open-access, and often charge the authors a fee to cover expense related to publishing. In turn this means that the cost associated with accessing a scientific article is shifted from the reader to the writer. However, several scientific funding programs now provide earmarked funds for this purpose, exemplified by the EU Seventh Framework Programme (“Open Access in FP7” 2014).

However, there are still major problems with the availability of raw experimental data as well as databases. Databases of SMs such as Antibase (Laatsch 2012) and MarinLit (“MarinLit” n.d.), are commercially available but are very expensive to acquire. The Dictionary of Natural Products(Press n.d.) hosts a free version containing a subset of the compounds available in the paid version and only allows for single compound look-up, requiring the paid version for batch look-up . Most of them require a subscription or release updated versions annually making them a recurring cost. The databases are the *de facto* standards in the field of natural product chemistry, and are therefore essential for any researcher working in the field. Even though the vast majority of the compounds in Antibase have been culled from published literature, the database does contain several unpublished compounds, such as the compound methyl pyrrole-2-carboxylate isolated from a marine *Actinomyces*. Because of this there is a great incentive to keep using these databases, thus perpetuating the cycle and increasing the power of the publishers. There is no readily apparent solution to the issue of these closed databases. Closed databases also make it hard for the community to share information about errors in the database. As discussed in section 2.2.4, Antibase contains an erroneous entry for the compound fungisporin among others. However, because of the closed nature of the databases, there is no way to disseminate information about errors. These errors can be reported by to the creators of the database, but would probably not be corrected until a new paid version is released.

Unlike scientific articles, it may be advantageous for a group of researchers not to publish their in-house databases, as it can give them an advantage over their competitors, for instance when performing dereplication of samples. One way to encourage researchers to publish these databases could be to establish funding specifically for database creation, or to require funded projects to publish data from experiments in specific open-source formats.

4 Conclusion

The focus of this thesis has been on investigation of SMs from microorganisms through development of new methods for analysis of LC-MS data as well as new experimental approaches for investigation of the biosynthesis of these metabolites. The methods developed, and the results obtained through their use are described in chapter 3, divided into the subjects: targeted analysis, biosynthesis studies using stable isotope labeling, and untargeted analysis.

For targeted analysis, two methods were described, both based on screening of extracts from microorganisms using prepared libraries of known metabolites using LC-MS and LC-MS/MS data, respectively. Approaches for the study of biosynthesis of fungal metabolites using SIL compounds were described. Lastly, a metabolomics approach was developed to assess the biosynthetic potential of a collection of marine bacteria. Several of the developed data analysis methods and experimental approaches were applied in combination, leveraging the developed screening methods to speed up data analysis.

Isotopic labeling for investigation of biosyntheses proved very effective as a means to investigate compounds of both known and unknown origin using LC-MS. Investigation of the PK yanuthone D lead to characterization of its biosynthesis, including the biosynthetic genes responsible for its production, and identification of several new analogs. The experimental approach developed was further generalized and used to successfully investigate PK biosynthesis in a range of different fungal genera.

An approach combining SILAAs and molecular networking for the detection and structure elucidation of NRPs was developed and demonstrated using extracts from filamentous fungi. Results from the study resulted in the identification of several new NRPs, for which the biosynthesis could be linked to a single NRPS. This NRPS had previously been shown to produce other NRPs, demonstrating the usefulness of the combined approach in both detecting and identifying compounds.

Finally, a metabolomics based approach was developed to characterize the biosynthetic potential of marine bacteria. The developed methodologies could be used to select organism for further studies, by prioritizing strains based on their expressed metabolites, but also by coupling these metabolites to their biosynthetic genes.

Based on these results, the data analysis methods and methodologies developed during these studies have proven very effective and applicable to a wide range of microorganisms, not only restricted to fungi. The developed methods have revealed new insights into microbial SMs, and it is clear that further discoveries still wait.

5 References

Adrio, J. L., & Demain, A. L. (2003). Fungal biotechnology. *International microbiology : the official journal of the Spanish Society for Microbiology*, *6*(3), 191–9. doi:10.1007/s10123-003-0133-0

Agilent Techologies. (n.d.). Agilent Mass Profiler Professional. http://www.chem.agilent.com/en-US/products-services/Software-Informatics/Mass-Profiler-Professional-Software/Pages/default.aspx. Accessed 8 December 2014

Ali, H., Ries, M. I., Lankhorst, P. P., van der Hoeven, R. a M., Schouten, O. L., Noga, M., et al. (2014). A non-canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in *Penicillium chrysogenum*. *PloS one*, *9*(6), e98212. doi:10.1371/journal.pone.0098212

America, A. H. P., & Cordewener, J. H. G. (2008). Comparative LC-MS: a landscape of peaks and valleys. *Proteomics*, *8*(4), 731–49. doi:10.1002/pmic.200700694

Andolfi, A., Maddau, L., Basso, S., Linaldeddu, B. T., Cimmino, A., Scanu, B., et al. (2014). Diplopimarane, a 20-nor-ent-pimarane produced by the oak pathogen *Diplodia quercivora*. *Journal of natural products*, *77*(11), 2352–60. doi:10.1021/np500258r

Andrews, G. L., Simons, B. L., Young, J. B., Hawkridge, A. M., & Muddiman, D. C. (2011). Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). *Analytical chemistry*, *83*(13), 5442–6. doi:10.1021/ac200812d

Author guidelines for submission to the Journal of Natural Products. (n.d.). http://pubs.acs.org/paragonplus/submission/jnprdf/jnprdf_authguide.pdf. Accessed 10 September 2014

Bandu, M. L., Watkins, K. R., Bretthauer, M. L., Moore, C. A., & Desaire, H. (2004). Prediction of MS/MS data. 1. A focus on pharmaceuticals containing carboxylic acids. *Analytical chemistry*, *76*(6), 1746–53. doi:10.1021/ac0353785

Bennett, J. W., & Klich, M. (2003). Mycotoxins. *Clinical Microbiology Reviews*, *16*(3), 497–516. doi:10.1128/CMR.16.3.497-516.2003

Benson, D. a, Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2011). GenBank. *Nucleic acids research*, *39*(Database issue), D32–7. doi:10.1093/nar/gkq1079

Bezuidenhout, S. (1988). Structure elucidation of the fumonisins, mycotoxins from *Fusarium moniliforme*. *Journal of the Chemical Society, Chemical Communications*, *1730*, 743–745. http://pubs.rsc.org/en/content/articlehtml/1988/c3/c39880000743. Accessed 25 November 2014

Bijlsma, L., Sancho, J. V, Hernández, F., & Niessen, W. M. A. (2011). Fragmentation pathways of drugs of abuse and their metabolites based on QTOF MS/MS and MS(E) accurate-mass spectra. *Journal of mass spectrometry : JMS*, *46*(9), 865–75. doi:10.1002/jms.1963

Boccard, J., Kalousis, A., Hilario, M., Lantéri, P., Hanafi, M., Mazerolles, G., et al. (2010). Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana*. *Chemometrics and Intelligent Laboratory Systems*, *104*(1), 20–27. doi:10.1016/j.chemolab.2010.03.003

Bolser, D. M., Chibon, P.-Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V. D., et al. (2012). MetaBase--the wiki-database of biological databases. *Nucleic acids research*, *40*(Database issue), D1250–4. doi:10.1093/nar/gkr1099

Bonn, B., Leandersson, C., Fontaine, F., & Zamora, I. (2010). Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation. *Rapid communications in mass spectrometry : RCM*, *24*(21), 3127–38. doi:10.1002/rcm.4753

Borel, J., Feurer, C., Gubler, H., & Stähelin, H. (1994). Biological effects of cyclosporin A: a new antilymphocytic agent. *Agents and actions*, *43*, 468–475. http://link.springer.com/article/10.1007/BF01986686. Accessed 12 November 2014

Bouslimani, A., Sanchez, L. M., Garg, N., & Dorrestein, P. C. (2014). Mass spectrometry of natural products: current, emerging and future technologies. *Natural product reports*, *31*(6), 718–29. doi:10.1039/c4np00044g

Brown, S. C., Kruppa, G., & Dasseux, J.-L. (2005). Metabolomics applications of FT-ICR mass spectrometry. *Mass spectrometry reviews*, *24*(2), 223–31. doi:10.1002/mas.20011

Bugni, T. S., Abbanat, D., Bernan, V. S., Maiese, W. M., Greenstein, M., Van Wagoner, R. M., & Ireland, C. M. (2000). Yanuthones: Novel metabolites from a marine isolate of *Aspergillus niger*. *The Journal of Organic Chemistry*, *65*(21), 7195–7200. doi:10.1021/jo0006831

Bycroft, B. W. (1988). *Dictionary of antibiotics and related substances* (1st ed., p. 962). Cambridge: Chapman and Hall/CRC.

Callahan, D. L., & Elliott, C. E. (2013). Metabolomics tools for natural product discovery, *1055*, 57–70. doi:10.1007/978-1-62703-577-4

Challis, G. L., Ravel, J., & Townsend, C. A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & Biology*, *7*(3), 211–224. doi:10.1016/S1074-5521(00)00091-0

Chanda, A., Roze, L. V, Kang, S., Artymovich, K. a, Hicks, G. R., Raikhel, N. V, et al. (2009). A key role for vesicles in fungal secondary metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(46), 19533–8. doi:10.1073/pnas.0907416106

Chen, Y.-B., Chattopadhyay, A., Bergen, P., Gadd, C., & Tannery, N. (2007). The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System--a one-stop gateway to online bioinformatics databases and software tools. *Nucleic acids research*, *35*(Database issue), D780–5. doi:10.1093/nar/gkl781

Chiang, Y.-M., Szewczyk, E., Davidson, A. D., Entwistle, R., Keller, N. P., Wang, C. C. C., & Oakley, B. R. (2010). Characterization of the *Aspergillus nidulans* monodictyphenone gene cluster. *Applied and environmental microbiology*, *76*(7), 2067–2074. doi:10.1128/AEM.02187-09

Cho, K., Mahieu, N. G., Ivanisevic, J., Uritboonthai, W., Chen, Y.-J., Siuzdak, G., & Patti, G. J. (2014). isoMETLIN: A Database for Isotope-Based Metabolomics. *Analytical chemistry*. doi:10.1021/ac5029177

Christensen, C., Nelson, G., & Mirocha, C. (1965). Effect on the white rat uterus of a toxic substance isolated from *Fusarium*. *Applied microbiology*, *13*(5), 653–659. http://aem.asm.org/content/13/5/653.short. Accessed 10 December 2014

Collier, T. S., Hawkridge, A. M., Georgianna, D. R., Payne, G. a, & Muddiman, D. C. (2008). Top-down identification and quantification of stable isotope labeled proteins from *Aspergillus flavus* using online nano-flow reversed-phase liquid chromatography coupled to a LTQ-FTICR mass spectrometer. *Analytical chemistry*, *80*(13), 4994–5001. doi:10.1021/ac800254z

Collins, B. C., Gillet, L. C., Rosenberger, G., Röst, H. L., Vichalkovski, A., Gstaiger, M., & Aebersold, R. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nature methods*, *10*(12), 1246–53. doi:10.1038/nmeth.2703

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et biophysica acta*, *1830*(6), 3670–95. doi:10.1016/j.bbagen.2013.02.008

Dai, W., Yin, P., Zeng, Z., Kong, H., Tong, H., Xu, Z., et al. (2014). Nontargeted modification-specific metabolomics study based on liquid chromatography-high-resolution mass spectrometry. *Analytical chemistry*, *86*(18), 9146–53. doi:10.1021/ac502045j

De Hoffmann, E., & Stroobant, V. (2007). Tandem mass spectrometry. In *Mass Spectrometry* (3rd ed., pp. 189–216). Hoboken: John Wiley & Sons.

Demain, A. L., & Fang, A. (2000). The natural functions of secondary metabolites. *Advances in biochemical engineering/biotechnology*, *69*, 1–39. http://www.ncbi.nlm.nih.gov/pubmed/11036689. Accessed 3 August 2011

Ding, X., Ghobarah, H., Zhang, X., Jaochico, A., Liu, X., Deshmukh, G., et al. (2013). High-throughput liquid chromatography/mass spectrometry method for the quantitation of small molecules using accurate mass technologies in supporting discovery drug screening. *Rapid Communications in Mass Spectrometry*, *27*(3), 401–408. doi:10.1002/rcm.6461

Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical biology*, *5*(1), 011001. doi:10.1088/1478-3975/5/1/011001

Eichner, J., Rosenbaum, L., Wrzodek, C., Häring, H.-U., Zell, A., & Lehmann, R. (2014). Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, *966*, 77–82. doi:10.1016/j.jchromb.2014.04.030

El-Elimat, T., Figueroa, M., Ehrmann, B. M., Cech, N. B., Pearce, C. J., & Oberlies, N. H. (2013). High-resolution MS, MS/MS, and UV database of fungal secondary metabolites as a dereplication protocol for bioactive natural products. *Journal of natural products*, *76*(9), 1709–16. doi:10.1021/np4004307

Eliasson, M., Rännar, S., Madsen, R., Donten, M. A., Marsden-Edwards, E., Moritz, T., et al. (2012). Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach. *Analytical chemistry*, *84*(15), 6869–76. doi:10.1021/ac301482k

Endo, A. (1985). Compactin (ML-236B) and related compounds as potential cholesterol-lowering agents that inhibit HMG-CoA reductase. *Journal of medicinal chemistry*, *28*(4), 401–405. http://pubs.acs.org/doi/abs/10.1021/jm00382a001. Accessed 12 November 2014

Eugster, P., Guillarme, D., Rudaz, S., Veuthey, J.-L., Carrupt, P.-A., & Wolfender, J.-L. (2011). Ultra high pressure liquid chromatography for crude plant extract profiling. *Journal of AOAC …*, *94*(1), 51–70. http://www.ingentaconnect.com/content/aoac/jaoac/2011/00000094/00000001/art00008. Accessed 31 October 2014

Fiehn, O. (2002). Metabolomics - The link between genotypes and phenotypes. *Plant Molecular Biology*, *48*(1-2), 155–171. ISI:000173211000011

Finking, R., & Marahiel, M. a. (2004). Biosynthesis of nonribosomal peptides. *Annual review of microbiology*, *58*, 453–88. doi:10.1146/annurev.micro.58.030603.123615

Forner, D., Berrué, F., Correa, H., Duncan, K., & Kerr, R. G. (2013). Chemical dereplication of marine actinomycetes by liquid chromatography–high resolution mass spectrometry profiling and statistical analysis. *Analytica Chimica Acta*, *805*, 70–79. doi:10.1016/j.aca.2013.10.029

Frisvad, J. C., Rank, C., Nielsen, K. F., & Larsen, T. O. (2009). Metabolomics of *Aspergillus fumigatus*. *Medical mycology : official publication of the International Society for Human and Animal Mycology*, *47 Suppl 1*, S53–71. doi:10.1080/13693780802307720

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., et al. (2010). SMPDB: The Small Molecule Pathway Database. *Nucleic acids research*, *38*(Database issue), D480–7. doi:10.1093/nar/gkp1002

Georgianna, D. R., Hawkridge, A. M., Muddiman, D. C., & Payne, G. A. (2008). Temperature-dependent regulation of proteins in *Aspergillus flavus*: whole organism stable isotope labeling by amino acids. *Journal of proteome research*, *7*(7), 2973–9. doi:10.1021/pr8001047

Geris, R., & Simpson, T. J. (2009). Meroterpenoids produced by fungi. *Natural product reports*, *26*(8), 1063–94. doi:10.1039/b820413f

Glauser, G., Veyrat, N., Rochat, B., Wolfender, J.-L., & Turlings, T. C. J. (2012). Ultra-high pressure liquid chromatography-mass spectrometry for plant metabolomics: A systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. *Journal of chromatography. A*, *1292*, 151–159. doi:10.1016/j.chroma.2012.12.009

GnPS: Global Natural Products Social Molecular Networking. (n.d.). http://gnps.ucsd.edu. Accessed 10 July 2014

Goliński, P., Wnuk, S., Chełkowski, J., Visconti, A., & Schollenberger, M. (1986). Antibiotic Y: biosynthesis by *Fusarium avenaceum* (Corda ex Fries) Sacc., isolation, and some physicochemical and biological properties. *Applied and environmental microbiology*, *51*(4), 743–5.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=238958&tool=pmcentrez&rendertype=abstract

Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., et al. (2014). Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Analytical chemistry*, *86*(14), 6931–9. doi:10.1021/ac500734c

Griffith, G. (2004). The use of stable isotopes in fungal ecology. *Mycologist*, *18*(November), 177–183. doi:10.1017/S0269915XO4004082

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., et al. (2014). The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & cellular proteomics : MCP*, *13*(10), 2765–75. doi:10.1074/mcp.O113.036681

Grunwald, H., Hargreaves, P., Gebhardt, K., Klauer, D., Serafyn, A., Schmitt-Hoffmann, A., et al. (2013). Experiments for a systematic comparison between stable-isotope-(deuterium) labeling and radio-((14)C) labeling for the elucidation of the in vitro metabolic pattern of pharmaceutical drugs. *Journal of pharmaceutical and biomedical analysis*, *85*, 138–44. doi:10.1016/j.jpba.2013.07.004

Gu, W., Zhang, Y., Hao, X.-J., Yang, F.-M., Sun, Q.-Y., Morris-Natschke, S. L., et al. (2014). Indole alkaloid glycosides from the aerial parts of *Strobilanthes cusia*. *Journal of natural products*, 1–5. doi:10.1021/np5003274

Guo, C.-J., Sun, W.-W., Bruno, K. S., & Wang, C. C. C. (2014). Molecular genetic characterization of terreic acid pathway in *Aspergillus terreus*. *Organic letters*, *16*(20), 5250–3. doi:10.1021/ol502242a

Halabalaki, M., Vougogiannopoulou, K., Mikros, E., & Skaltsounis, A. L. L. (2014). Recent advances and new strategies in the NMR-based identification of natural products. *Current opinion in biotechnology*, *25*, 1–7. doi:10.1016/j.copbio.2013.08.005

Hanahan, D., & Al-Wakil, S. (1952). The biosynthesis of ergosterol from isotopic acetate. *Archives of biochemistry and biophysics*, *37*(1), 167–171. http://www.sciencedirect.com/science/article/pii/0003986152901768. Accessed 13 October 2014

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2013). MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, *41*(Database issue), D781–6. doi:10.1093/nar/gks1004

Helmstaedt, K., Braus, G. H., Braus-Stromeyer, S., Busch, S., Hofmann, K., Goldman, G. H., & Draht, O. W. (2007). Amino acid supply of *Aspergillus*. In G. H. Goldman & S. A. Osmani (Eds.), *The Aspergilli Genomics, Medical Aspects, Biotechnology, and Research Methods* (1st ed., pp. 143–175). Boca Raton: CRC Press. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Amino+Acid+Supply+of+Aspergillus#5. Accessed 14 August 2014

Hendriks, M. M. W. B., Eeuwijk, F. A. va., Jellema, R. H., Westerhuis, J. a., Reijmers, T. H., Hoefsloot, H. C. J., & Smilde, A. K. (2011). Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry*, *30*(10), 1685–1698. doi:10.1016/j.trac.2011.04.019

Hertweck, C. (2009). The biosynthetic logic of polyketide diversity. *Angewandte Chemie-International Edition*, *48*(26), 4688–4716. ISI:000267494500004

Honoré, A. H., Thorsen, M., & Skov, T. (2013). Liquid chromatography-mass spectrometry for metabolic footprinting of co-cultures of lactic and propionic acid bacteria. *Analytical and bioanalytical chemistry*, *405*(25), 8151–70. doi:10.1007/s00216-013-7269-3

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, *45*(7), 703–14. doi:10.1002/jms.1777

Hou, Y., Braun, D. R., Michel, C. R., Klassen, J. L., Adnani, N., Wyche, T. P., & Bugni, T. S. (2012). Microbial strain prioritization using metabolomics tools for the discovery of natural products. *Analytical chemistry*, *84*(10), 4277–83. doi:10.1021/ac202623g

Huang, X., Chen, Y.-J., Cho, K., Nikolskiy, I., Crawford, P. a, & Patti, G. J. (2014). X13CMS: global tracking of isotopic labels in untargeted metabolomics. *Analytical chemistry*, *86*(3), 1632–9. doi:10.1021/ac403384n

Hufsky, F., Rempt, M., Rasche, F., Pohnert, G., & Böcker, S. (2012). De novo analysis of electron impact mass spectra using fragmentation trees. *Analytica chimica acta*, *739*, 67–76. doi:10.1016/j.aca.2012.06.021

Hufsky, F., Scheubert, K., & Böcker, S. (2014a). Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends in Analytical Chemistry*, *53*, 41–48. doi:10.1016/j.trac.2013.09.008

Hufsky, F., Scheubert, K., & Böcker, S. (2014b). New kids on the block: novel informatics methods for natural product discovery. *Natural product reports*, 807–817. doi:10.1039/c3np70101h

Hunter, P. (2009). Reading the metabolic fine print. The application of metabolomics to diagnostics, drug research and nutrition might be integral to improved health and personalized medicine. *EMBO reports*, *10*(1), 20–3. doi:10.1038/embor.2008.236

Ito, T., & Masubuchi, M. (2014). Dereplication of microbial extracts and related analytical technologies. *The Journal of antibiotics*, *67*(5), 353–60. doi:10.1038/ja.2014.12

Jones, K. A., Kim, P. D., Patel, B. B., Kelsen, S. G., Braverman, A., Swinton, D. J., et al. (2013). Immunodepletion plasma proteomics by tripleTOF 5600 and Orbitrap elite/LTQ-Orbitrap Velos/Q exactive mass spectrometers. *Journal of proteome research*, *12*(10), 4351–65. doi:10.1021/pr400307u

Junot, C., Fenaille, F., Colsch, B., & Bécher, F. (2014). High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. *Mass spectrometry reviews*, *33*(6), 471–500. doi:10.1002/mas.21401

Kaever, A., Landesfeind, M., Feussner, K., Morgenstern, B., Feussner, I., & Meinicke, P. (2014). Meta-analysis of pathway enrichment: combining independent and dependent omics data sets. *PloS one*, *9*(2), e89297. doi:10.1371/journal.pone.0089297

Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., & Hill, H. H. (2008). Ion mobility-mass spectrometry. *Journal of mass spectrometry : JMS*, *43*(1), 1–22. doi:10.1002/jms.1383

Katajamaa, M., Miettinen, J., & Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, *22*(5), 634–6. doi:10.1093/bioinformatics/btk039

Katajamaa, M., & Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of chromatography. A*, *1158*(1-2), 318–28. doi:10.1016/j.chroma.2007.04.021

Kaufmann, A. (2011). The current role of high-resolution mass spectrometry in food analysis. *Analytical and bioanalytical chemistry*. doi:10.1007/s00216-011-5629-4

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2012). WikiPathways: building research communities on biological pathways. *Nucleic acids research*, *40*(Database issue), D1301–7. doi:10.1093/nar/gkr1074

Keller, N. P., Turner, G., & Bennett, J. W. (2005). Fungal secondary metabolism - from biochemistry to genomics. *Nature reviews. Microbiology*, *3*(12), 937–947. doi:10.1038/nrmicro1286

Kempken, F., & Rohlfs, M. (2010). Fungal secondary metabolite biosynthesis - a chemical defence strategy against antagonistic animals? *Fungal Ecology*, *3*(3), 107–114. ISI:000279414700001

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, *7*, 234. doi:10.1186/1471-2105-7-234

Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, *8*, 105. doi:10.1186/1471-2105-8-105

Kluger, B., Bueschl, C., Neumann, N. K. N., Stueckler, R., Doppler, M., Chassy, A. W., et al. (2014). Untargeted profiling of tracer derived metabolites using stable isotopic labeling and fast polarity switching LC-ESI-HRMS. *Analytical chemistry*. doi:10.1021/ac503290j

Konishi, Y., Kiyota, T., Draghici, C., Gao, J.-M., Yeboah, F., Acoca, S., et al. (2007). Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Analytical chemistry*, *79*(3), 1187–97. doi:10.1021/ac061391o

Krauss, M., Singer, H., & Hollender, J. (2010). LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Analytical and bioanalytical chemistry*, *397*(3), 943–51. doi:10.1007/s00216-010-3608-9

Laatsch, H. (2012). Antibase 2012: The natural compound identifier. In *Antibase 2012: The natural compound identifier*.

Lang, G., Mitova, M. I., Ellis, G., van der Sar, S., Phipps, R. K., Blunt, J. W., et al. (2006). Bioactivity profiling using HPLC/microtiter-plate analysis: application to a New Zealand marine alga-derived fungus, Gliocladium sp. *Journal of natural products*, *69*(4), 621–4. doi:10.1021/np0504917

Lange, E., Tautenhahn, R., Neumann, S., & Gröpl, C. (2008). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC bioinformatics*, *9*, 375. doi:10.1186/1471-2105-9-375

Larsen, T. O., & Hansen, M. A. E. (2007). Dereplication and discovery of natural products by UV spectrosopy. In R. Molyneux & S. Colegate (Eds.), *Bioactive Natural Products* (2nd ed., pp. 221–244). CRC Press. doi:10.1201/9781420006889

Lee, D.-K., Yoon, M. H., Kang, Y. P., Yu, J., Park, J. H., Lee, J., & Kwon, S. W. (2013). Comparison of primary and secondary metabolites for suitability to discriminate the origins of *Schisandra chinensis* by GC/MS and LC/MS. *Food chemistry*, *141*(4), 3931–7. doi:10.1016/j.foodchem.2013.06.064

Lehner, S. M., Neumann, N. K. N., Sulyok, M., Lemmens, M., Krska, R., & Schuhmacher, R. (2011). Evaluation of LC-high-resolution FT-Orbitrap MS for the quantification of selected mycotoxins and the simultaneous screening of fungal metabolites in food. *Food additives & contaminants. Part A, Chemistry, analysis, control, exposure & risk assessment*, *28*(10), 1457–68. doi:10.1080/19440049.2011.599340

Li, S., Kang, L., & Zhao, X.-M. (2014). A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *BioMed research international*, *2014*, 362738. doi:10.1155/2014/362738

Lin, X., Wang, Q., Yin, P., Tang, L., Tan, Y., Li, H., et al. (2011). A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, *7*(4), 549–558. doi:10.1007/s11306-011-0274-7

Lin, X., Yang, F., Zhou, L., Yin, P., Kong, H., Xing, W., et al. (2012). A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, *910*, 149–55. doi:10.1016/j.jchromb.2012.05.020

Liou, J., Wu, T.-Y., Thang, T. D., Hwang, T., Wu, C., Cheng, Y.-B., et al. (2014). Bioactive 6S-Styryllactone Constituents of *Polyalthia parviflora*. *Journal of natural products*. doi:10.1021/np5004577

Liu, W.-T., Ng, J., Meluzzi, D., Bandeira, N., Gutierrez, M., Simmons, T. L., et al. (2009). Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Analytical chemistry*, *81*(11), 4200–9. doi:10.1021/ac900114t

Lommen, A. (2009). MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical chemistry*, *81*(8), 3079–86. doi:10.1021/ac900036d

López-Pérez, J. L., Therón, R., del Olmo, E., & Díaz, D. (2007). NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics (Oxford, England)*, *23*(23), 3256–7. doi:10.1093/bioinformatics/btm516

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Analytical chemistry*, *80*(19), 7562–70. doi:10.1021/ac800954c

Mamyrin, B. A. (2001). Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*, *206*(3), 251–266. doi:10.1016/S1387-3806(00)00392-4

MarinLit. (n.d.). New Zealand: University of Canterbury. http://pubs.rsc.org/marinlit. Accessed 10 December 2014

McIntyre, C., Scott, F., Simpson, T., Trimble, L., & Vederas, J. (1989). Application of stable isotope labelling methodology to the biosynthesis of the mycotoxin, terretonin, by *Aspergillus terreus*: Incorporation of 13C-labelled acetates and methionine, 2H- and 13C, 18O-labelled ethyl 3,5-dimethylorsellinate and oxygen-. *Tetrahedron*, *45*(8), 2307–2321.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. a, et al. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, *39*(Web Server issue), W339–46. doi:10.1093/nar/gkr466

MetaBase. (2014). www.metabase.org. Accessed 3 November 2014

Meyer, F. M., Gerwig, J., Hammer, E., Herzberg, C., Commichau, F. M., Völker, U., & Stülke, J. (2011). Physical interactions between tricarboxylic acid cycle enzymes in *Bacillus subtilis*: evidence for a metabolon. *Metabolic engineering*, *13*(1), 18–27. doi:10.1016/j.ymben.2010.10.001

Miller, T. H., Musenga, A., Cowan, D. A., & Barron, L. P. (2013). Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. *Analytical chemistry*, *85*(21), 10330–7. doi:10.1021/ac4024878

Miyao, K. (1955). 14. Studies on Fungisporin. Part 2. *Journal of the Agricultural Chemical Society of Japan*, *19*(1), 86–91. http://www.tandfonline.com/doi/abs/10.1080/03758397.1955.10857269. Accessed 16 October 2014

Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., de Groot, J., van Beek, T. A., et al. (2006). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant physiology*, *141*(4), 1205–18. doi:10.1104/pp.106.078428

Moore, G. (1965). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*(1), 114–117. http://web.eng.fiu.edu/npala/eee6397ex/gordon_moore_1965_article.pdf. Accessed 10 December 2014

Moschet, C., Piazzoli, A., Singer, H., & Hollender, J. (2013). Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Analytical chemistry*, *85*(21), 10312–20. doi:10.1021/ac4021598

Nesbitt, B. F., J., O., Sargeant, K., & Sheridan, A. (1962). Toxic Metabolites of *Aspergillus flavus*. *Nature*, *195*(4846), 1062–1063. doi:10.1038/1951062a0

Ng, J., Bandeira, N., Liu, W.-T., Ghassemian, M., Simmons, T. L., Gerwick, W. H., et al. (2009). Dereplication and de novo sequencing of nonribosomal peptides. *Nature methods*, *6*(8), 596–9. doi:10.1038/nmeth.1350

Nielsen, K. F., Månsson, M., Rank, C., Frisvad, J. C., & Larsen, T. O. (2011). Dereplication of microbial natural products by LC-DAD-TOFMS. *Journal of natural products*, *74*(11), 2338–48. doi:10.1021/np200254t

Nielsen, K. F., & Smedsgaard, J. (2003). Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry methodology. *Journal of Chromatography A*, *1002*(1-2), 111–136. ISI:000183799200011

Nielsen, M. M. L., Nielsen, J. B. J. B., Rank, C., Klejnstrup, M. L., Holm, D. K., Brogaard, K. H., et al. (2011). A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in *Aspergillus nidulans*. *FEMS microbiology letters*, *321*(2), 157–66. doi:10.1111/j.1574-6968.2011.02327.x

Nordström, A., O'Maille, G., Qin, C., & Siuzdak, G. (2006). Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. *Analytical chemistry*, *78*(10), 3289–3295. http://pubs.acs.org/doi/abs/10.1021/ac060245f. Accessed 22 November 2014

Open Access in FP7. (2014). http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1300&lang=1. Accessed 15 September 2014

Patti, G. J. (2011). Separation strategies for untargeted metabolomics. *Journal of separation science*, *34*(24), 3460–9. doi:10.1002/jssc.201100532

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews. Molecular cell biology*, *13*(4), 263–9. doi:10.1038/nrm3314

Petersen, L. M., Holm, D. K., Knudsen, P. B., Nielsen, K. F., Gotfredsen, C. H., Mortensen, U. H., & Larsen, T. O. (2014). Characterization of four new antifungal yanuthones from *Aspergillus niger*. *The Journal of antibiotics*, (April), 1–5. doi:10.1038/ja.2014.130

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, *11*, 395. doi:10.1186/1471-2105-11-395

Press, C. (n.d.). Dictionary of natural products. http://dnp.chemnetbase.com/intro/. Accessed 12 November 2014

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org

Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., van Schaik, R., & Vervoort, J. (2012). Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid communications in mass spectrometry : RCM*, *26*(20), 2461–71. doi:10.1002/rcm.6364

Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology*, *32*(3), 219–23. doi:10.1038/nbt.2841

Sansone, S., Fan, T., & Goodacre, R. (2007). The metabolomics standards initiative. *Nature biotechnology*, *25*(8), 846–848. http://www.nature.com/nbt/journal/v25/n8/full/nbt0807-846b.html. Accessed 27 November 2014

Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nature reviews. Drug discovery*, *4*(1), 45–58. doi:10.1038/nrd1608

Simpson, T. (1998). Application of isotopic methods to secondary metabolic pathways. *Biosynthesis*, *195*, 1–48. http://link.springer.com/chapter/10.1007/3-540-69542-7_1. Accessed 8 October 2014

Simpson, T., & Cox, R. (2012). Polyketides in fungi. In N. Civjan (Ed.), *Natural Products in Chemical Biology* (1st ed., pp. 143–161). Hoboken: John Wiley & Sons. http://books.google.com/books?hl=en&lr=&id=0SX_GoqzEQ0C&oi=fnd&pg=PA143&dq=Polyketides+in+fungi&ots=jVH0l69XAc&sig=zUJBMxdSVP1Ei11MGhYnanqh4NY. Accessed 7 December 2014

Smith, C. a, O'Maille, G., Want, E. J., Qin, C., Trauger, S. a, Brandon, T. R., et al. (2005). METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring*, *27*(6), 747–51. http://www.ncbi.nlm.nih.gov/pubmed/16404815. Accessed 3 November 2014

Stanstrup, J., Gerlich, M., Dragsted, L. O., & Neumann, S. (2013). Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. *Analytical and bioanalytical chemistry*, *405*(15), 5037–48. doi:10.1007/s00216-013-6954-6

Stephanopoulos, G. N., Aristidou, A. A., & Nielsen, J. (1998). Review of cellular metabolism. In *Metabolic engineering - principles and methodologies* (1st ed., pp. 21–79). San Diego, CA: Academic Press.

Steyn, P. S., Vleggaar, R., & Simpson, T. J. (1984). Stable Isotope Labelling Studies. *Journal of the Chemical Society, Chemical Communications*, *3*, 765–767.

Strife, R. (2011). Orbitrap high-resolution applications. In B. N. Pramanik, M. S. Lee, & G. Chen (Eds.), *Characterization of impurities and degradants using mass spectrometry* (1st ed., pp. 109–134). John Wiley & Sons. http://onlinelibrary.wiley.com/doi/10.1002/9780470921371.ch4/summary. Accessed 14 November 2014

Suber, P. (2012). *Open Access* (First., p. 242). Masschusetts: MIT Press essential knowledge.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. a, Glass, C. K., et al. (2007). LMSD: LIPID MAPS structure database. *Nucleic acids research*, *35*(Database issue), D527–32. doi:10.1093/nar/gkl838

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., & Tomita, M. (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Current bioinformatics*, *7*(1), 96–108. doi:10.2174/157489312799304431

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. a, et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society*, *3*(3), 211–221. doi:10.1007/s11306-007-0082-2

Sysoev, A. a, Chernyshev, D. M., Poteshin, S. S., Karpov, A. V, Fomin, O. I., & Sysoev, A. a. (2013). Development of an atmospheric pressure ion mobility spectrometer-mass spectrometer with an orthogonal acceleration electrostatic sector TOF mass analyzer. *Analytical chemistry*, *85*(19), 9003–12. doi:10.1021/ac401191k

Tanenbaum, S. W., & Bassett, E. W. (1959). The Biosynthesis of Patulin: III. Rearrangement of the aromatic ring. *Journal of Biological Chemistry*, *234*(7), 1861–1866.

Tang, J. K.-H., You, L., Blankenship, R. E., & Tang, Y. J. (2012). Recent advances in mapping environmental microbial metabolisms through 13C isotopic fingerprints. *Journal of the Royal Society, Interface / the Royal Society*, *9*(76), 2767–80. doi:10.1098/rsif.2012.0396

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., & Siuzdak, G. (2012). An accelerated workflow for untargeted metabolomics using the METLIN database. *Nature Biotechnology*, *30*(9), 826–828. doi:10.1038/nbt.2348

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. E. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry*. doi:10.1021/ac300698c

Townsend, C., & Christensen, S. (1983). Stable isotope studies of anthraquinone intermediates in the aflatoxin pathway. *Tetrahedron*, *39*(21), 3575–3582. http://www.sciencedirect.com/science/article/pii/S0040402001886683. Accessed 15 August 2014

Urry, W., Wehrmeister, H., Hodge, E., & Hidy, P. (1966). The structure of zearalenone. *Tetrahedron Letters*, (27), 3109–3114. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+structure+of+zearalenone#0. Accessed 25 November 2014

Vaclavik, L., Krynitsky, A. J., & Rader, J. I. (2014). Targeted analysis of multiple pharmaceuticals, plant toxins and other secondary metabolites in herbal dietary supplements by ultra-high performance liquid chromatography-quadrupole-orbital ion trap mass spectrometry. *Analytica chimica acta*, *810*, 45–60. doi:10.1016/j.aca.2013.12.006

Van der Merwe, K. J., Steyn, P. S., Fourie, L., Scott, D. B., & Theron, J. J. (1965). Ochratoxin A, a Toxic Metabolite produced by *Aspergillus ochraceus* Wilh. *Nature*, *205*(4976), 1112–1113. doi:10.1038/2051112a0

Vélot, C., Mixon, M., Teige, M., & Srere, P. (1997). Model of a quinary structure between Krebs TCA cycle enzymes: a model for the metabolon. *Biochemistry*, *36*(47), 14271–14276. http://pubs.acs.org/doi/abs/10.1021/bi972011j. Accessed 5 December 2014

Walsh, C. T., & Fischbach, M. A. (2010). Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society*, *132*(8), 2469–2493. doi:10.1021/ja909118a

Wang, Y., Kora, G., Bowen, B. P., & Pan, C. (2014). MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Analytical chemistry*, *86*(19), 9496–503. doi:10.1021/ac5014783

Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., et al. (2012). Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), E1743–52. doi:10.1073/pnas.1203689109

Wehrens, R., Carvalho, E., & Fraser, P. D. (2014). Metabolite profiling in LC–DAD using multivariate curve resolution: the alsace package for R. *Metabolomics*. doi:10.1007/s11306-014-0683-5

WikiPathways. (n.d.). www.wikipathways.org. Accessed 10 August 2014

Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, *11*, 148. doi:10.1186/1471-2105-11-148

Wolfender, J.-L., Marti, G., & Ferreira Queiroz, E. (2010). Advances in techniques for profiling crude extracts and for the rapid identification of natural products: Dereplication, quality control and metabolomics. *Current Organic Chemistry*, *14*(16), 1808–1832. doi:10.2174/138527210792927645

Wolfender, J.-L., Marti, G., Thomas, A., & Bertrand, S. (2014). Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*. doi:10.1016/j.chroma.2014.10.091

Wolfender, J.-L., Ndjoko, K., & Hostettmann, K. (2003). Liquid chromatography with ultraviolet absorbance–mass spectrometric detection and with nuclear magnetic resonance spectrometry: a powerful combination for the on-line structural investigation of plant metabolites. *Journal of Chromatography A*, *1000*(1-2), 437–455. doi:10.1016/S0021-9673(03)00303-0

Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N., et al. (2013). Molecular networking as a dereplication strategy. *Journal of natural products*, *76*(9), 1686–99. doi:10.1021/np400413s

Yang, X., Neta, P., & Stein, S. E. (2014). Quality control for building libraries from electrospray ionization tandem mass spectra. *Analytical chemistry*, *86*(13), 6393–400. doi:10.1021/ac500711m

Yoshizawa, Y., Li, Z., Reese, P. B., & Vederas, J. C. (1990). Intact incorporation of acetate-derived di- and tetraketides during biosynthesis of dehydrocurvularin, a macrolide phytotoxin from *Alternaria cinerariae*. *Journal of the American Chemical Society*, *112*(8), 3212–3213. doi:10.1021/ja00164a053

Yue, S., Duncan, J. S., Yamamoto, Y., & Hutchinson, C. R. (1987). Macrolide biosynthesis. Tylactone formation involves the processive addition of three carbon units. *Journal of the American Chemical Society*, *109*(4), 1253–1255. doi:10.1021/ja00238a050

Zhang, J., McCombie, G., Guenat, C., & Knochenmuss, R. (2005). FT-ICR mass spectrometry in the drug discovery process. *Drug discovery today*, *10*(9), 635–642. http://www.sciencedirect.com/science/article/pii/S1359644605034380. Accessed 12 November 2014

Zhang, W., Chang, J., Lei, Z., Huhman, D., Sumner, L. W., & Zhao, P. X. (2014). MET-COFEA: A liquid chromatography/mass spectrometry data processing platform for metabolite compound feature extraction and annotation. *Analytical chemistry*, *86*(13), 6245–53. doi:10.1021/ac501162k

Zheng, H., Clausen, M. R., Dalsgaard, T. K., Mortensen, G., & Bertram, H. C. (2013). Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. *Analytical chemistry*, *85*(15), 7109–16. doi:10.1021/ac4020325

Zhu, X., Chen, Y., & Subramanian, R. (2014). Comparison of information-dependent acquisition, SWATH, and MS(All) techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Analytical chemistry*, *86*(2), 1202–9. doi:10.1021/ac403385y

Zhu, Z.-J., Schultz, A. W., Wang, J., Johnson, C. H., Yannone, S. M., Patti, G. J., & Siuzdak, G. (2013). Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nature protocols*, *8*(3), 451–60. doi:10.1038/nprot.2013.004

Zubarev, R. a, & Makarov, A. (2013). Orbitrap mass spectrometry. *Analytical chemistry*, *85*(11), 5288–96. doi:10.1021/ac4001223

6 Papers

6.1 Paper 1 – Aggressive dereplication using UHPLC-DAD-QTOF: Screening extracts for up to 3000 fungal secondary metabolites

Klitgaard, A., Iversen, A., Andersen, M. R., Larsen, T. O., Frisvad, J. C., & Nielsen, K. F.

Paper accepted in Analytical and Bioanalytical Chemistry (2014)

Anal Bioanal Chem (2014) 406:1933–1943
DOI 10.1007/s00216-013-7582-x

RESEARCH PAPER

Aggressive dereplication using UHPLC–DAD–QTOF: screening extracts for up to 3000 fungal secondary metabolites

Andreas Klitgaard · Anita Iversen · Mikael R. Andersen · Thomas O. Larsen · Jens Christian Frisvad · Kristian Fog Nielsen

Received: 11 September 2013 / Revised: 3 December 2013 / Accepted: 14 December 2013 / Published online: 18 January 2014

Abstract In natural-product drug discovery, finding new compounds is the main task, and thus fast dereplication of known compounds is essential. This is usually performed by manual liquid chromatography-ultraviolet (LC-UV) or visible light-mass spectroscopy (Vis-MS) interpretation of detected peaks, often assisted by automated identification of previously identified compounds. We used a 15 min high-performance liquid chromatography–diode array detection (UHPLC–DAD)–high-resolution MS method (electrospray ionization (ESI)$^+$ or ESI$^-$), followed by 10–60 s of automated data analysis for up to 3000 relevant elemental compositions. By overlaying automatically generated extracted-ion chromatograms from detected compounds on the base peak chromatogram, all major potentially novel peaks could be visualized. Peaks corresponding to compounds available as reference standards, previously identified compounds, and major contaminants from solvents, media, filters etc. were labeled to differentiate these from compounds only identified by elemental composition. This enabled fast manual evaluation of both known peaks and potential novel-compound peaks, by manual verification of: the adduct pattern, UV–Vis, retention time compared with log D, co-identified biosynthetic related compounds, and elution order. System performance, including adduct patterns, in-source fragmentation, and ion-cooler bias, was investigated on reference standards, and the overall method was used on extracts of *Aspergillus carbonarius* and *Penicillium melanoconidium*, revealing new nitrogen-containing biomarkers for both species.

Keywords Metabolomics · Mycotoxin · NRPS · LC–MS · UPLC · Polyketide · Nonribosomal peptide

Introduction

Fungi are an immense source of diverse natural products that can be used as drugs, food and feed additives, and industrial chemicals [1, 2]. Unfortunately fungi also have a negative side, producing mycotoxins which include some of the most immunotoxic, estrogenic, cytotoxic, and carcinogenic compounds known [3, 4].

Fast and accurate dereplication of previously described compounds is an essential and resource-saving aspect of working with natural products [1, 5–9]. The alternative, isolation and subsequent NMR-based structure elucidation, is time consuming and costly [7], and is thus primarily used in important cases, e.g. for compounds with known bioactivity.

Currently, dereplication is mainly performed by liquid chromatography–mass spectrometry (LC–MS) analysis of extracts, followed by a search of all ions of interest performed by entering the monoisotopic mass into appropriate databases. For microbial compounds, the most comprehensive database is AntiBase (Wiley-VCH, Weinheim, Germany) the 2012 version of which contains 41,000 recorded compounds. In dereplication, obtaining an elemental composition is the most efficient first step because it reduces the number of hits from a database search 3–10-fold compared with searching for a nominal mass [9–11]. For compounds below 400–600 Da, high-resolution MS (HRMS) instruments can often provide the elemental composition unambiguously if they have < 0.5–

Electronic supplementary material The online version of this article (doi:10.1007/s00216-013-7582-x) contains supplementary material, which is available to authorized users.

A. Klitgaard · A. Iversen · M. R. Andersen · T. O. Larsen ·
J. C. Frisvad · K. F. Nielsen (✉)
Department of Systems Biology, Søltofts Plads, Technical University of Denmark, 2800 Kgs., Lyngby, Denmark
e-mail: kfn@bio.dtu.dk

A. Iversen
Current address: Danish Emergency Management Agency,
Universitetsparken 2, 2100 Copenhagen, Denmark

1.5 ppm mass accuracy. In addition, time of flight (TOF)-based mass spectrometers can now provide an accurate isotope pattern, enabling an even higher degree of certainty for identification of elemental compositions [9, 12, 13].

An important extra detector is the UV–Vis diode array detection (DAD) detector, which provides information on the conjugated double-bond systems found in most secondary metabolites. This can be used to confirm or reject candidates from a database search [14, 15]. Finally, log D-based calculations can be used to predict the chromatographic elution order of compounds of interest [9].

Dereplication of peaks in extracts from genera, including *Aspergillus*, *Penicillium*, and *Fusarium*, which are known to produce many different compounds often results in many hits (1724, 1726, and 611 compounds, respectively, listed in AntiBase). Because of this, identifying compounds on the basis of UV–Vis, chromatographic retention, elution order, and comparison to biosynthetically related compounds is a slow (0.5–3 h per extract) and tedious task.

A solution could be to use MS–MS libraries [16] to identify compounds automatically. This is the preferred strategy in forensic science and toxicology, for which subjects commercial compound libraries are available [17]. However, no natural-product MS–MS libraries are currently available, because including an MS–MS spectrum for future dereplication is unfortunately not a prerequisite for publishing new structures. Because of this, only a few percent of described compounds from fungi are commercially available, and therefore only small in-house databases are available [9, 18, 19].

Another complication is that the compound adduct pattern and possible fragmentations need to be correctly interpreted, because unnoticed loss of water or addition of sodium or ammonium ions will invalidate a subsequent database search. Unambiguous determination of the accurate mass of fungal metabolites on the basis of adduct formation, dimers, and mutably charged ions can be challenging [9], but software including ACDs intelliXtract [19] and some instrument vendor software packages have algorithms for this.

To reduce the analysis time for known fungal compounds in complex extracts, we decided to test the TargetAnalysis software from Bruker Daltonics (similar software available from Waters, Thermo, Agilent, and Advanced Chemical Developments). The program was originally developed for pesticide [20] and forensic analysis [21]. TargetAnalysis can screen an extract for 3000 compounds, on the basis of mass accuracy, isotope fit, and retention time (RT), within 10–60 s, depending on how small peaks are integrated. The screening software was interfaced with our internal compound database, containing approximately 7100 compounds [9], via an in-house-built Excel application that generated automatic search lists for TargetAnalysis, and made it possible to search for the most likely adduct and/or fragment ions and to only include taxonomically relevant compounds if wanted.

Using this approach, we are able to rapidly screen extracts from several different fungi, and to annotate chromatographic peaks corresponding to known compounds. The approach makes it possible to easily identify chromatographic peaks that do not correspond to known compounds, thereby enabling one to quickly ascertain which compounds might be novel.

Materials and methods

Chemicals

Solvents were LC–MS grade, and all other chemicals were analytical grade. All were from Sigma-Aldrich (Steinheim, Germany) unless otherwise stated. Water was purified using a Milli-Q system (Millipore, Bedford, MA). ESI–TOF tune mix was purchased from Agilent Technologies (Torrance, CA, USA).

Reference standards of mycotoxins and microbial metabolites (approximately 1500, 95 % of fungal origin) had been collected over the last 30 years [9, 22, 23], either from commercial sources, as gifts from other research groups, or from our own projects. Approximately one-third of the standards were purchased from Sigma-Aldrich, Axxora (Bingham, UK), Cayman (Ann Arbor, MI), TebuBio (Le-Perray-en-Yvelines, France), Biopure (Tulln, Austria), Calbiochem, (San Diego, CA), and ICN (Irvine, CA). Standards were maintained dry at −20 °C, and were compared with original UV–VIS data, accurate mass, and relative RT from previous studies [22].

Culture extracts in the examples originated from three-point cultures on solid media, incubated for seven days in darkness at 25 °C, and extracted using a (3:2:1) (ethyl acetate:dichloromethane:methanol) mixture [24]. *Penicillium melanoconidium* IBT 30549 (IBT culture collection, author's address) was grown on CYA, and *A. carbonarius* IBT 31236 (ITEM5010) was grown on YES [24].

UHPLC–DAD–QTOFMS

A UHPCL–DAD–QTOF method was set up for screening, with typical injection volumes of 0.1–2 μl extract. Separation was performed on a Dionex Ultimate 3000 UHPLC system (Thermo Scientific, Dionex, Sunnyvale, California, USA) equipped with a 100×2.1 mm, 2.6 μm, Kinetex C_{18} column, held at a temperature of 40 °C, and using a linear gradient system composed of A: 20 mmol L^{-1} formic acid in water, and B: 20 mmol L^{-1} formic acid in acetonitrile. The flow was 0.4 ml min^{-1}, 90 % A graduating to 100 % B in 10 min, 100 % B 10–13 min, and 90 % A 13.1–15 min.

Time-of-flight detection was performed using a maXis 3G QTOF orthogonal mass spectrometer (Bruker Daltonics, Bremen, Germany) operated at a resolving power of ~50000 full width at half maximum (FWHM). The instrument was equipped with an orthogonal electrospray ionization source, and mass spectra were recorded in the range *m/z* 100–1000 as centroid spectra, with five scans per second. For calibration, 1 µl 10 mmol L^{-1} sodium formate was injected at the beginning of each chromatographic run, using the divert valve (0.3–0.4 min). Data files were calibrated post-run on the average spectrum from this time segment, using the Bruker HPC (high-precision calibration) algorithm.

For ESI^+ the capillary voltage was maintained at 4200 V, the gas flow to the nebulizer was set to 2.4 bar, the drying temperature was 220 °C, and the drying gas flow was 12.0 L min^{-1}. Transfer optics (ion-funnel energies, quadrupole energy) were tuned on HT-2 toxin to minimize fragmentation. For ESI^- the settings were the same, except that the capillary voltage was maintained at −2500 V. Unless otherwise stated, ion-cooler settings were: transfer time 50 µs, radio frequency (RF) 55 V peak-to-peak (Vpp), and pre-pulse storage time 5 µs. After changing the polarity, the mass spectrometer needed to equilibrate the power supply temperature for 1 h to provide stable mass accuracy.

Construction of the compound database

The database was constructed in ACD Chemfolder (Advanced Chemistry Development, Toronto, Canada) from:

1. reference standards (~1500) [9];
2. tentatively identified compounds (~500) [25–27];
3. compound peaks appearing in blank samples; and
4. all compounds in AntiBase2012 listed as coming from: *Aspergillus*, *Fusarium*, *Trichoderma*, *Penicillium*, *Chaetomium*, *Stachybotrys*, *Alternaria*, and *Cladosporium.*

A detailed description of the database construction can be found in the Electronic Supplementary Material, Section "Introduction".

For each compound, the known or suspected major adducts were registered as: $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M+K]^+$, $[M+H+CH_3CN]^+$, $[M+Na+CH_3CN]^+$, $[M+H-H_2O]^+$, $[M+H-2H_2O]^+$, $[M+H-H_2]^+$ (sterols), $[M+H-HCOOH]^+$, $[M+H-CH_3COOH]^+$, $[M+2H]^{2+}$, $[M+Na+H]^{2+}$ or $[M+2Na]^{2+}$ or "No ionization" in ESI^+, and in ESI^-: $[M-H]^-$, $[M-H+HCOOH]^-$, and $[M+Cl]^-$.

Creating search lists for targetanalysis

A Microsoft Excel application was created for sorting the Chemfolder database into a taxonomically relevant search-list for TargetAnalysis (elemental composition and charge state of desired adduct, and name of compound).

For labeling peaks in Bruker DataAnalysis 4.0 (DA), compounds that were available as reference standards were labeled "S-x" in front of the name. A description of the database creation procedure can be found in the Electronic Supplementary Material, Section "Introduction".

Automated screening of fungal samples

TargetAnalysis 1.2 (Bruker Daltonics, Bremen, Germany), was used to process data-files, with the following typical settings:

A) retention time (if known) as ± 1.2 min as broad, 0.8 min as medium, and 0.3 min as narrow range;
B) SigmaFit; 1000 (broad) (isotope fit not used), 40 (medium), and 20 (narrow); and
C) mass accuracy of the peak assessed at 4 ppm (broad), 2.5 ppm (medium), and 1.5 ppm (narrow).

Area cut-off was set to 3000 counts as default, but was often adjusted for very concentrated or dilute samples.

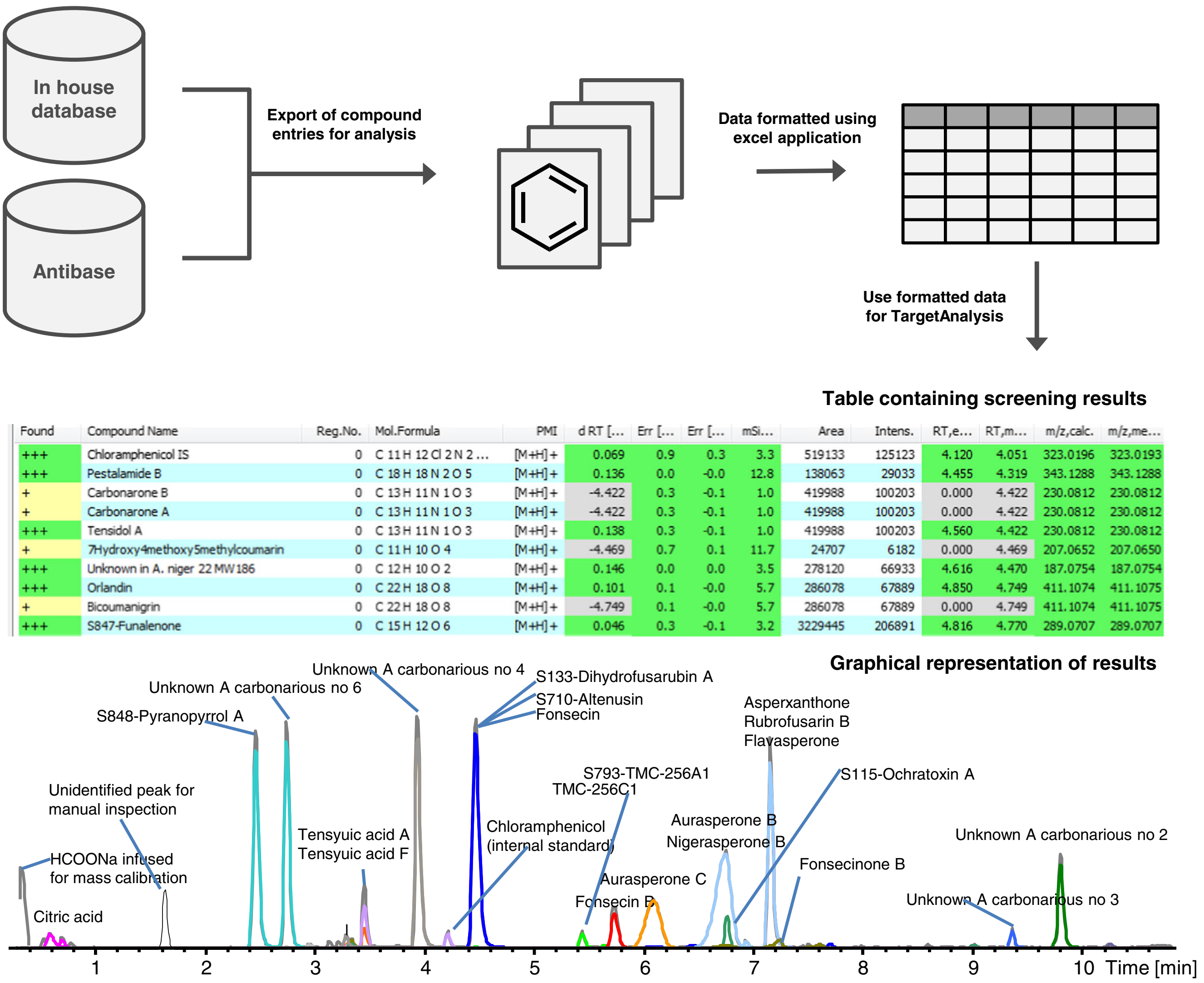
The software DataAnalysis (DA) from Bruker Daltonics was used for manual comparison of all extracted-ion chromatograms (EIC) generated by TargetAnalysis to the base peak chromatograms (BPC), to identify non-detected major peaks.

Results and discussion

The database

The database used for screening comprised 7100 compounds, of which 1500 were available reference standards and 500 were tentatively identified compounds. The database was handled in ACD Chemfolder, using a custom interface shown in Fig. S1, Electronic Supplementary Material. The database also contained legacy data from older HPLC–DAD [22], HPLC–DAD–TOFMS [9, 23], and pKa data [9] if available. Records from AntiBase needed proofreading, because we found that approximately 2–3 % of the structures had incorrect elemental compositions. We also estimate that approximately 5 % of structures published annually are not indexed.

Because TargetAnalysis could not extract both targeted and untargeted data and combine them, the fastest workflow was to overlay all the identified compounds from TargetAnalysis on the BPC chromatograms. All major non-identified peaks could then easily be observed visually (as shown in Fig. 1), dereplicated, and added to the database as a tentatively identified [9, 25] or unknown compound. Subsequently it was clear that the signals from compounds originating from filters, media blanks etc. were most efficiently handled by including them in the database, so that they would be annotated and



Found	Compound Name	Reg.No.	Mol.Formula	PMI	d RT [...	Err [...	Err [...	mSi...	Area	Intens.	RT,e...	RT,m...	m/z,calc.	m/z,me...
+++	Chloramphenicol IS	0	C 11 H 12 Cl 2 N 2 ...	[M+H]+	0.069	0.9	0.3	3.3	519133	125123	4.120	4.051	323.0196	323.0193
+++	Pestalamide B	0	C 18 H 18 N 2 O 5	[M+H]+	0.136	0.0	-0.0	12.8	138063	29033	4.455	4.319	343.1288	343.1288
+	Carbonarone B	0	C 13 H 11 N 1 O 3	[M+H]+	-4.422	0.3	-0.1	1.0	419988	100203	0.000	4.422	230.0812	230.0812
+	Carbonarone A	0	C 13 H 11 N 1 O 3	[M+H]+	-4.422	0.3	-0.1	1.0	419988	100203	0.000	4.422	230.0812	230.0812
+++	Tensidol A	0	C 13 H 11 N 1 O 3	[M+H]+	0.138	0.3	-0.1	1.0	419988	100203	4.560	4.422	230.0812	230.0812
+	7Hydroxy4methoxy5methylcoumarin	0	C 11 H 10 O 4	[M+H]+	-4.469	0.7	0.1	11.7	24707	6182	0.000	4.469	207.0652	207.0650
+++	Unknown in A. niger 22 MW186	0	C 12 H 10 O 2	[M+H]+	0.146	0.0	0.0	3.5	278120	66933	4.616	4.470	187.0754	187.0754
+++	Orlandin	0	C 22 H 18 O 8	[M+H]+	0.101	0.1	-0.0	5.7	286078	67889	4.850	4.749	411.1074	411.1075
+	Bicoumanigrin	0	C 22 H 18 O 8	[M+H]+	-4.749	0.1	-0.0	5.7	286078	67889	0.000	4.749	411.1074	411.1075
+++	S847-Funalenone	0	C 15 H 12 O 6	[M+H]+	0.046	0.3	-0.1	3.2	3229445	206891	4.816	4.770	289.0707	289.0707

Fig. 1 Example of workflow for screening of fungal extracts, in this case an extract from *Aspergillus niger*. The database maintained at our center contains 7100 records, comprising reference standards and their associated MS and UV data. For a specific analysis it is possible to export relevant entries from the database and, via an in-house-built Excel application, convert these to a format that can be imported into TargetAnalysis. Analysis via TargetAnalysis then yields both a graphical interpretation of the results and a table of the data

labeled by TargetAnalysis. This led to labeling peaks with the reference standard number (Fig. 1), indicating whether a compound was available as a reference standard for subsequent reanalysis.

The results from the analysis of an extract from *A. niger* are depicted in Fig. 1, illustrating the major disadvantage of the method. It can be seen that several compounds have been annotated to the same chromatographic peak, because numerous compounds in the search list had the same elemental composition and unknown RT. This is the major reason for not including, e.g., all 41,000 compounds from AntiBase2012 in the search list, because it contains up to 130 compounds with the same elemental composition [9]. For each experiment it is therefore important to use a search list from which highly unlikely compounds, for example metabolites from other organisms, are restricted. If no compounds are found, reanalysis can be conducted using a list of all elemental compositions in the database of choice.

Handling adducts and in-source fragmentation

Early analytical work (results not shown), using atmospheric-pressure chemical ionization (APCI)$^+$, APCI$^-$, ESI$^+$ and ESI$^-$ ionization for analysis of extracts from *A. niger* and *A. nidulans*, did not reveal superior ionization by APCI over ESI for any compound. Thus APCI was not further pursued, although there must be some apolar and/or semi-volatile compounds that are better ionized by APCI.

Adduct formation on the maXis 3G ion-source was surprisingly different from that observed on our 10-years-older Waters Micromass LCT (z-spray source) [9], even though exactly the same eluents were used. In ESI$^+$ mode we observed many compounds using the maXis, e.g. chloramphenicol and several anthraquinones, which were not previously detected by the LCT system using ESI$^+$. It remains to be investigated whether this was caused by the grounded needle (and thus a potential of −42000 V over the source), the ion-funnel, or other changes in the source. Ammonium adducts were also far less abundant on the maXis, and formation seemed to be efficiently suppressed by the drying gas, leading to spectra with abundant $[M+H]^+$ and $[M+Na]^+$, because most compounds with high affinity for ammonium also have a high affinity for sodium [9].

An interesting phenomenon observed with ESI$^+$ was that in the end of the gradient, when the acetonitrile content was close to 100 %, ionization seemed to favor formation of $[2M+Na]^+$ ions. For such analytes as the variecoxanthones and emericellin (Fig. S2, Electronic Supplementary Material) the $[2M+Na]^+$ ion (*m/z* 839.3766) had a 5–10-fold-higher intensity than $[M+H]^+$. This was presumably caused by the high acetonitrile content, which would have facilitated fast evaporation, and acidic compounds may thus hold the residual Na^+ by ion exchange before evaporation from the droplet.

Macrocyclic trichothecenes in extracts from *Baccharis megapotamica* [28] revealed that the adduct pattern was concentration-dependent, with the highest intensity $[M+Na]^+$ occurring at low concentrations of the analyte (Fig. S3, Electronic Supplementary Material). This is probably the result of limited Na^+, and thus $[M+H]^+$ is most abundant when Na^+ is depleted. On full-scan instruments this phenomenon can be regarded as *adduct displacement*, whereas it will be observed as ion suppression on MS–MS instruments if only one of $[M+H]^+$ or $[M+Na]^+$ is measured. For MS–MS characterization of compounds that favor sodium adducts, we have in several applications used ammonium formate as buffer to depress sodium adduct formation. In one example we also changed the sodium formate calibration solution to a polyethylene glycol mixture, and switched the glass water-solvent bottle to plastic.

Ergosterol and related steroles were, surprisingly, detected as $[M+H-H_2]^+$ ions, whereas, e.g., cholesterol was detected as $[M+H-H_2O]^+$.

ESI$^-$ ionized acidic compounds (carboxylic acids, enoles and phenols) well, because of easy disassociation of H^+, and also proved superior to ESI$^+$ unless the target compounds also contained amine or amide functionalities. Compounds without acidic protons, that were observed as $[M+HCOO]^-$ on both Waters LCT z-spray source instrumentation [9] and an Agilent 6550 QTOF, were often not detected at all using the maXis system.

Ion-source fragmentation was unavoidable for very fragile molecules, but was mainly observed as water loss for compounds that formed sodium adducts: jumping from $[M+Na]^+$ to $[M+H-H_2O]^+$, with *m/z* 39.9925, and occasionally also to $[M+H-2H_2O]^+$, with *m/z* 58.0031. Thus the sodium adducts could be an advantage when screening fragile compounds. Cases where $[M+H]^+$ was not observed were much more predominant on the maXis than on the Waters LCT (z-spray source). In-source fragmentation could be minimized by lowering the potential of the quadrupole and between the funnels, but could not be abolished because this would lead to >10 % loss of sensitivity. We therefore included $[M+H-H_2O]^+$ and $[M+H-2H_2O]^+$ in the database of compounds losing H_2O during ESI$^+$ (often an alcohol group with α-carbon was available for elimination via double-bond formation) [9].

The screening process was also performed, using similar samples, on an Agilent 1290 UHPLC–6550 QTOF system, using Agilent Masshunter's *Find By Formula* option. This function could handle different adducts and simple losses, for example water loss, theoretically ensuring that no compounds were overlooked. This, however, also resulted in many more false positives, because all peaks are believed to correspond to, e.g., an $[M+H-H_2O]^+$ ion, even if the peaks also fit the $[M+H]^+$ of another compound. ACD's MS Workbook Suite intelliXtract function (v. 12) was also tested. The software could assign the whole adduct, multimer and fragment pattern for a peak, but required the presence of a $[M+H]^+$ or $[M-H]^-$ ion. This software was approximately 50–100 times more time-consuming than Brukers TargetAnalysis for a list of 3000 compounds, but does work for smaller databases [19].

Molecules with masses above 1000 Da, which include many NRPs (e.g. lipopeptides and peptaibols), all produced doubly and often also triply charged ions, thus appearing in the scan window of *m/z* 100–1000. The only two exceptions were special cyclic peptides, for example cereulide and valinomycin, which are very strong K^+-ionophores and therefore only produced $[M+Na]^+$ and $[M+K]^+$ ions [29].

The adduct formation behavior of some compounds can however be hard to predict. This was observed for an extract of *Phoma levellei* [30] (incorrectly identified as *Cladosporium uredinicola*), for which the ESI$^-$ spectrum of 3-Hydroxy-2,5-dimethylphenyl 3-[(2,4-Dihydroxy-3,6-dimethylbenzoyl)oxy]-6-hydroxy-2,4-dimethylbenzoate (Fig. 2) indicated the presence of several co-eluting compounds. Deconvolution of the ions revealed that ions labeled A–D came from the same compound. Ion C corresponded to $[M-H]^-$, A and B were fragments, and D was a composite ion of $[M-H]^-$ and one fragment-ion A.

Ion-cooler bias

The maXis 3G is equipped with a hexapole ion-cooler, which collects the ions, reduces their kinetic energy, and ejects them into the orthogonal accelerator in the TOF mass analyzer. Our results reveal that the ion cooler settings have a significant effect on the intensities of the ions in the measured mass range (Fig. S4, Electronic Supplementary Material).

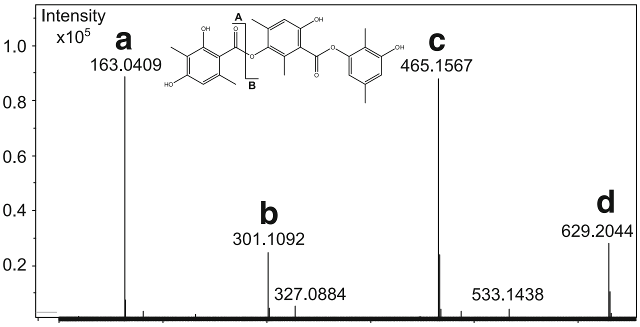


Fig. 2 ESI$^-$ spectrum of 3-Hydroxy-2,5-dimethylphenyl 3-[(2,4-Dihydroxy-3,6-dimethylbenzoyl)oxy]-6-hydroxy-2,4-dimethylbenzoate, showing M−H]$^-$ (C) and fragment ions **a** and **b**. **d** is a composite of ions **a** and **c**

Three variables were important:

1. the *ion-cooler radio frequency* (RF), which sets the voltage for the ion-cooler;
2. the transfer time, which is the time window wherein ions are transmitted into the TOF; and
3. the *pre-pulse storage time*, which will apply a low mass limit and is a delay between the transfer time and the TOF pulser. Higher values favored the transfer of higher *m*/*z* ions, but also discriminated low *m*/*z* ions.

Figure S4 (Electronic Supplementary Material) shows selected results from analysis using seven different transfer times. The results revealed that the ion-cooler “window” for low mass compounds is narrow, and the settings used to obtain an optimum signal for lower *m*/*z* ions resulted in low intensities of higher *m*/*z* ions, and vice versa. For analytes with *m*/*z* lower than 100 (data not shown), the optimum settings excessively discriminated the signal intensity of higher *m*/*z* values. At an ion cooler RF value of 30 Vpp, the signal of *m*/*z* 91 was highly suppressed at all transfer times.

Our in-house database contained 7100 compounds with a $[M+H]^+$ in the range *m*/*z* 100–1000. Of these, 14 % will have a $[M+H]^+<226$ *m*/*z* and will reach only 30 % of their maximum intensity using standard screening settings. For ions smaller than *m*/*z* 130 the signal suppression will be extensive, but luckily less than 1 % of the compounds in our in-house database and AntiBase have masses this low [9]. If a target compound was in the mass range below *m*/*z* 130, the optimum ion-cooler settings resulted in an intensity of less than 10 % for compounds with an *m*/*z*>226, and of only 5 % of the signal from compounds with an *m*/*z*>600. It is important to be aware of this signal discrimination in some mass ranges under different ion-cooler settings.

Effect of detector overload on isotope pattern and mass accuracy

Because fungal extracts contain many different compounds with varying concentrations and ionization efficiencies, screening of extracts routinely resulted in analysis of compounds with intensities higher than $2–3\times10^6$ counts, which overloaded the detector of the maXis QTOF (this problem was much more severe on older TOF instruments [9]). This caused an *m*/*z* shift to higher values, which in the worst case resulted in an increase of up to 3–4 ppm. This also led to a distorted isotopic pattern, where the A+1, A+2 isotopomers were too intense relative to the A isotopomer. To avoid false negative results in TargetAnalysis, it was thus crucial to set a wide range (5 ppm) on the isotope fit and mass accuracy. However, these high-intensity peaks could be easily spotted by the peak height in the results table, after which data for the chromatographic peak could be examined from scans where the detector was not overloaded. The isotope fit was highly dependent on a weekly detector tuning, and the medium and narrow-range settings had to be increased twofold when the detector had not been tuned within the week.

Aggressive dereplication reveals new metabolites from highly toxic spoilage fungus *Aspergillus carbonarius*

A. carbonarius is a physiologically very well investigated species because of its contamination of grapes, and the subsequent contamination of wine and raisins, with ochratoxin A [31]. However, other compounds from the fungus have attracted little attention. As well as this toxin, it is capable of producing carbonarones and pestalamide A (former tensidol B) [32], pyranonigrins, carbonarins, organic acids, and aurasperones [26].

Extracts from *A. carbonarius* cultivated on YES agar were screened for 3000 compounds:

1. compounds from *Aspergillus* (with an emphasis on *Aspergillus* section Nigri compounds) and *Penicillium*;
2. all standards available in our collection; and
3. all unidentified peaks registered in our database.

With a high area cut-off of 10,000 counts, 66 peaks were integrated (Table 1); however, 16 of these compounds were from peaks assigned to several compounds (up to five) and thus only 45 true peaks were annotated. The major peaks in the sample are displayed in Fig. 3.

Citric acid was detected as the sodium adduct and as two peaks because of poor retention on the column, which occurred because the LC–MS method is not well suited to such polar compounds. Kojic acid was incorrectly identified as another compound with the same elemental composition, because neither the RT nor the characteristic UV spectrum matched a reference standard.

Three interesting nitrogen-containing biomarkers for this species, with elemental compositions $C_{11}H_{11}NO_5$ and

Table 1 Results from the aggressive dereplication of an extract of *Aspergillus carbonarius* grown on YES agar

Peak	Class	Comment	Compound name	Molecular formula	Err (ppm)	mSigma	Area (arbitrary units)	RT measured (min)	RT expected (min)
A	+++	OK double peak caused by injection	Citric acid	$C_6H_7NaO_7$	0.1	8	351577	0.609	0.61
B	+++	OK double peak caused by injection	Citric acid	$C_6H_7NaO_7$	0.1	3	256614	0.719	0.72
C	+++		BL-UK Cla no 60 pos. blank	$C_{10}H_{13}N_5O_4$	0.9	7	22958	0.722	0.72
D	+	Wrong, UV and RT do not fit	S96-Kojic acid	$C_6H_6O_4$	0.9	9	14965	0.791	1.2
E	+++		BL-UK Cla no 72 pos. blank	$C_{10}H_{16}N_2O_2$	0.2	11	15379	1.807	1.75
F	+++		BL-UK Cla no 95 pos. blank	$C_7H_{14}N_2O_3$	1.2	6	15141	2.243	2.1
G	+++	OK	S848-Pyranonigrin A	$C_{10}H_9N_1O_5$	0.9	19	5428853	2.475	2.36
H	+++		UK in A. ni 2	$C_{10}H_9N_1O_4$	0.4	17	24641	2.756	2.906
I	+++	Interesting new biomarker	UK A car no 6	$C_{11}H_{11}N_1O_5$	0.6	17	5203919	2.756	2.751
J	+++		UK in A. ni 19	$C_{18}H_{37}NaO_{10}$	0.2	10	13945	2.892	2.844
K	+++		BL-UK Cla no 11 pos. blank	$C_{11}H_{18}N_2O_2$	1.3	10	29484	2.912	3.09
L	+++		UK in A. ni 2	$C_{10}H_9N_1O_4$	1.2	1	90082	2.962	2.906
M	+++		BL-UK Cla no 12 pos. blank	$C_{11}H_{18}N_2O_2$	0.2	5	44764	3.14	3.09
N	+++	Interesting new biomarker	UK A car no 4	$C_{18}H_{21}N_1O_2$	0.1	16	350827	3.295	3.288
O	+++		UK in A. ni 16	$C_{22}H_{45}NaO_{12}$	0.6	18	13611	3.299	3.25
P	+	No confused by the A isomer	Tensyuic acid A	$C_{11}H_{16}O_6$	0.2	7	96858	3.344	0
P	+	Presumably OK	Tensyuic acid F	$C_{11}H_{16}O_6$	0.2	7	96858	3.344	0
Q	++		UK A car no 4	$C_{18}H_{21}N_1O_2$	0.1	15	48785	3.592	3.288
Q	++		UK A car no 1	$C_{18}H_{21}N_1O_2$	0.1	15	48785	3.592	3.923
R	+++		UK in A. ni 5	$C_{21}H_{44}O_{11}$	0.3	14	10039	3.63	3.581
S	+	OK but may be the C isomer	Pyranonigrin B	$C_{11}H_{11}N_1O_6$	0.5	9	55596	3.76	0
S	+	OK but may be the B isomer	Pyranonigrin C	$C_{11}H_{11}N_1O_6$	0.5	9	55596	3.76	0
T	+++		UK in A. ni 7	$C_{23}H_{47}NaO_{12}$	0.4	37	17040	3.767	3.72
U	++		UK A car no 4	$C_{18}H_{21}N_1O_2$	0.7	15	5265217	3.944	3.288
U	+++		UK A car no 1	$C_{18}H_{21}N_1O_2$	0.7	15	5265217	3.944	3.923
V	+		Pyranonigrin D	$C_{11}H_9N_1O_5$	0.2	9	17070	3.946	0
W	+++	Internal standard	Chloramphenicol IS	$C_{11}H_{12}Cl_2N_2O_5$	0.2	31	326301	4.219	4.12
X	+++	No confused by Fonsecin	S133-Dihydrofusarubin A	$C_{15}H_{14}O_6$	1.1	25	6829770	4.47	4.75
X	++	Wrong, UV and RT do not fit	S710-Altenusin	$C_{15}H_{14}O_6$	1.1	25	6829770	4.47	4.908
X	+++	OK	Fonsecin	$C_{15}H_{14}O_6$	1.1	25	6829770	4.47	4.45
Y	+	OK but one must be a new isomer	Tensyuic acid B	$C_{12}H_{18}O_6$	1.1	24	21361	4.554	0
Z	+	OK but one must be a new isomer	Tensyuic acid B	$C_{12}H_{18}O_6$	1	22	10189	4.681	0
AA	+++	OK	S133-Dihydrofusarubin A	$C_{15}H_{14}O_6$	1	46	10340	5.031	4.75
AA	+++	Wrong, UV and RT do not fit	S710-Altenusin	$C_{15}H_{14}O_6$	1	46	10340	5.031	4.908
AB	++	No confused by Dihydrofusarubin A	Fonsecin	$C_{15}H_{14}O_6$	1	46	10340	5.031	4.45
AC	++		Aurasperone C	$C_{31}H_{28}O_{12}$	0.5	37	15414	5.249	5.94
AD	+++	No confused by TMC-256A1	TMC-256C1	$C_{15}H_{12}O_5$	0.6	18	349791	5.437	5.67
AD	+++	OK	S793-TMC-256A1	$C_{15}H_{12}O_5$	0.6	18	349791	5.437	5.37
AE	++		Aurasperone C	$C_{31}H_{28}O_{12}$	0.4	41	19423	5.494	5.94
AF	+++	OK	TMC-256C1	$C_{15}H_{12}O_5$	0.3	7	65429	5.641	5.67
AF	+++	No confused by TMC-256C1	S793-TMC-256A1	$C_{15}H_{12}O_5$	0.3	7	65429	5.641	5.37

Table 1 (continued)

Peak	Class	Comment	Compound name	Molecular formula	Err (ppm)	mSigma	Area (arbitrary units)	RT measured (min)	RT expected (min)
AG	+++		Fonsecin B	$C_{16}H_{16}O_6$	0.8	30	1055089	5.729	5.66
AH	+	Wrong water-loss ion of C isomer	Niasperone C	$C_{31}H_{26}O_{11}$	1	9	76397	6.08	0
AH	+++	Wrong water-loss ion of C isomer	Aurasperone F	$C_{31}H_{26}O_{11}$	1	9	76397	6.08	6.303
AH	+++		Aurasperone C	$C_{31}H_{28}O_{12}$	1.1	23	3247597	6.081	5.94
AI	++		UK in A. ni 23	$C_{15}H_{33}N_{17}O_6$	0.2	62	39935	6.344	6.23
AJ	++		UK in A. ni 20	$C_{28}H_{36}N_4O_5$	0.9	25	49747	6.397	6.043
AK	+	OK but may be a different isomer	Niasperone C	$C_{31}H_{26}O_{11}$	0.8	11	115620	6.434	0
AK	+++	OK but may be a different isomer	Aurasperone F	$C_{31}H_{26}O_{11}$	0.8	11	115620	6.434	6.303
AL	+++	Wrong water-loss ion of B isomer	Aurasperone E	$C_{32}H_{28}O_{11}$	0.9	23	186091	6.728	6.62
AL	++	Wrong water loss ion of B isomer	Aurasperone E-isomer	$C_{32}H_{28}O_{11}$	0.9	23	186091	6.728	7.104
AL	++	Wrong water loss ion of B isomer	Fonsecinone B	$C_{32}H_{28}O_{11}$	0.9	23	186091	6.728	7.472
AL	+	OK but may be a different isomer	Niasperone B	$C_{32}H_{30}O_{12}$	1.3	22	6659679	6.728	0
AL	+++	OK but may be a different isomer	Aurasperone B	$C_{32}H_{30}O_{12}$	1.3	22	6659679	6.728	6.605
AM	+++	OK	S115-Ochratoxin A	$C_{20}H_{18}Cl_1N_1O_6$	0.7	50	693721	6.75	6.62
AN	+	OK but may be a different isomer	Niasperone C	$C_{31}H_{26}O_{11}$	1.5	9	62334	6.779	0
AN	++	OK but may be a different isomer	Aurasperone F	$C_{31}H_{26}O_{11}$	1.5	9	62334	6.779	6.303
AO	++	No rubrofusarin	Flavasperone	$C_{16}H_{14}O_5$	0.7	20	146028	6.923	7.2
AO	+++	OK	Rubrofusarin B	$C_{16}H_{14}O_5$	0.7	20	146028	6.923	7.029
AP	+++	OK	Flavasperone	$C_{16}H_{14}O_5$	0.6	14	4285585	7.145	7.2
AP	++	No flavasperone	Rubrofusarin B	$C_{16}H_{14}O_5$	0.6	14	4285585	7.145	7.029
AQ	++	OK but may be a different isomer	Aurasperone E	$C_{32}H_{28}O_{11}$	0.2	35	300587	7.221	6.62
AQ	+++	OK but may be a different isomer	Aurasperone E-isomer	$C_{32}H_{28}O_{11}$	0.2	35	300587	7.221	7.104
AQ	+++	OK but may be a different isomer	Fonsecinone B	$C_{32}H_{28}O_{11}$	0.2	35	300587	7.221	7.472
AR	+++	OK but may be a different isomer	Fonsecinone B	$C_{32}H_{28}O_{11}$	0.7	15	156648	7.588	7.472
AS	+++		S598-Linoleic acid	$C_{18}H_{32}O_2$	0.6	11	104992	10.23	10.17

mSigma, fit of isotope pattern (see text for more details); RT, retention time

$C_{18}H_{21}NO_2$ (two isomers), were detected (unknown 1, 4, and 6), and these were not detected for other black *Aspergilli* (results not shown). Ochratoxin A, which was produced in very high amounts, is an interesting case because its precursors, ochratoxin α and B, were not detected even in trace amounts, indicating that the biosynthetic enzymes are very efficient.

Several closely eluting same-elemental-composition groups were observed and needed manual verification. For example, the rationale for identifying peak AA, as seen in Table 1, was:

1. Altenusin $C_{15}H_{14}O_6$ was from *Alternaria* and thus taxonomically unlikely. RT was within the limits where a reference standard should be co-analyzed in the sequence for verification. Inspection of the UV–Vis data led to easy elimination, and so did the presence of a perfectly co-eluting $[M+Na]^+$ ion with $M=C_{15}H_{16}O_7$.
2. Fonsecin could be eliminated by the same arguments.
3. Finally, dihydrofusarubin A was identified as the correct compound, on the basis of its perfectly matching UV–Vis spectrum and its $[M+H-H_2O]^+$ and $[M+Na]^+$ ions. However, dihydrofusarubin A was only detected because

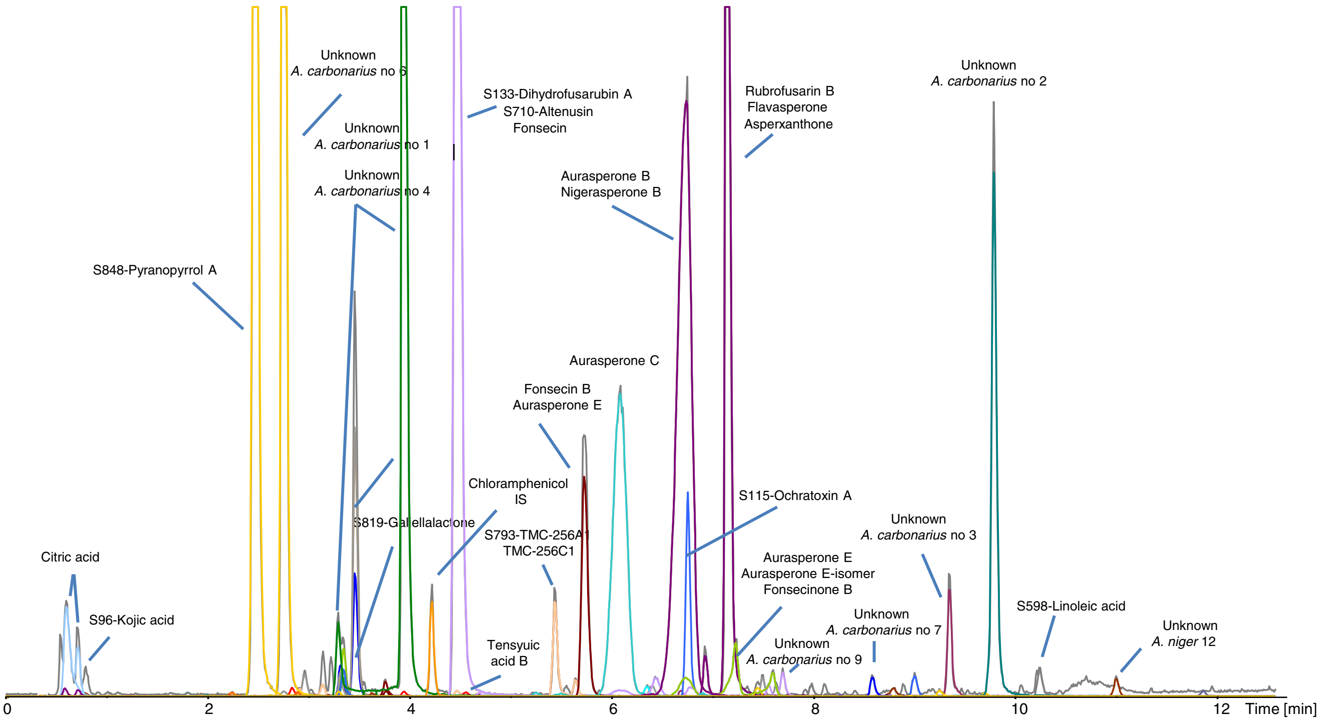


Fig. 3 Analyzed fungal extract from *A. carbonarius* cultivated on YES media. The chromatogram is overlaid with EIC from detected compounds, facilitating easy dereplication. The chromatogram has been scaled to better illustrate the smaller peaks

it was registered in the database in the form $[M+H-H_2O]^+$.

The AL peak (Table 1) must be niasperone B or aurasperone B, but could not be differentiated without a reference standard. In that case, water-loss ions led to the peak being wrongly assigned to aurasperone E and one of its isomers, and to fonsecinone B.

The pair flavasperone and rubrofusarin B should both be produced when the dimeric naphtho-γ-pyrones are produced, and a log D calculation revealed that rubrofusarin B should elute first.

Differentiating the tensyuic acids was more ambiguous, because the reported elution pattern from reversed phase is F, A, B, C, D, and E [33], with F and B having the same elemental composition, and A and B almost co-eluting. Manual inspection of the screening results was therefore necessary to attempt to distinguish between the isomers. This revealed that the first-eluting tensyuic acid was most probably the F isomer (1.3 min to the B isomer). However, the B isomer could not be unambiguously assigned as one of the two peaks Y or Z, because only one compound with $C_{12}H_{18}O_6$ is described.

In conclusion, the method very quickly identified suspected compounds from *A. carbonarius*. Besides this, a novel group of nitrogen-containing compounds, and tensyuic acids and numerous other compounds from related species, were detected. This indicated that, from a toxicological perspective, more compounds needed to be considered. A problem is that many of the closely related niasperones, aurasperones, and fonsecinones have identical elemental compositions and UV–Vis spectra and are very difficult to differentiate. To enable differentiation, we are currently considering an MS–HRMS library approach, as done for a toxic substance library [17]. However, TargetAnalysis does not presently have the capability to handle MS–HRMS data or pseudo-MS–MS data including MS-E, MS-All and/or All-Ions [21]. A further example of aggressive dereplication applied to *Penicillium melanoconidium* can be found in Electronic Supplementary Material Section "Materials and methods" and Tables S1 and S2. Here, several families of compounds not previously seen in the species were detected (Fig. S5, Electronic Supplementary Material). This included the highly toxic verrucosidins, and a presumed novel dideoxyverrucosidin. Chrysogine, a compound often detected in cereal-infecting Fusaria, was also detected, indicating that this may be an important virulence factor. The example shows how the aggressive dereplication procedure was used to detect known compounds not previously detected from the fungus. The results illustrate that all major peaks in the chromatogram were overlaid with an EIC, proving the effectiveness of the procedure and also indicating that it is a chemically very well characterized species.

Conclusion

Screening fungal secondary metabolites on the basis of elemental composition and lists restricted to the same genus and related fungi was proved to be an efficient way to quickly investigate fungal extracts. By overlaying detected peaks and BPC chromatograms, the approach gives a visual overview of a sample and indicates whether it is a previously uninvestigated species by establishing how many peaks are unlabeled. This approach can also be used on other vendor instrumentations using analogous software packages, for example: TargetLynx (Waters), TraceFinder (Thermo), MassHunter Find By Formula (Agilent), and ACD intelliXtract (Advanced Chemical Developments).

Labeling of co-identified biosynthetic related compounds could also be directly identified from the peak, making it possible to quickly assess the elution order of such compounds.

However, adduct formation and simple fragmentations are still important challenges to address when working with analytes that do not only form $[M+H]^+$ or $[M-H]^-$. Using a database approach and learning from the spectrometric behavior of reference standards can minimize problems with false-negative results. More efficient adduct-analysis software will further improve this setup [9, 21].

A further improvement to be introduced is use of MS–MS [17, 19, 34] and/or pseudo-MS–MS (MS-All, MS-E, All Ions) [21] to obtain compound-specific fragment ions for confirmation of reference standards, reducing the need to run many thousands of reference standards on a daily basis. The addition of qualifier and/or fragment ions from libraries and literature data will help to minimize the number of wrongly annotated ions with the same elemental composition, which is the main disadvantage of this method.

Acknowledgements This work was supported by the Danish Research Agency for Technology and Production (grant 09-064967), and the EEC project MycoRed (KBBE-2007-222690-2). Dr Sven Meyer and Dr Verena Tellström from Bruker Daltonics are acknowledged for fruitful discussions and help on scripting and setting up TargetAnalysis.

References

1. Zengler K, Paradkar A, Keller M (2009) in: Zhang L and Demain AL (Eds.) Natural Products: Drug Discovery and Therapeutic Medicine, Humana Press Inc., Totowa.
2. Butler MS (2004) The Role of Natural Product Chemistry in Drug Discovery. J Nat Prod 67:2141–2153
3. Miller JD (2008) Mycotoxins in small grains and maize: Old problems, new challenges. Food Addit Contam 25:219–230
4. Shephard GS (2008) Impact of mycotoxins on human health in developing countries. Food Addit Contam 25:146–151
5. Bitzer J, Kopcke B, Stadler M, Heilwig V, Ju YM, Seip S, Henkel T (2007) Accelerated dereplication of natural products, supported by reference libraries. Chimia 61:332–338
6. Bobzin SC, Yang S, Kasten TP (2000) LC-NMR: A new tool to expedite the dereplication and identification of natural products. J Ind Microbiol Biotechnol 25:342–345
7. Cordell GA, Shin YG (1999) Finding the needle in the haystack. The dereplication of natural products extracts. Pure Appl Chem 71:1089–1094
8. Zhang L (2005) in: Zhang L and Demain AL (Eds.) Natural Products: Drug Discovery and Therapeutic Medicine, Humana Press Inc., Totowa.
9. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO (2011) Dereplication of microbial natural products by LC-DAD-TOFMS. J Nat Prod 74:2338–2348
10. Bueschl C, Kluger B, Berthiller F, Lirk G, Winkler S, Krska R, Schuhmacher R (2012) MetExtract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics, research. Bioinformatics 28:736–738
11. Sleno L (2012) The use of mass defect in modern mass spectrometry. J Mass Spectrom 47:226–236
12. Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinforma 7:234
13. Erve JC, Gu M, Wang Y, DeMaio W, Talaat RE (2009) Spectral Accuracy of Molecular Ions in an LTQ/Orbitrap Mass Spectrometer and Implications for Elemental Composition Determination. J Am Mass Spectr 20:2058–2069
14. Hansen ME, Smedsgaard J, Larsen TO (2005) X-Hitting: An Algorithm for Novelty Detection and Dereplication by UV Spectra of Complex Mixtures of Natural Products. Anal Chem 77:6805–6817
15. Larsen TO, Petersen BO, Duus JO, Sørensen D, Frisvad JC, Hansen ME (2005) Discovery of New Natural Products by Application of X-hitting, a Novel Algorithm for Automated Comparison of Full UV Spectra, Combined with Structural Determination by NMR Spectroscopy. J Nat Prod 68:871–874
16. Fredenhagen A, Derrien C, Gassmann E (2005) An MS/MS Library on an Ion-Trap Instrument for Efficient Dereplication of Natural Products. Different Fragmentation Patterns for [M + H] + and [M + Na] + Ions. J Nat Prod 68:385–391
17. Broecker S, Herre S, Wust B, Zweigenbaum J, Pragst F (2011) Development and practical application of a library of CID accurate mass spectra of more than 2,500 toxic compounds for systematic toxicological analysis by LC-QTOF-MS with data-dependent acquisition. Anal Bioanal Chem 400:101–117
18. Bijlsma L, Sancho JV, Hernandez F, Niessen WMA (2011) Fragmentation pathways of drugs of abuse and their metabolites based on QTOF MS/MS and MSE accurate-mass spectra. J Mass Spectrom 46:865–875
19. El-Elimat T, Figueroa M, Ehrmann BM, Cech NB, Pearce CJ, Oberlies NH (2013) High-Resolution MS, MS/MS, and UV Database of Fungal Secondary Metabolites as a Dereplication Protocol for Bioactive Natural Products. J Nat Prod 76:1709–1716
20. Meyer S, Ketterlinus R (2011) Confirming Multi-Target Screening Full Scan Workflows of Pesticides in Food. Lc Gc Europe S1:11
21. Ojanpera S, Pelander A, Pelzing M, Krebs I, Vuori E, Ojanpera I (2006) Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. Rapid Commun mass sp 20:1161–1167
22. Frisvad JC, Thrane U (1987) Standardised High-Performance Liquid Chromatography of 182 mycotoxins and other fungal metabolites based on alkylphenone retention indices and UV-VIS spectra (Diode Array Detection). J Chromatogr 404:195–214

23. Nielsen KF, Smedsgaard J (2003) Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for de-replication by standardised liquid chromatography-UV-mass spectrometry methodology. J Chromatogr A 1002:111–136
24. Samson RA, Houbraken J, Thrane U, Frisvad JC, Andersen B (2010) Food and Indoor Fungi. CBS Laboratory Manual Series 2, CBS, Utrecht.
25. Månsson M, Phipps RK, Gram L, Munro MH, Larsen TO, Nielsen KF (2010) Explorative Solid-Phase Extraction (E-SPE) for Accelerated Microbial Natural Product Discovery, Dereplication, and Purification. J Nat Prod 73:1126–1132
26. Nielsen KF, Mogensen JM, Johansen M, Larsen TO, Frisvad JC (2009) Review of secondary metabolites and mycotoxins from the *Aspergillus niger* group. Anal Bioanal Chem 395:1225–1242
27. Frisvad JC, Rank C, Nielsen KF, Larsen TO (2009) Metabolomics of *Aspergillus fumigatus*. Med Mycol 47:S53–S71
28. Oliveira-Filho JC, Carmo PMS, Iversen A, Nielsen KF, Barros CLS (2012) Experimental poisoning by *Baccharis megapotamica* var. *weirii* in buffalo. Pesquisa vet Brasil 32:383–390
29. Thorsen L, Paulin A, Hansen BM, Rønsbo MH, Nielsen KF, Hounhouigan DJ, Jacobsen M (2011) Formation of cereulide and enterotoxins by *Bacillus cereus* in fermented African locust beans. Food Microbiol 28:1441–1447
30. de Medeiros LS, Murgu M, de Souza AQL, Rodrigues-Fo E (2011) Antimicrobial Depsides Produced by Cladosporium uredinicola, an Endophytic Fungus Isolated from Psidium guajava Fruits. Helv Chim Acta 94:1077–1084
31. Abarca ML, Accensi F, Bragulat MR, Castella G, Cabanes FJ (2003) *Aspergillus carbonarius* as the main source of ochratoxin A contamination in dried vine fruits from the Spanish market. J Food Prot 66: 504–506
32. Henrikson JC, Ellis TK, King JB, Cichewicz RH (2011) Reappraising the Structures and Distribution of Metabolites from Black Aspergilli Containing Uncommon 2-Benzyl-4H-pyran-4-one and 2-Benzylpyridin-4(1H)-one Systems. J Nat Prod 74: 1959–1964
33. Hasegawa Y, Fukuda T, Hagimori K, Tomoda H, Omura S (2007) Tensyuic acids, new antibiotics produced by *Aspergillus niger* FKI-2342. Chem Pharm Bull 55:1338–1341
34. Guthals A, Watrous JD, Dorrestein PC, Bandeira N (2012) The spectral networks paradigm in high throughput mass spectrometry. Mol Biosyst 8:2535–2544

Analytical and Bioanalytical Chemistry

Electronic Supplementary Material

Aggressive dereplication using UHPLC-DAD-QTOF: screening extracts for up to 3000 fungal secondary metabolites

Andreas Klitgaard, Anita Iversen, Mikael R. Andersen, Thomas O. Larsen, Jens Christian Frisvad, Kristian Fog Nielsen

Section 1. Construction of compound database

The database was constructed in ACD Chemfolder (Advanced Chemistry Development, Toronto, Canada) from: i) our in-house collection of reference standards (~1500 compounds) [1]; ii) compounds tentatively identified during the last 30 years (~500 compounds) [2-5]; iii) compound-peaks appearing in blank samples; iv) putative biosynthetic intermediates mainly from *A. niger* and *A. nidulans* and PKS pathways; and v) all compounds in AntiBase2012 which were listed as coming from: *Aspergillus, Fusarium, Trichoderma, Penicilium, Chaetomium, Stachybotrys, Alternaria* and *Cladosporium*, as well as their teleomorphic genera. Records of compounds reported from studies where the fungal culture was considered incorrectly identified were corrected, before addition to the compound database. When obtained from our own data or the literature, the full UV/VIS spectrum was linked to the record.

Many compounds were further registered to sub-genus level / species group level based on taxonomic data and chemotaxonomic studies [4-10]. In *Aspergillus* these were: *A. niger* complex; *A. nidulans* complex; and *A. fumigatus* complex. In *Fusarium* these were: Arthrosporiella (*F. incarnatum*); Discolor (*F. graminearum*); Elegans (*F. oxysporum*); Eupionnotes (*F. merismoides*); Gibbosum (*F. equiseti*); Lateritium (*F. lateritium*); Liseola (*F. verticillioides*); Martiella (*F. solani*); Roseum (*F. avenaceum*); and Sporotrichiella (*F. poae*).

From our work on metabolite profiling genera such as *Aspergillus, Fusarium, Penicillium, Alternaria* and *Cladosporium*, approximately 400 unknown compounds were added to the database as "unknowns" and registered via their elemental composition and from which species the compounds were detected.

For each compound the known or suspected major adducts, based on analysis of reference standards, were listed as: $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M+K]^+$, $[M+H+CH_3CN]^+$, $[M+Na+CH_3CN]^+$, $[M+H-H_2O]^+$, $[M+H-2H_2O]^+$, $[M+H-H_2]^+$(sterols), $[M+H-HCOOH]^+$, $[M+H-CH_3COOH]^+$, $[M+2H]^{2+}$, $[M+Na+H]^{2+}$ or $[M+2Na]^{2+}$ or "No ionization" in ESI^+, and in ESI^-: $[M-H]^-$, $[M-H+HCOOH]^-$, and $[M+Cl]^-$.

Creating search lists for Target Analysis (TA)

A Microsoft Excel application was created so the whole Chemfolder data-base (without structures) could be copied into one of the Excel sheets, and then sorted to include one or more genera, subspecies, known impurities, or all compounds with unknown retention time (RT). These data were transferred to a data search-list for TA containing: RT (if known), elemental composition and charge state of desired adduct, and name of compound.

For labelling of peaks in Bruker DataAnalysis 4.0 (DA) (Bruker Daltonics, Bremen, Germany), compounds that were available as reference standards were labelled "S-x" in front of the name where x is the reference standard number in our database. Compounds observed in sample blanks, were labelled "Bl-" in front of the name. Finally, compounds not tentatively identified were labelled as "Unknown"-"producing species"-number in the species, e.g. "Unknown-Aspergillus nidulans No. 3".

Automated screening of fungal samples

TA 1.2 (Bruker Daltonics, Bremen, Germany), was used to process data-files with the following typical parameters: A) retention time (if known) as ± 1.2 min (broad range), 0.8 min (medium range) and 0.3 min (narrow range); B) SigmaFit; broad 1000 (isotope fit not used), 40 as medium, and 20 as narrow range; and C) mass accuracy of the peak assessed at 4 ppm (broad range), 2.5 ppm (medium range), and 1.5 ppm (narrow range). Area cut off was set to 3000 counts as default, but was often adjusted in case of very concentrated or dilute samples.

The Software DA was used for manual comparison of all the extracted-ion-chromatograms (EIC), generated by TA, to the BPC chromatograms in order to identify non-detected major peaks.

Section 2. Aggressive dereplication (AD) of a *Penicillium melanoconidium* extract detects nearly all known compounds

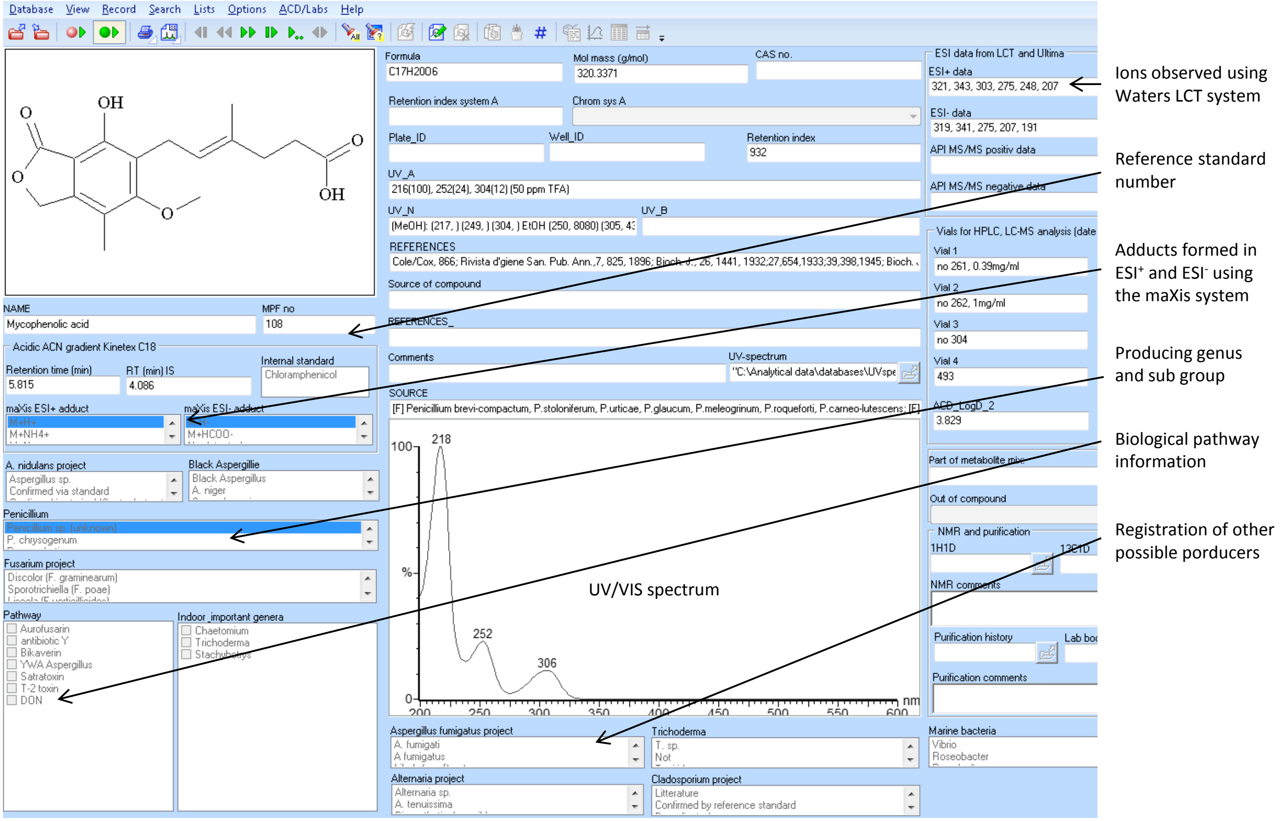
P. melanoconidium has formerly been reported to produce penitrem A, sclerotigenin, roquefortine C, meleagrin, oxaline, penicillic acid, verrucosidin and xanthomegnin, based on HPLC-DAD [40].

The extract was examined by the AD method searching for a subset of ~1700 *Penicillium* compounds and additional 700 compounds, and was found to produce a large number of secondary metabolites, see the figure (Fig. S5, Tables S1 and S2).

Previously detected metabolites along with additional families of secondary metabolites are listed in the Table S1 and the full search results list can be seen in the Table S2. Twenty five secondary metabolites could be assigned with a high degree of confidence. Chrysogine, 6-oxopiperidine-2-carboxylic acid, and 8-(methoxycarbonyl)-1-hydroxy-9-oxo-9H-xanthene-3-carboxylic acid were detected for the first time in *P. melanoconidium*, but been found in related *Penicillium* species [41;42]. Eight members of the roquefortine biosynthetic family (end products oxalines) were found, and also further confirmed by UV spectra and retention times. Concerning the penitrems, taxonomic and biosynthetic considerations, in connection with polarity and literature data, were used to verify the presence of penitrem A-F. Furthermore the UV spectrum and RT was the same for the authentic standard of penitrem A. Isomeric compounds of penitrem A such as pennigritrem and the acid hydrolysis products thomitrem A [43] could be excluded based on UV spectra different from that of penitrem A or because they were minor compounds (pennigritrem) as compared to the main product penitrem A [44;45]. PF1101A and B had the penitrem A UV spectrum which is different from the shearinine and janthitrems [46] and penitrems molecules were therefore much more likely candidates. Biosynthetic and taxonomic considerations also dictate that it must be the penitrems that are produced by *P. melanoconidium*.

The polyketides penicillic acid and verrucosidins were also found in *P. melanoconidium*. Verrucosidin had the same molecular formula as atranone A ($C_{24}H_{32}O_6$) [12], but the UV spectrum easily verified the right one. The finding of normethylverrucosidin and deoxyverrucosidin [47] also confirms that the verrucosidin biosynthetic family was produced by *P. melanoconidium,* which is likely as the closely related *P. polonicum* and *P. aurantiogriseum* also produce these [40]. A metabolite with the formula $C_{24}H_{32}O_4$ was annotated as 6-farnesyl-5,7-dihydroxcy-4-methylphthalide. However this metabolite has a mycophenolic acid chromophore, which has never been found in *P. melanoconidium*. The formula could be hypothesized to be a “dideoxyverrucosidin”, but this has to be confirmed.

Primary metabolites were few, and included: choline-O-sulfate, linoleic acid, phenylalanine and 1,2-dilininoyl-n-glycero-3-phosphocholine, which could be annotated based on reference standards. In conclusion several new families of compounds were which are highly toxic, especially the verrucosidins, but also chrysogine a compound often detected in cereal infecting fungi, e.g. *Fusarium*. Such information is valuable for future comparative genomics for revealing biosynthetic pathways.



NAME
Mycophenolic acid

MPF no
108

Acidic ACN gradient Kinetex C18

Retention time (min)
5.815

RT (min) IS
4.086

Internal standard
Chloramphenicol

maXis ESI+ adduct
M+H+
M+NH4+

maXis ESI- adduct
M-H-
M+HCOO-

A. nidulans project
Aspergillus sp.
Confirmed via standard

Black Aspergillie
Black Aspergillus
A. niger

Penicillium
Penicillium sp. (unknown)
P. chrysogenum

Fusarium project
Discolor (F. graminearum)
Sporotrichiella (F. poae)

Pathway
□ Aurofusarin
□ antibiotic Y
□ Bikaverin
□ YWA Aspergillus
□ Satratoxin
□ T-2 toxin
□ DON

Indoor _important genera
□ Chaetomium
□ Trichoderma
□ Stachybotrys

Formula
C17H20O6

Mol mass (g/mol)
320.3371

CAS no.

Retention index system A

Chrom sys A

Plate_ID

Well_ID

Retention index
932

UV_A
216(100), 252(24), 304(12) (50 ppm TFA)

UV_N
(MeOH): (217,) (249,) (304,) EtOH (250, 8080) (305, 4

UV_B

REFERENCES
Cole/Cox, 866; Rivista d'giene San. Pub. Ann.,7, 825, 1896; Bioch. J., 26, 1441, 1932;27,654,1933;39,398,1945; Bioch.

Source of compound

REFERENCES_

Comments

UV-spectrum
"C:\Analytical data\databases\UVspe

SOURCE
[F] Penicillium brevi-compactum, P.stoloniferum, P.urticae, P.glaucum, P.meleogrinum, P.roqueforti, P.carneo-lutescens; [F]

UV/VIS spectrum

Aspergillus fumigatus project
A. fumigati
A fumigatus

Trichoderma
T. sp.
Not

Alternaria project
Alternaria sp.
A. tenuissima

Cladosporium project
Litterature
Confirmed by reference standard

ESI data from LCT and Ultima

ESI+ data
321, 343, 303, 275, 248, 207

ESI- data
319, 341, 275, 207, 191

API MS/MS positiv data

API MS/MS negative data

Vials for HPLC, LC-MS analysis (date

Vial 1
no 261, 0.39mg/ml

Vial 2
no 262, 1mg/ml

Vial 3
no 304

Vial 4
493

ACD_LogD_2
3.829

Part of metabolite mix:

Out of compound

NMR and purification

1H1D

13C1D

NMR comments

Purification history

Lab boo

Purification comments

Marine bacteria
Vibrio
Roseobacter

Ions observed using Waters LCT system

Reference standard number

Adducts formed in ESI$^+$ and ESI$^-$ using the maXis system

Producing genus and sub group

Biological pathway information

Registration of other possible porducers

Fig. S1. Compound registration in the compound database

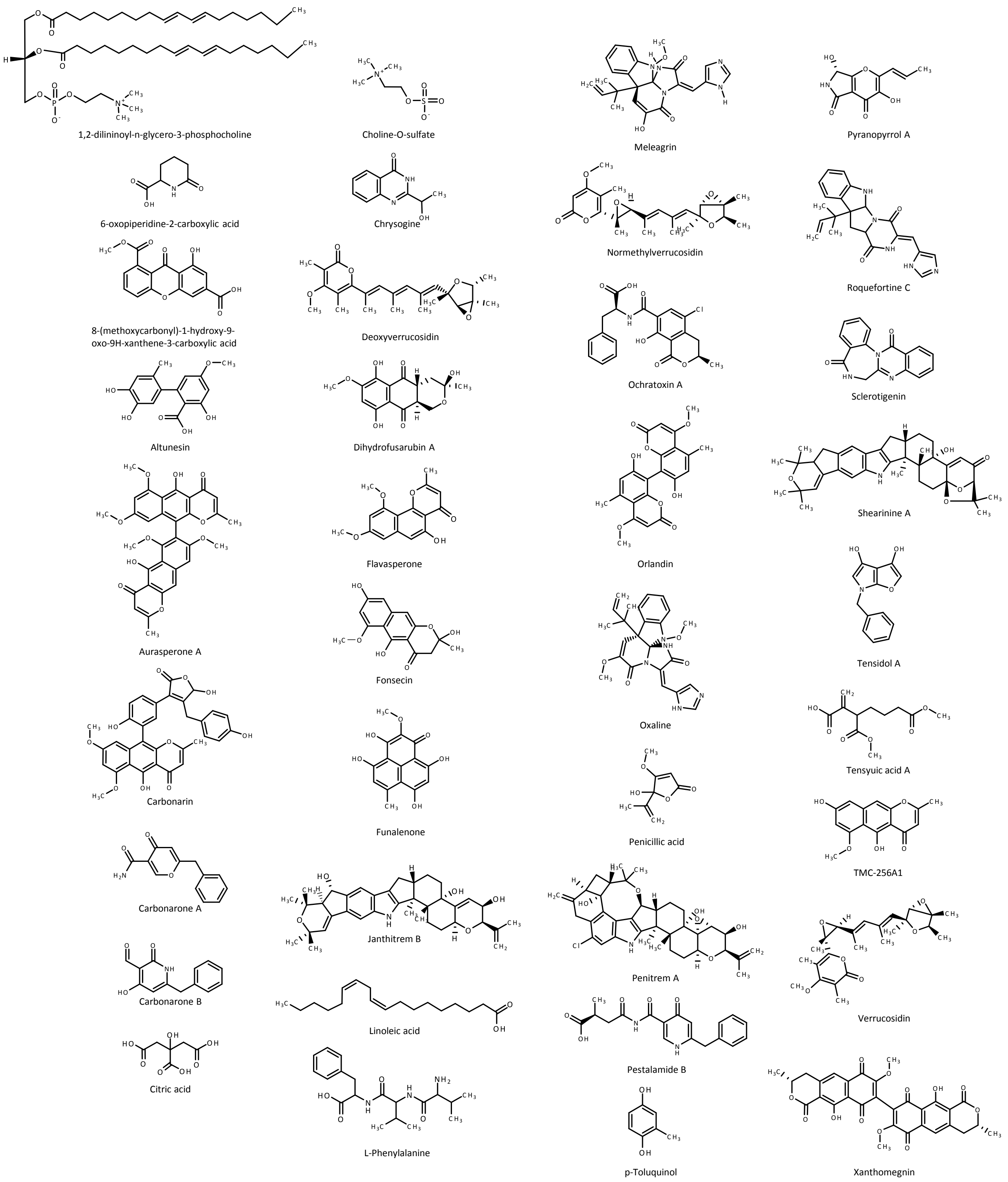


Fig. S2. Chemical structures of compounds mentioned in the text. The structures are shown in alphabetical order in columns from left to right. Only one example for each biosynthetic family is depicted

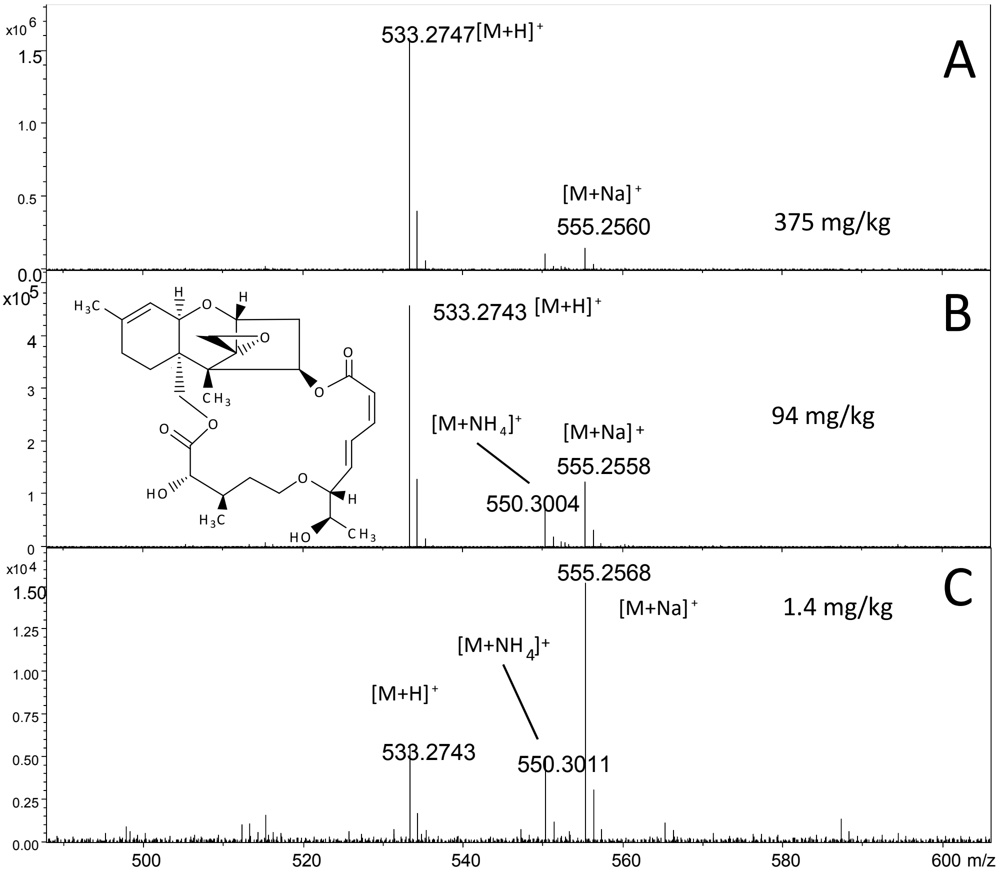


Fig. S3. ESI$^+$ spectrum of roridin A in crude extracts of *Baccharis megapotamica* spiked with (A) 375, (B) 94 and (C) 1.4 mg/kg roridin A

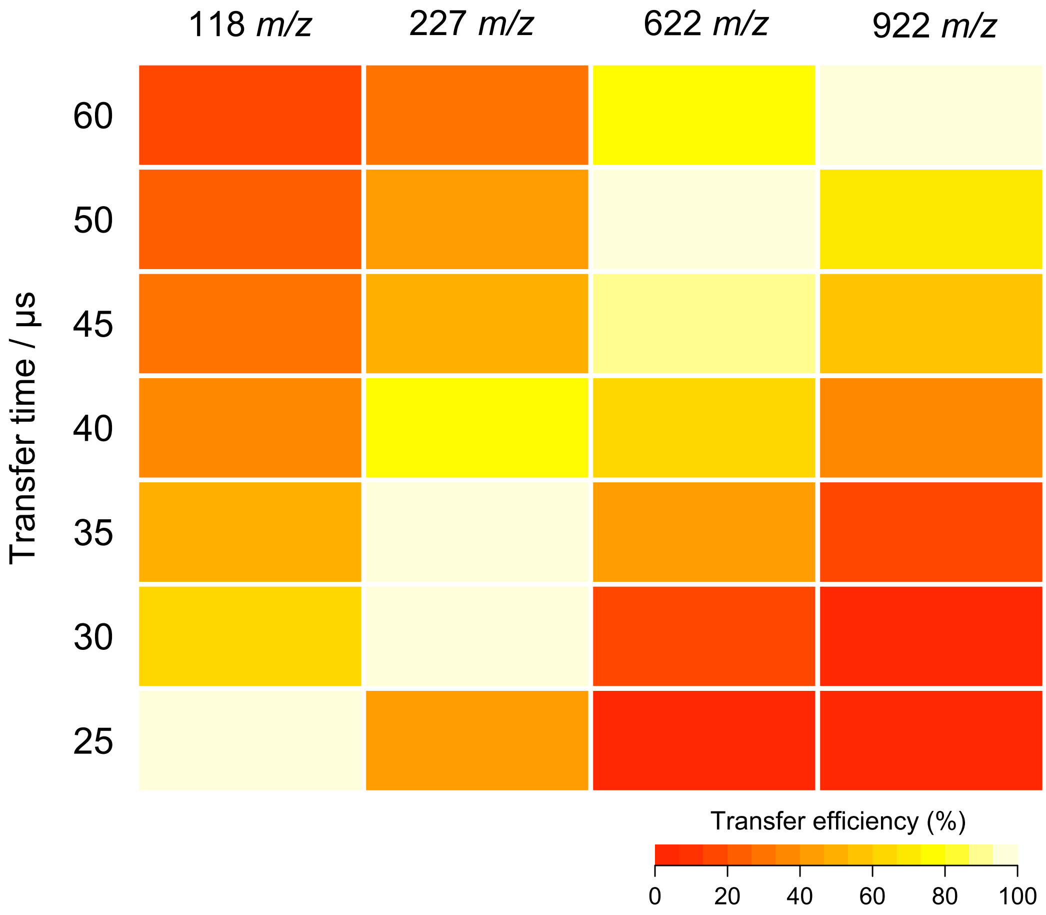


Fig. S4. Transfer efficiency (%) of selected ions from *m/z* 118-922 (relative to maximum)

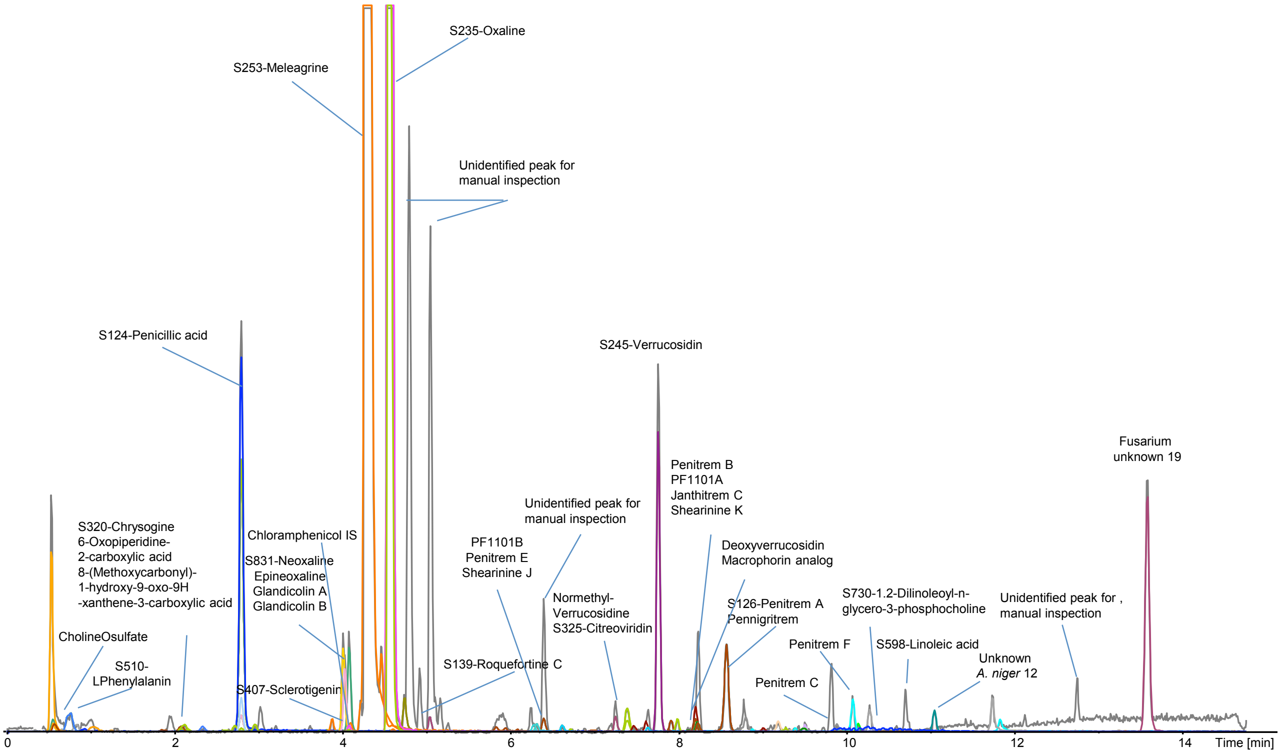


Fig. S5. Analyzed fungal extract from *Penicillium melanoconidium* (IBT 30549) cultivated on CYA media. The chromatogram is overlaid with EICs from detected compounds facilitating easy dereplication. The chromatogram has been scaled to better illustrate the presence of smaller peaks

Table S1. UHPLC-HRMS detection of secondary metabolites produced by *Penicillium melanoconidium* IBT 30549 grown on CYA agar for 7 days at 25°C in darkness

Biosynthetic family	Name of metabolite	Formula	Retention time (min.)
Chrysogines	Chrysogine	$C_{10}H_{10}N_2O_2$	2.337
Sclerotigenins	Sclerotigenin	$C_{16}H_{11}N_3O_2$	3.876
Roquefortines	Roquefortine C	$C_{22}H_{23}N_5O_2$	4.738
	Roquefortine F	$C_{23}H_{25}N_5O_3$	5.038
	E-3-H-Imidazol-4-yl-methylene-6-1H-indole-3-yl-methyl-2,5-piperazinedione	$C_{17}H_{15}N_5O_2$	1.038
	Glandicolin A	$C_{22}H_{21}N_5O_3$	4.092
	Glandicolin B	$C_{22}H_{21}N_5O_4$	4.008
	Meleagrin	$C_{23}H_{23}N_5O_4$	4.291
	Epi-Meleagrin)	$C_{23}H_{23}N_5O_4$	4.456
	Epi-Neoxaline	$C_{23}H_{25}N_5O_4$	4.028
	Oxaline	$C_{24}H_{25}N_5O_4$	4.560
Penitrems	Penitrem A	$C_{37}H_{44}ClNO_6$	8.563
	Penitrem B	$C_{37}H_{45}NO_5$	8.217
	Penitrem C	$C_{37}H_{44}ClNO_4$	9.876
	Penitrem D	$C_{37}H_{45}NO_4$	7.980
	Penitrem E	$C_{37}H_{45}NO_6$	6.613
	Penitrem F	$C_{37}H_{44}ClNO_5$	10.065
	Thomitrem A	$C_{37}H_{44}ClNO_6$	8.226
	PF1101A	$C_{37}H_{47}NO_4$	6.391
	(PF1101A-isomer)	$C_{37}H_{47}NO_4$	8.194
	PF1101B	$C_{37}H_{47}NO_6$	6.309
Penicillic acids	Penicillic acid	$C_8H_{10}O_4$	2.795
Verrucosidins [61]	Verrucosidin	$C_{24}H_{32}O_6$	7.752
	Normethylverrucosidin	$C_{23}H_{30}O_6$	7.245
	Deoxyverrucosidin	$C_{24}H_{32}O_5$	8.197
	Dideoxyverrucosidin	$C_{24}H_{32}O_4$	9.494
Unknown	8-(Methoxycarbonyl)-1-hydroxy-9-oxo-9H-xanthene-3-carboxylic acid	$C_{16}H_{10}O_7$	2.118
Unknown	Toluquinol	$C_7H_8O_2$	2.794
Primary metabolites	Cholin-O-sulfate	$C_5H_{13}NO_4S$	0.561
	Phenylalanine	$C_9H_{11}NO_2$	0.757
	1,2-dilininoyl-n-glycero-3-phosphocholine	$C_{44}H_{80}NO_8P$	10.237
	Linoleic acid	$C_{18}H_{32}O_2$	10.265

Table S2. Table S2 – AD of extract of *P. melanoconidium* grown on CYA agar (crude results)

Peak	Class	Comment	Compound Name	Mol.Formula	Err ppm	mSigma	Area	RT meassured	RT expected
A	+++		Unknown A nidulans no 37 Diana	C6H13NaO6	0.9	9	240479	0.54	0.64
B	+++		BL-UK Cla no 32 possible blank	C7H13NO2	0.4	19	19325	0.558	0.57
C	+		CholineOsulfate	C5H13NO4S1	0.1	25	12403	0.561	0.00
D	+++		BL-UK Cla no 60 possible blank	C10H13N5O4	0.4	32	14268	0.577	0.72
E	+		S510-LPhenylalanin	C9H11NO2	2.4	2	53072	0.757	
E	++		BL-UK Cla no 54 possible blank	C9H11NO2	2.4	2	53072	0.757	0.85
F	+	Detected for the first time in P. melanoconidium	6Oxopiperidine2carboxylic acid	C6H9NO3	1.2	12	16102	0.834	
G	+		E31HImidazol4ylmethylen61Hindol3ylmethyl2.5piperazindiol	C17H15N5O2	1	50	17266	1.038	
G	+		E31HImidazol4ylmethylene61Hindole3ylmethyl2.5piperazinediol	C17H15N5O2	1	50	17266	1.038	
H	+++		BL-UK Cla no 95 possible blank	C7H14N2O3	1.6	17	11833	2.084	2.10
H	+++		BL-UK Cla no 94 possible blank	C7H14N2O3	1.6	17	11833	2.084	1.91
I	+	Detected for the first time in P. melanoconidium	8Methoxycarbonyl1hydroxy9oxo9Hxanthene3carboxylic acid	C16H10O7	1.7	32	19042	2.118	
J	+++	Detected for the first time in *P. melanoconidium*	S320-Chrysogine	C10H10N2O2	1.5	28	10415	2.337	2.56
K	+	No, confused with toloquinol	2.3Dihydroxy toluene	C7H8O2	2.3	3	32022	2.794	
K	+		S297-Hydroquinone. methyl 6CI.8CI	C7H8O2	2.3	3	32022	2.794	
K	+		2Acetyl5methylfuran	C7H8O2	2.3	3	32022	2.794	
K	+		S502-3.5dihydrotoluen	C7H8O2	2.3	3	32022	2.794	
L	+		S124-Penicillic acid	C8H10O4	2.2	15	660761	2.795	
M	+		8betaHydroxy7oxocurvularin	C16H18O7	1.2	11	51254	2.796	
M	+		11aHydroxy12oxocurvularin	C16H18O7	1.2	11	51254	2.796	
M	+		S103-6Methylsalicylic acid	C8H8O3	0.6	15	484020	2.796	
M	+		S601-3hydroxy4methylbenzoic acid	C8H8O3	0.6	15	484020	2.796	
M	+		S621-2hydroxy3methoxybenzaldehyde	C8H8O3	0.6	15	484020	2.796	
M	+		S620-3hydroxy4methoxybenzaldehyde	C8H8O3	0.6	15	484020	2.796	
M	+		S570-pHydroxybenzoic acid methyl ester	C8H8O3	0.6	15	484020	2.796	
M	+		S616-12.6dihydroxyphenylethanone	C8H8O3	0.6	15	484020	2.796	
M	+		S499-3Methylsalicylic acid	C8H8O3	0.6	15	484020	2.796	
N	+++		BL-UK Cla no 11 possible blank	C11H18N2O2	0.1	21	12754	2.957	2.85
N	+++		BL-UK Cla no 12 possible blank	C11H18N2O2	0.1	21	12754	2.957	3.09
O	+++		S407-Sclerotigenin	C16H11N3O2	0.4	10	18700	3.876	3.88
P	+		Sorbicillactone B	C21H25NO8	2.2	6	124492	4.008	
P	+++		Glandicolin B	C22H21N5O4	0.9	20	123717	4.008	4.01
Q	+++		S831-Neoxaline	C23H25N5O4	0.5	18	86551	4.028	4.28
Q	+		epiNeoxaline	C23H25N5O4	0.5	18	86551	4.028	
R	+++	Internal standard	Chloramphenicol IS	C11H12Cl2N2O5	0.2	26	129956	4.078	4.12
S	+++		Glandicolin A	C22H21N5O3	0.1	4	14081	4.092	4.09
T	+++		S253-Meleagrin	C23H23N5O	1	24	1E+07	4.291	4.29

				4					
U	+++		S253-Meleagrin	C23H23N5O4	0	24	319359	4.456	4.29
V	++		UK Cla no 61	C20H32O11	1.1	54	2092310	4.56	5.33
V	+++		S235-Oxaline	C24H25N5O4	1.3	7	7560329	4.56	4.56
V	+		S470-Oxaline	C24H25N5O4	1.3	7	7560329	4.56	
X	+		S340-PF3	C22H23N5O2	0.4	7	59076	4.738	
X	+++		S139-Roquefortine C	C22H23N5O2	0.4	7	59076	4.738	4.97
X	+		S338-PF1	C22H23N5O2	0.4	7	59076	4.738	
Y	++		Fusarium solani unknown 15	C22H29N1O7	3.2	11	21556	5.038	5.52
Y	+++		Roquefortine F	C23H25N5O3	0	17	21760	5.038	5.04
Z	++		Unknown in A. niger 20	C28H36N4O5	2.4	23	10745	6.273	6.04
AA	+		PF1101B	C37H47NO6	1.6	50	12191	6.309	
AA	+	No, confused with penitrem-like compound	Shearinine J	C37H47NO6	1.6	50	12191	6.309	
AB	+	No, confused with penitrem-like compound	Shearinine K	C37H47NO4	1	19	19716	6.391	
AB	+		PF1101A	C37H47NO4	1	19	19716	6.391	
AB	+	No, confused with penitrem-like compound	Janthitrem C	C37H47NO4	1	19	19716	6.391	
AC	+		Thomitrem E	C37H45NO6	1.9	59	10690	6.613	
AC	+		S387-Penitremone A	C37H45NO6	1.9	59	10690	6.613	
AC	+	No, confused with penitrem-like compound	Shearinine D	C37H45NO6	1.9	59	10690	6.613	
AC	++		Penitrem E	C37H45NO6	1.9	59	10690	6.613	6.61
AD	+++		Normethylverrucosidine	C23H30O6	0.6	12	21959	7.245	7.25
AD	+		S37-Citreoviridin	C23H30O6	0.6	12	21959	7.245	
AD	+		S325-Citreoviridin	C23H30O6	0.6	12	21959	7.245	
AE	+		IsocitreohybridoneB	C29H38O8	2.5	22	44729	7.383	
AE	+		Citreohybridone B	C29H38O8	2.5	22	44729	7.383	
AF	+++		Unknown in A. niger 21	C27H40O8	1.9	29	20444	7.384	7.49
AG	+++		Unknown A carbonarius no 9	C29H41N7O2	0.9	16	14010	7.608	7.69
AH	++	No, confused with verrucosidin	S452-AtranoneA	C24H32O6	0.5	34	439654	7.752	7.23
AH	+++		S245-Verrucosidin	C24H32O6	0.5	34	439654	7.752	7.75
AI	++		Fusarium solani unknown 11	C18H31NaO4	0.4	21	17273	7.904	7.29
AI	++		Fusarium solani unknown 10	C18H31NaO4	0.4	21	17273	7.904	7.16
AJ	+++		Penitrem D	C37H45NO4	1.1	14	19122	7.98	7.98
AK	+	No, confused with penitrem-like compound	Shearinine K	C37H47NO4	1.4	20	18744	8.194	
AK	+		PF1101A	C37H47NO4	1.4	20	18744	8.194	
AK	+	No, confused with penitrem-like compound	Janthitrem C	C37H47NO4	1.4	20	18744	8.194	
AL	+		Macrophorin analog	C24H32O5	1	8	38721	8.197	
AL	+++		Deoxyverrucosidin	C24H32O5	1	8	38721	8.197	8.20
AM	+++		Penitrem B	C37H45NO5	1.9	36	15171	8.217	8.22
AM	+		Shearinine F	C37H45NO5	1.9	36	15171	8.217	
AM	+		Penitremone C	C37H45NO5	1.9	36	15171	8.217	
AM	+		ShearinineA	C37H45NO5	1.9	36	15171	8.217	

AN	+	No, confused with penitrem A	Pennigritrem	C37H44ClN1O6	2	26	12720	8.226	
AN	+	No, confused with penitrem A	Thomitrem A	C37H44ClN1O6	2	26	12720	8.226	
AN	++	No, confused with penitrem A	S126-Penitrem A	C37H44ClN1O6	2	26	12720	8.226	8.56
AO	+		Pennigritrem	C37H44ClN1O6	2.1	42	172731	8.563	
AO	+		Thomitrem A	C37H44ClN1O6	2.1	42	172731	8.563	
AO	++		S126-Penitrem A	C37H44ClN1O6	2.1	42	172731	8.563	8.56
AP	++		Unknown in A. niger 18	C16H21NaO4	3.4	14	31776	8.794	8.85
AQ	+++		Unknown in A. niger 24	C28H42	1.8	23	18486	9.179	8.93
AR	+		6-Farnesyl-5,7-dihydroxy-4-methylphthalide	C24H32O4	2.9	49	10051	9.494	
AS	+++		Penitrem C	C37H44ClNO4	0.1	33	11419	9.876	9.88
AT	+++		Penitrem F	C37H44ClNO5	1.6	25	53231	10.07	10.07
AU	++		Unknown A nidulans no 36 Diana	C19H37NaO4	1.3	37	12976	10.13	9.67
AV	+++		S730-1,2-Dilinoleoyl-sn-glycero-3-phosphocholine	C44H80NO8P	0.2	17	24565	10.24	10.24
AX	+++		S598-Linoleic acid	C18H32O2	1.5	7	38041	10.27	10.17
AY	+++		Unknown in A. niger 12	C21H41NaO4	0.5	9	35385	11.04	11.04
AZ	+++		Fusarium solani unknown 2	C24H37NaO4	0.1	10	62599	11.73	11.67
AAA	+++		BL-UK Cla no 83 possible blank	C22H43NO	0.5	10	28484	11.82	11.84
AAB	++		Fusarium unknown 19	C27H21N2NaO9	1.5	82	470182	13.57	13.49

mSigma: Fit of isotope pattern, see text for more. RT Retention time (min).

References

1. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO (2011) Dereplication of microbial natural products by LC-DAD-TOFMS. J Nat Prod 74:2338-2348

2. Rank C, Klejnstrup ML, Petersen LM, Kildgaard S, Frisvad JC, Godtfredsen CH, Larsen TO (2012) Comparative chemistry of *Aspergillus oryzae* (RIB40) and *A. flavus* (NRRL 3357). Metabolites 2:39-56

3. Månsson M, Phipps RK, Gram L, Munro MH, Larsen TO, Nielsen KF (2010) Explorative Solid-Phase Extraction (E-SPE) for Accelerated Microbial Natural Product Discovery, Dereplication, and Purification. J Nat Prod 73:1126-1132

4. Nielsen KF, Mogensen JM, Johansen M, Larsen TO, Frisvad JC (2009) Review of secondary metabolites and mycotoxins from the *Aspergillus niger* group. Anal Bioanal Chem 395:1225-1242

5. Frisvad JC, Rank C, Nielsen KF, Larsen TO (2009) Metabolomics of *Aspergillus fumigatus*. Med Mycol 47:S71

6. Rank C, Nielsen KF, Larsen TO, Varga J, Samson RA, Frisvad JC (2011) Distribution of sterigmatocystin in filamentous fungi. Fungal Biology 115:406-420

7. Frisvad JC, Andersen B, Thrane U (2008) The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. Mycol Res 112:231-240

8. Andersen B, Sørensen JL, Nielsen KF, van den Ende B, de Hoog S (2009) A polyphasic approach to the taxonomy of the *Alternaria infectoria* species-group. Fungal Genet Biol 46:642-656

9. Andersen B, Dongo A, Pryor BM (2008) Secondary metabolite profiling of Alternaria dauci, A. porri, A. solani, and A. tomatophila. Mycol Res 112:241-250

10. Frisvad JC, Smedsgaard J, Larsen TO, Samson RA (2004) Mycotoxins, drugs and other extrolites produced by species in *Penicillium* subgenus *Penicillium*. Stud Mycol 49:201-241

6.2 Paper 2 – Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library

Kildgaard, S., Mansson, M., Dosen, I., Klitgaard, A., Frisvad, J. C., Larsen, T. O., & Nielsen, K. F.

Paper accepted in Marine Drugs (2014)

Mar. Drugs **2014**, *12*, 3681-3705; doi:10.3390/md12063681

OPEN ACCESS

marine drugs

ISSN 1660-3397

www.mdpi.com/journal/marinedrugs

Article

Accurate Dereplication of Bioactive Secondary Metabolites from Marine-Derived Fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library

Sara Kildgaard, Maria Mansson, Ina Dosen, Andreas Klitgaard, Jens C. Frisvad, Thomas O. Larsen and Kristian F. Nielsen *

Department of Systems Biology, Technical University of Denmark, Soeltofts Plads 221,
Kgs. Lyngby DK-2800, Denmark; E-Mails: sarki@bio.dtu.dk (S.K.); maj@bio.dtu.dk (M.M.);
idos@bio.dtu.dk (I.D.); ankl@bio.dtu.dk (A.K.); jcf@bio.dtu.dk (J.C.F.); tol@bio.dtu.dk (T.O.L.)

***** Author to whom correspondence should be addressed; E-Mail: kfn@bio.dtu.dk;
Tel.: +45-4525-2602.

Received: 28 February 2014; in revised form: 23 May 2014 / Accepted: 11 June 2014 / Published: 20 June 2014

Abstract: In drug discovery, reliable and fast dereplication of known compounds is essential for identification of novel bioactive compounds. Here, we show an integrated approach using ultra-high performance liquid chromatography-diode array detection-quadrupole time of flight mass spectrometry (UHPLC-DAD-QTOFMS) providing both accurate mass full-scan mass spectrometry (MS) and tandem high resolution MS (MS/HRMS) data. The methodology was demonstrated on compounds from bioactive marine-derived strains of *Aspergillus*, *Penicillium*, and *Emericellopsis*, including small polyketides, non-ribosomal peptides, terpenes, and meroterpenoids. The MS/HRMS data were then searched against an in-house MS/HRMS library of ~1300 compounds for unambiguous identification. The full scan MS data was used for dereplication of compounds not in the MS/HRMS library, combined with ultraviolet/visual (UV/Vis) and MS/HRMS data for faster exclusion of database search results. This led to the identification of four novel isomers of the known anticancer compound, asperphenamate. Except for very low intensity peaks, no false negatives were found using the MS/HRMS approach, which proved to be robust against poor data quality caused by system overload or loss of lock-mass. Only for small polyketides, like patulin, were both retention time and UV/Vis spectra necessary for unambiguous identification. For the ophiobolin family with many structurally similar analogues partly co-eluting, the peaks could be assigned correctly by combining MS/HRMS data and *m*/*z* of the $[M + Na]^+$ ions.

Keywords: marine-derived; MS/MS; dereplication; library; peptaibiotics; metabolomics

1. Introduction

Due to the cosmopolitan occurrence of many bioactive compounds, most natural product extracts contain compounds that have previously been characterized, despite intelligent selection of new organisms. This is of particular importance in primary screens where the target is usually non-selective, which inevitably leads to a high rediscovery rate of generally toxic compounds [1,2].

Microorganisms from the marine environment are a promising source of new bioactive compounds based on new chemical scaffolds [3–5], with the majority of known compounds originating from bacterial species such as *Salinospora* [6], *Pseudoalteromonas* [7,8], and *Vibrio* [9]. However, the subject of marine fungi is of much debate as most marine isolates have been found in mangrove and intertidal zones [4,10,11], rather than in true marine habitats; thus, no strict definition of "true marine fungi" currently exists [12]. Nonetheless, marine-derived fungal strains have yielded a plethora of biologically active compounds [5,13], with isolates of *Penicillium* and *Aspergillus* as the most common sources. These have mainly been isolated from substrates such as driftwood [14] and macroalgae [15], but also in deep-sediments [3,16,17]. *Aspergillus sydowii* is probably the most well-known example, identified as the cause of sea fan disease [18], but also the source of bioactive compounds [19]. It remains obscure whether these represent true marine isolates or just opportunistic strains that have adapted to the marine conditions [12]. From a drug discovery perspective, this might be of less importance, if the opportunistic strains produce different bioactive compounds than their terrestrial counterparts.

Several approaches to the dereplication process exist; for fast screening of extracts the aggressive dereplication approach can be very efficient [20]. This approach is based on accurate mass, isotopic patterns, and preferably selective adducts used for large batch searches of possible metabolites (up to 3000 compounds), e.g., based on all compounds described by a single genus. Yet, it returns false positives that need to be sorted away. The approach is currently not suited for organisms with limited taxonomic information. False positives can be circumvented by adding tandem MS with accurate mass determination of fragment ions (MS/HRMS) which can be automatically co-acquired using auto-MS/HRMS experiments (data-dependent acquisition of MS/HRMS spectra) [21]. This can now be achieved on both time-of-fight (TOFMS) and fourier transform (FTMS) mass spectrometers as well as Orbitrap and Q-Exactive instruments [22–25]. To achieve high quality MS/MS spectra, Agilent Technologies have chosen to acquire spectra at three different fragmentation energies, 10, 20 and 40 eV, as this often provides significant higher quality than e.g., a ramped spectrum from 10 to 40 eV [26]. The acquired MS/HRMS data can then be matches with the possible candidates using *in silico* fragmentation tools that can sort out poor matches [27,28].

For fast tentative identification of natural products, an automatic MS/HRMS spectral library search would be very efficient, if suitable natural products libraries existed. However, Massbank [29] and Metlin metabolomics library [30] (~10,000 compounds with spectra) only contain few microbial natural products. The current status will persist until it is required to publish MS/MS data with novel structures, for which there are now public depositories such as MetLin, Massbank and/or Global

Natural Products Social Molecular Networking (GnPS) [31] in the making at time of writing). Nevertheless, a major barrier is that MS/MS spectra of small molecules are inconsistent between instruments, in particular between ion-trap and collision cell-based instruments [32]. Also, compared to fragmentation of linear peptides [33] and lipids [34], fragmentation of natural products are much less predictable, since they often contain more condensed and highly complex ring systems: In consequence *in silico* predictors cannot predict a fragmentation spectrum, but to some extent, verify some fragments from a structure in a spectrum [27,28].

For smaller natural products libraries, different algorithms have been used to search MS/MS spectra for the tentative identification (absolute identification always requires a nuclear magnetic resonance (NMR) validated reference standard). Fredenhagen *et al.* [35] searched low resolution MS/MS data with the National Institute of Standards and Technology (NIST) algorithm developed for full scan EI^+ spectra and the Mass Frontier software for MS^n spectra and found the latter to be superior. El-Elimat *et al.* [2] used ACD-IntelliXtract that also includes accurate mass of the fragments, but does not use the parent ion data as search entry. A comprehensive review on algorithms can be found in Hufsky *et al.* [28]. Recently, a networking MS/MS strategy has been published from the Dorrestein/ Bandeira labs [36,37], where MS/MS spectra are compared pairwise to yield clusters of structurally related compounds. However, back integration/deconvolution of raw data to find corresponding full scan data and linking MS/MS spectra of adducts belonging to the same molecular feature as well as retention time still needs to be done manually and is thus very time consuming.

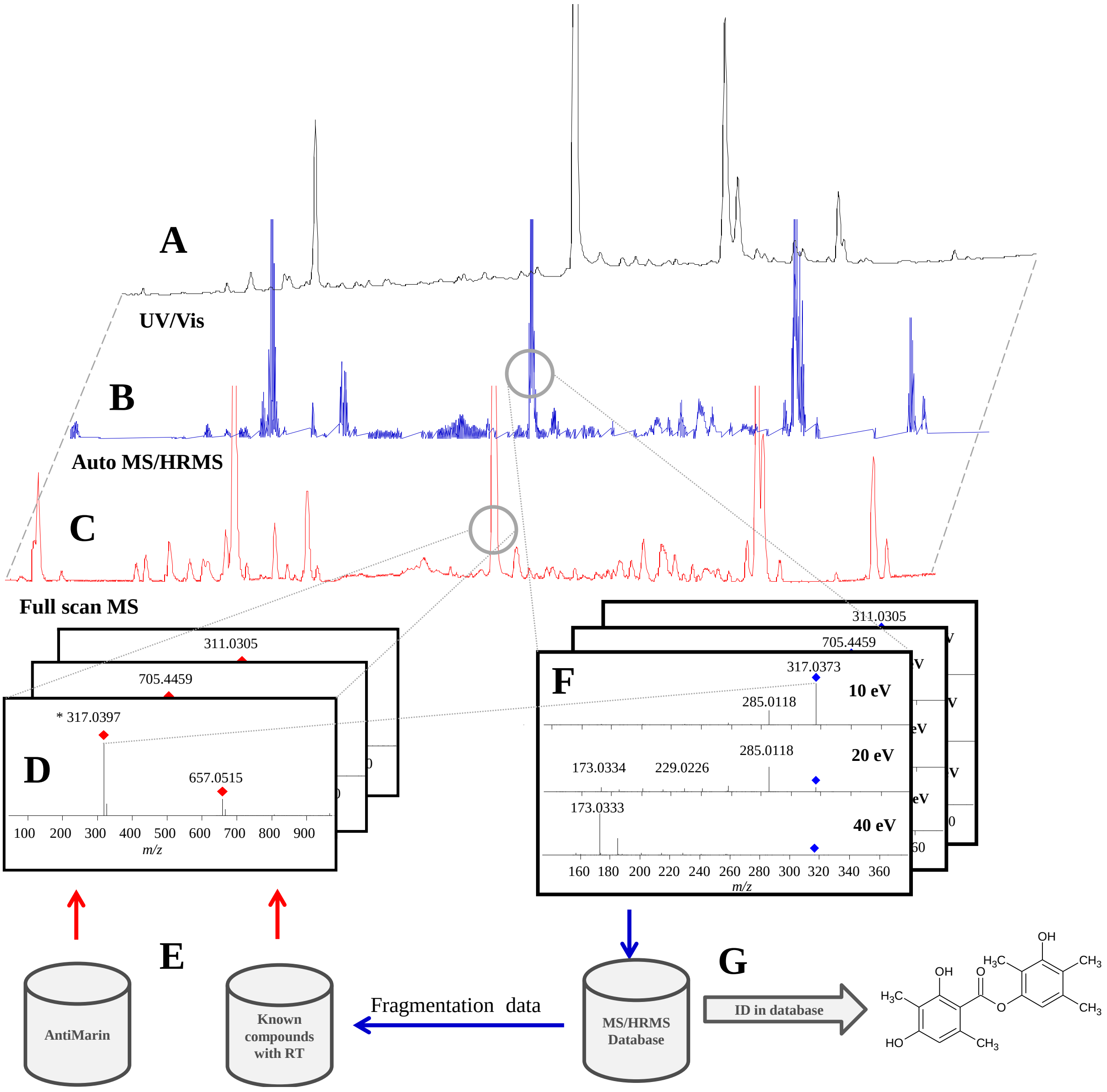
In this current study, we demonstrate the use of our MS/HRMS library search to dereplicate known compounds in bioactive extracts from marine-derived *Aspergillus*, *Penicillium*, and *Emericellopsis* strains. Extracts were selected from a screening conducted as a part of the PharmaSea project [38].

Ultra-high performance liquid chromatography-diode array detection-quadrupole time of flight mass spectrometry (UHPLC-DAD-HRMS) with auto- tandem high resolution mass spectrometry (MS/HRMS) analysis was used to screen the extracts and subsequently, MS/HRMS data was matched against a newly constructed library of 1300 compounds (10, 20, and 40 eV spectra) using the Agilent search algorithm. This algorithm is an integral part of the Agilent MassHunter software, which can subtract background and merge spectra over a chromatographic peak into a single spectrum prior to automatic search against the library. To assess the limitations and inherent bias of the library, we compare the results with the aggressive dereplication approach [20] based on accurate mass, isotope pattern, and lists of taxonomically relevant compounds. Specificity is tested on a number of small polar analytes, showing the importance of including retention time and appropriate search parameters for compounds with less characteristic spectra. Finally, comparison with UV/Vis detection was done for a number of poorly ionizing compounds showing the value of this additional cheap detector.

2. Results and Discussion

Figure 1 illustrates the overall screening concept used in this study, where UHPLC-DAD-QTOF data are analyzed in three different ways: (i) MS/HRMS data searched directly in MS/HRMS library; (ii) aggressive dereplication of the full scan HRMS data using search lists of known compounds; (iii) UV/Vis detection for poorly ionizing compounds. Finally, an unbiased peak-picking algorithm was used to highlight completely novel compounds. For dereplication of previously described compounds and novel isomers, all four approaches were combined as illustrated in the examples of *Penicillium bialowiezense* (Section 2.2.1) and *Aspergillus insuetus* (Section 2.2.2). Specificity problems with MS/HRMS searching are illustrated for patulin and compounds with the same elemental composition (Section 2.1.5).

Figure 1. Overview of the screening setup where ultra-high performance liquid chromatography (UHPLC) with three detection methods is used. (**A**) ultraviolet/visual (UV/Vis) for poorly ionizing compounds; (**C**,**D**) full scan high resolution mass spectrometry (HRMS) screening; (**B**,**F**) MS/HRMS identification using the MS/HRMS library (**G**). Elemental compositions from compounds known from literature and previous studies were searched for in the full scan data (**E**,**D**).



2.1. Data Acquisition and Library Creation

2.1.1. Chromatographic Separation

The gradient was developed to provide the highest peak capacity in extracts from *Aspergillus niger* and *A. nidulans* with emphasis on not losing polar alkaloids (e.g., pyranonigrins and nigragillins) and small organic acids. This led to the use of the more polar phenyl-hexyl Poroshell column (compared to C_{18}) as well as a low start of the gradient (10% acetonitrile). This retains highly polar compounds such as patulin and type B trichothecenes slightly better than C_{18}. However, the long column required a longer gradient and equilibration time leading to half the productivity, but better opportunities for more MS/MS experiments. The high temperature of 60 °C was needed in order to keep the back pressure below the limit of the 2.7 μm Poroshell column. The method yielded an excellent peak distribution and narrow peak width compared to other methods [2], which allowed for higher quality spectra of most compounds in an extract. Injection volume had to be kept low (1 μL) to avoid peak broadening of polar peaks as samples were dissolved in methanol. However, in some projects less had to be injected (as little as 0.1 μL) as strongly ionizing compounds in high concentration resulted in broad peaks due to peak broadening in the ion-source which was further enhanced by the limited linearity of the time of flight (TOF) detector.

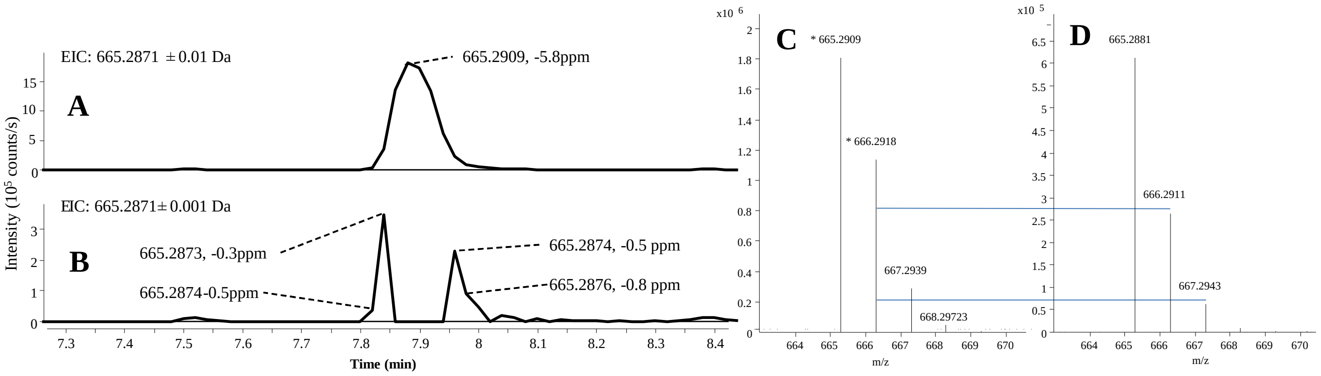
2.1.2. Mass Accuracy and Isotopic Ratio

Currently, time of flight mass spectrometry (TOFMS) and fourier transform mass spectrometry (FTMS) instruments provide similar mass accuracy when using a lock mass, but the TOFMS instruments still have problems with detector overload [39,40] as illustrated in Figure 2, where the mass accuracy and isotope ratio is compared between overloaded and non-overloaded parts of a chromatographic peak. As high intensity peaks are unavoidable, MassHunter was set to handle this by using only non-overloaded MS scans from the front and end of the chromatographic peaks during the peak picking and integration, similar to other TOFMS manufacturers like Waters. Currently, this cannot be handled by any third-party software like ACD-IntelliXtract or open source software like XCMS and MZmine.

On the up-side, quadruple time of flight mass spectrometry (QTOFMS) instruments have a much higher scan frequency of both full scan and MS/HRMS scans without losing resolution as is the case on the FTMS instruments (resolution proportional to scan time). When not using overloaded ion clusters (Figure 2) our data provided isotopic ratios $< \pm 2\%$ relative to the theoretical distribution as also observed elsewhere [20] while for Orbitrap data it might be as much as ±35% [24]. Since an accurate isotope ratio is the most efficient way to differentiate candidate elemental compositions within the instrument accuracy [41], the QTOFMS instruments are superior to the FTMS instruments in this point.

In some samples, high intensity peaks suppressed the lock mass ions in certain scans, resulting in up to 100 ppm mass error in cases where the instrument had not been tuned and calibrated for several days. Since MassHunter cannot automatically find scans with intact lock mass in other places in the data file, one needs to be aware of this problem to manually correct it if needed. Here, the MS/HRMS library still identified the correct compounds, underlining how this approach is very robust against mass errors from over-loaded peaks.

Figure 2. Ultra high performance liquid chromatography-electrospray ionization extracted ion chromatograms (UHPLC-ESI$^+$ EIC) of asperazine $[M + H]^+$ in an extract from *Aspergillus tubingensis*, showing the excellent mass accuracy until saturation in the peak apex. (**A**) EIC at ±0.01 Da; (**B**) EIC at ±0.001 Da; (**C**) spectrum at peak apex; and (**D**) spectrum at a non-saturated part of the peak. The vertical lines between **C** and **D** indicate the theoretical isotopic abundance of the A + 1 and A + 2 isotopomers.



2.1.3. Precursor Selection

A major challenge when using liquid chromatography-mass spectrometry (LC-MS)/MS libraries is the reproducibility of fragmentation patterns between different instruments and different instrument manufactures [42]. A major goal for the establishment of this library was to minimize variability due to changes in mobile phase composition and ion-source settings. This was done by including not only MS/HRMS from $[M + H]^+$, but also from the other predominant pseudo molecular ions such as $[M + Na]^+$ and $[M + NH_4]^+$ (for intensities >50% of $[M + H]^+$). Likewise, were MS/HRMS spectra of $[M + H - (H_2O)_n]^+$, $[M + H - HCOOH]^+$, and $[M + H - CH_3COOH]^+$ ions included when the full scan signal(s) were more intense than $[M + H]^+$, similar to Fredenshagen *et al.* [35]. When fragmentation of $[M + Na]^+$ only resulted in the loss of Na^+ to give the neutral molecule, the search algorithm gave false positives from any ion at the right *m*/*z*. MS/HRMS data from the stable $[M + Na]^+$ was therefore only included when resulting in specific fragments (~50% of the cases). Still, *m*/*z* of $[M + Na]^+$ is important for correct mass assignment of fragile molecules, where $[M + H]^+$ is not present due to spontaneous losses. Furthermore, in cases where in-source fragmentation of Compound A coincidentally results in production of Compound B also present in the library, the *m*/*z* of $[M + Na]^+$ can assist in correct assignment, as demonstrated.

In the negative ionization mode, $[M - H]^-$ is most often the dominant ion detected [43] while the formation of $[M - HCOO]^-$ (if formate is used as a buffer) seems to be very interface dependent [20], but very important for molecules not containing any acidic protons, and it was included when more than 50% of $[M - H]^-$ occurred. This resulted in the detection of highly active compounds like Type A and C trichothecenes, patulin, and aphidicolins not detected in other studies [2,35].

Part of the library (277 compounds) is available in PCDL format for download from the homepage of the Technical University of Denmark [44].

2.1.4. Fragmentation

In order to compensate for the high variation in energy needed to fragment natural products, the library was based on three distinct fragmentation energies (10, 20, and 40 eV) unlike existing microbial MS/MS libraries that are based on a single, fixed energy [2,35]. The high energy of 40 eV is needed to fragment larger, more stable molecules, while 10 and 20 eV are more gentle settings for smaller, more fragile molecules. This combination of energies also meant that the forensic science [26] and the Metlin libraries [35] which are also available for the MassHunter could be directly used. The latter, in particular, contains many lipids, prostaglandins, intracellular primary metabolites, small aromatics, amino acids, vitamins, *etc.*, which are also produced by fungi. The only cases where insufficient fragmentation was observed for all three energies were fusigen, SMTP-7 and 8, where only low intensity losses of formate and one other ion were observed in ESI$^+$. Thus, projects analyzing compounds with masses above 1000 Da should include additional fragmentation energies of e.g., 60–80 eV, as large single charged molecules are less disposed to fragment on the collisions with N_2. This is mainly due to simple energy kinetics (E_{kin} = ½ × m/z × v^2) where the ion-velocity in the collision cell is proportional to the square root of the mass.

Small (<200 Da) aromatic acids, pyrones, and lactones will statistically have less specific fragmentation reactions, which is observed in practice as loss of H_2O, HCOOH, and CO_2 [43]. Combined with an increase in the number of natural products with the same mass with decreasing mass (down to 220 Da) there is a double bias towards poor specificity of MS/MS of low mass compounds [43].

2.1.5. Library Scoring

Searching MS/HRMS spectra against the MS/HRMS library in MassHunter allows for three types of scorings: (i) using the parent mass and forward scoring that matches peaks in the unknown spectrum against the library spectrum; (ii) using parent mass and reverse scoring that matches peaks in the library spectrum against the unknown spectrum [28]; (iii) using reverse scoring but not the parent mass, called similarity, for finding compounds sharing fragment ions but having different molecular masses.

The pitfalls of scoring can be illustrated with patulin, a bioactive "nuisance" compound that is widely distributed in fungi that cause interference in many types of bioassays [45–47]. Patulin was identified in ESI$^-$ in marine-derived strains of *Penicillium antarcticum* (Figure 3). Patulin shares the same elemental composition ($C_7H_6O_4$) with six other compounds included in the library (Table 1), which all to a certain extent exhibited similar fragmentation patterns under the same CID condition. Using reverse and forward scoring, all library spectra belonging to compounds with the same elemental composition are in the matching pool. For reverse scoring there is an increased risk of wrong compound identification compared to forward scoring as the search algorithm in this case only looks for peaks present in the library spectra, disregarding peaks present in the unknown spectrum that are not present in the library spectra.

As seen in Figure 3, patulin and 2,3-dihydroxybenzoic acid had a similar ratio of the *m*/*z* 109.0287 fragment ion corresponding to the loss of CO_2 (CID 10 eV). 2,3-dihydroxybenzoic acid does not show any additional peaks in the 10 eV spectrum while patulin produces several. Reverse scoring only matched the two shared peaks in the unknown spectrum, resulting in 2,3-dihydroxybenzoic acid as the best match, while forward scoring, where all peaks in the spectrum are matched with the library spectrum, yielded patulin as the best match (Figure 3). The identification was verified by an authentic standard of patulin matching full scan MS, MS/HRMS, retention time and the UV spectrum where the slow slope from 200 to 240 nm prior to the main absorption at 276 nm. Deconvolution of all ions in the patulin full scan spectrum showed that it was not a false positive detection due to two or more co-eluting compounds.

Figure 3. UV/Vis spectrum (**A**) and MS/HRMS spectrum at 10 eV (ESI$^-$); (**D**) of unknown peak identified as patulin, compared to the patulin reference standard (**B**,**E**); and 2,3-dihydroxybenzoic acid (**C**,**F**). Identity was confirmed by correct retention time.

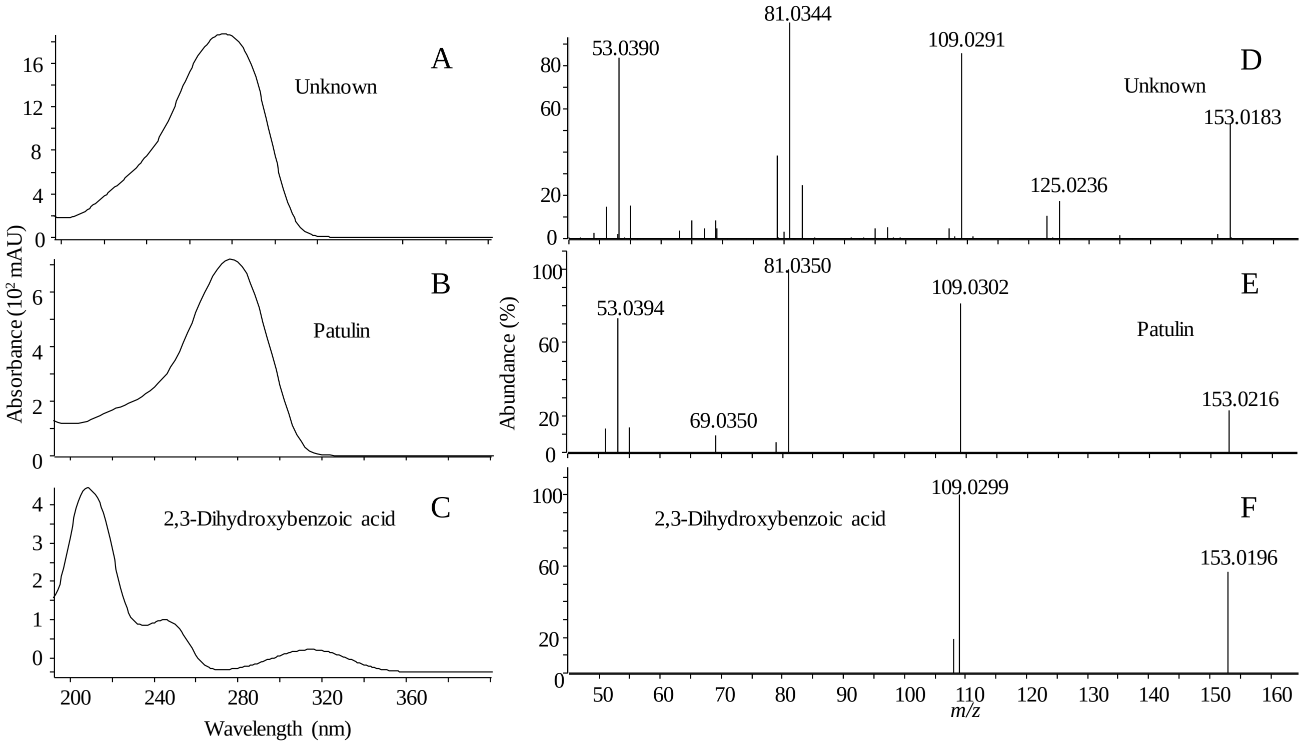


Table 1. Comparison of MS/HRMS spectra of all $C_7H_6O_4$ compounds in the MS/HRMS database against each other using forward and reverse scoring.

Name	RT (min)	Compound	Forward/Reverse Scoring (%)						
			1	2	3	4	5	6	7
Patulin	3.15	1	100	28/50	20/62	27/65	29/90	28/87	25/52
2,3-dihydroxybenzoic acid	3.85	2	28/32	100	60/68	63/71	97/90	76/86	0/0
2,4-dihydroxybenzoic acid	3.74	3	20/29	60/86	100	86/86	55/78	88/88	6/14
2,6-dihydroxybenzoic acid	3.87	4	21/29	63/86	86/98	100	58/79	80/92	6/14
3,4-dihydroxybenzoic acid	2.80	5	29/33	97/97	55/61	58/64	100	78/87	0/0
3,5-dihydroxybenzoic acid	2.63	6	29/33	97/97	55/61	58/64	78/87	100	0/0
Terreic acid	3.99	7	25/44	0/0	6/39	6/38	0/0	0/0	100

Inevitably, an unknown spectrum will contain more noise from co-eluting compounds compared to the library spectrum, another reason why reverse scoring is also valuable. This underlines the importance of using both forward and reverse scoring when evaluating matches from library searches in order to get the correct identification, thus multiple search types are recommended (e.g., using a minimal score of 50% for forward and 70% for reverse). It is further demonstrated that there is a need for orthogonal data like UV/Vis for dereplication of certain compound classes.

2.2. Dereplication of Marine-Derived Fungi

Fifteen marine-derived strains from different species belonging to *Penicillium*, *Aspergillus*, and *Emericellopsis* were fractionated and screened for their anti-microbial [48], anti-inflammatory [49], central nervous system (CNS) [50], and anticancer activity (unpublished assay based on glioblastoma stem cells), that resulted in 35 active fractions to be evaluated for their chemistry. Here, we present three of those as cases to illustrate the advantages and challenges using a MS/HRMS library for screening and dereplicating active fractions during a screening campaign. Analyzing the data file for MS/HRMS data, including peak picking, integration, and the final matching against 1300 compounds took 30–60 s on a standard laptop, thus providing a fast and easy first examination of active fractions.

2.2.1. Active Components from a Marine-Derived *Penicillium bialowiezense* Strain

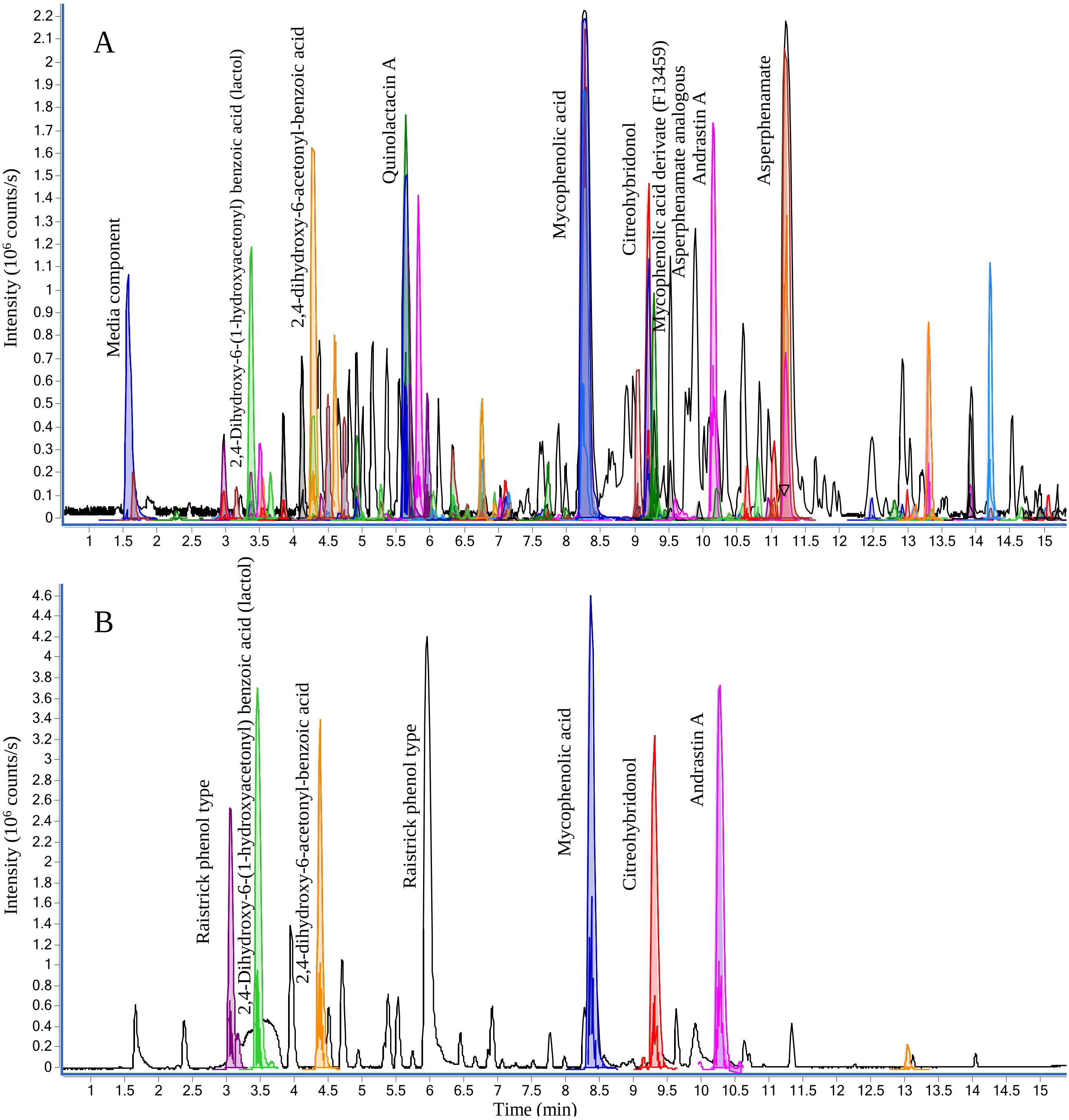
The extract of a *Penicillium bialowiezense* strain (IBT 28294) from a North Sea water sample displayed activity in a CNS assay [51] and anticancer assay (unpublished assay). *P. bialowiezense* is closely related to *P. brevicompactum*, and they are morphologically, genetically, and chemically very difficult to differentiate [52]. They are cosmopolitan species found across an amazing number of habitats such as seaweed, humid indoor environments, soil, and various vegetables and fruits [52,53]. Thus the marine-derived isolate used in this study is likely an opportunist in the marine environment, making the exclusion of known compounds even more important.

The crude extract analyzed in both positive and negative ionization mode can be seen in Figure 4 with the tentative identification of all major peaks: mycophenolic acid, mycophenolic acid derivate (F13459), asperphenamates, andrastin A, quinolactacin A, citreohybridonol, and raistrick phenols [54], all of which have previously been reported from terrestrial fungi [55].

The active fractions were found to contain mycophenolic acid (Figures 4 and 5), which is the active compound in the prodrug CellCept® (Mycophenalate mofetil) used as immunosuppressant in transplant medicine [56]. Several other activities have been reported including antiviral, antitumor [57], and CNS [56,58], in line with the activity observed in this study. The extracted MS/HRMS spectra (Figure 5A) in ESI$^+$ were compared to the mycophenolic acid standard in the MS/HRMS library (Figure 5B) with high scores (>90%) using both reverse and forward searching based on the accuracy of the parent ion (−0.31 ppm for $[M + H]^+$ 321.1328) and specific and abundant fragment ions at *m*/*z* 207.0649 $[C_{11}H_{11}O_4]^+$ and 159.0436 $[C_{10}H_7O_2]^+$.

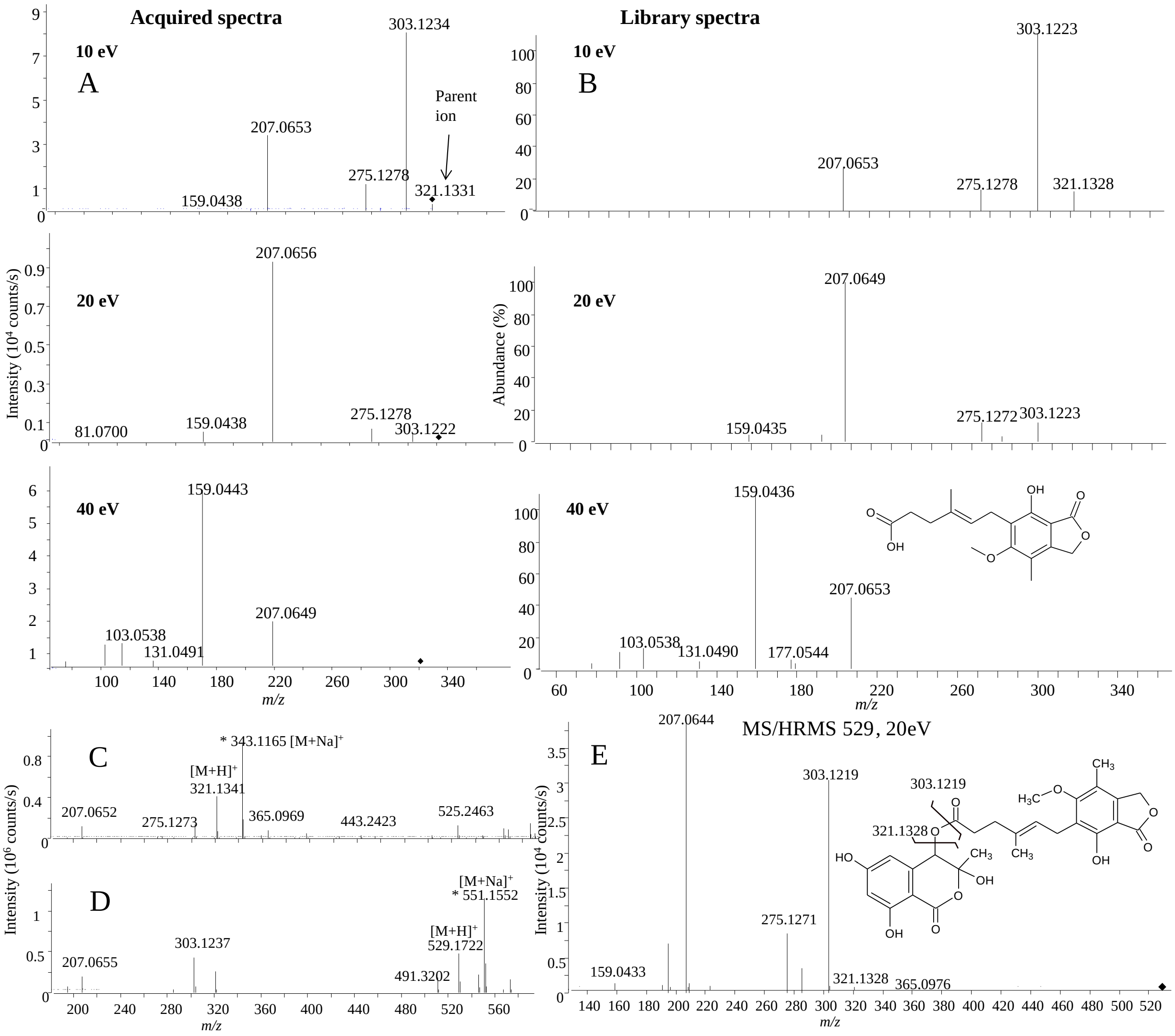
In addition, with ESI$^+$ reverse and forward scoring, a second compound was detected as mycophenolic acid itself but at a wrong retention time and not producing a $[M + Na]^+$ ion (Figure 5D), indicating that it was a fragment from a larger molecule. HRMS of the $[M + H]^+$ at *m*/*z* 529.1722 (Figure 5D) was used to tentatively identify the compound as a mycophenolic acid derivate F13459 previously isolated as a racemate from *Penicillium* sp. [59,60].

Figure 4. Base peak chromatograms (BPC) of the crude extract of *P. bialowiezense* in both positive (**A**) and negative (**B**) ESI modes. Peaks of compounds identified by MS/HRMS using forward scoring are colored.



The identity was verified by MS/HRMS fragmentation of *m*/*z* 529.1722 into the same ions observed from MS/HRMS of $[M + H]^+$ for mycophenolic acid (Figure 5E). F13459 might act as a natural prodrug that by hydrolysis loses the isocoumarin portion, leaving the active compound, mycophenolic acid (Figure 5E). The lost portion corresponds to the lactol form of the raistrick phenol, 2,4-dihydroxy-6-(1-hydroxyacetonyl) benzoic acid(Figure 5E) that was also detected in the extract (Figure 4).

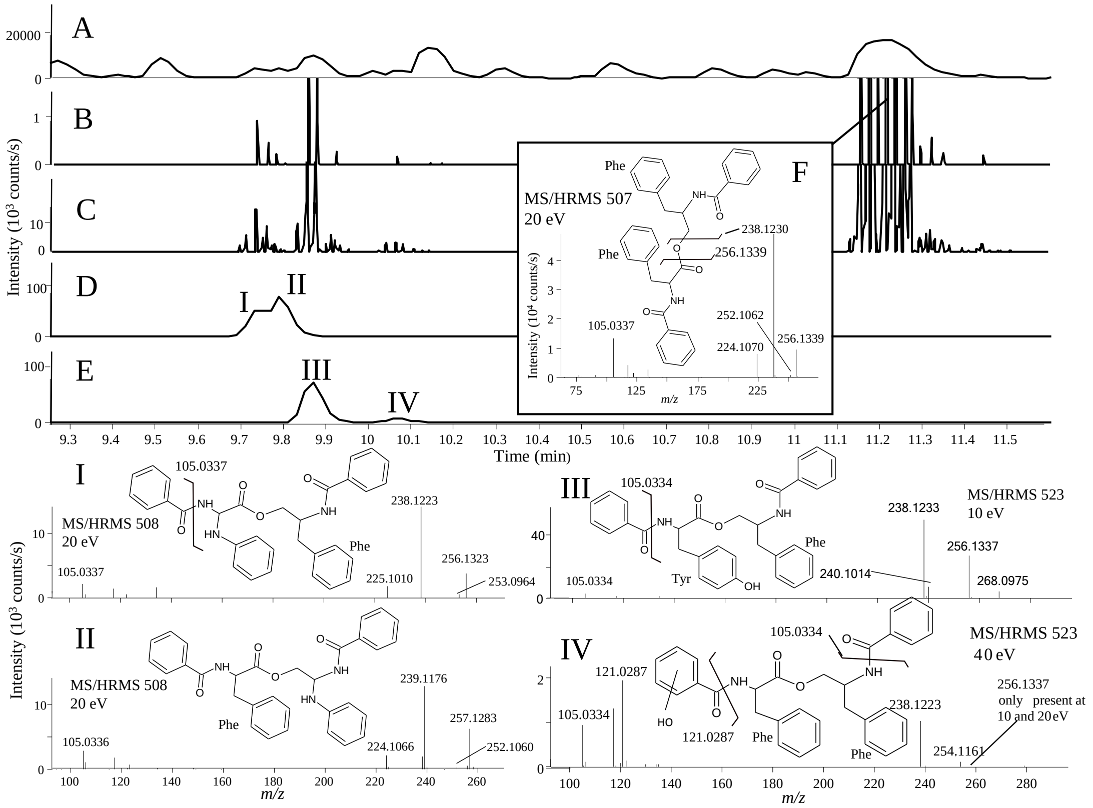
Figure 5. MS/HRMS spectra (*m*/*z* 321) for mycophenolic acid in the active fraction (**A**) compared to library spectra (**B**) at 10, 20 and 40 eV; (**C**) full scan spectrum of mycophenolic acid showing a [M + Na]$^+$ ion at *m*/*z* 343; (**D**) full scan spectrum of F13459 showing a [M + Na]$^+$ ion at *m*/*z* 551; (**E**) MS/HRMS at 20 eV for [M + H]$^+$ of F13459 including structure of the compound.



In the fraction displaying anticancer activity (unpublished assay), the library analysis led to the tentative identification of the fungal anticancer metabolite, asperphenamate (Figure 6F) [61]. A group of peaks eluting close to asperphenamate shared their major fragment ions as found using similarity searching (parent ion not used), which showed the presence of four novel asperphenamate analogues with the tentative structures I to IV (Figure 6). Unambiguous structure verification of course requires isolation and elucidation using nuclear magnetic resonance (NMR) spectroscopy.

Asperhenamate and three of the analogues (I, III, and IV) shared dominant fragment ions at *m*/*z* 238.1230 and 256.1339 (Figure 6B,C), corresponding to [$C_{16}H_{18}NO_2$]$^+$ and [$C_{16}H_{16}NO$]$^+$ formed from the right side of the molecule by cleavage of the ester-bond followed by water loss. The most abundant asperphenamate analogue (III) had a [M + H]$^+$ with *m*/*z* 523.2211, corresponding to an addition of an oxygen atom. This indicated replacement of the phenylalanine by a tyrosine in the asperphenamate skeleton, corroborated by the fragment ions at *m*/*z* 268.0975 $[C_{16}H_{14}NO_3]^+$ and 240.1014 $[C_{15}H_{14}NO_2]^+$ (Figure 6III) as opposed to 252.1062 $[C_{16}H_{14}NO_2]^+$ and 224.1070 $[C_{15}H_{14}NO]^+$ in asperhenamate (Figure 6F). These fragments matched the left side of the molecule formed from the ester cleavage followed by the loss of CO. The fragment 105.0334 $[C_7H_5O]^+$ corresponding to the benzoyl part was present in both asperphenamate and the analogues, and the lack of an ion at *m*/*z* 121.0287 also supported the presence of the tyrosine (Figure 6III).

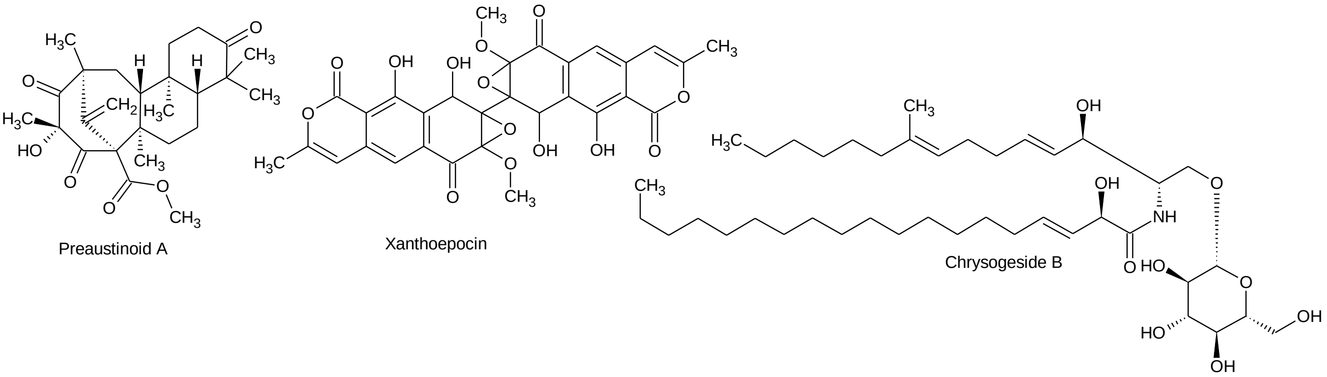
Figure 6. BPC chromatogram of the crude *P. bialowiezense* extract (**A**); EIC from MS/HRMS showing fragment ions (**B**) *m*/*z* 256.1333 and (**C**) 238.123; EIC full scan showing (**D**) *m*/*z* 508.2232 ± 0.005 and (**E**) 523.2211 ± 0.005; (**F**) MS/HRMS spectrum at 20 eV of asperphenemate. (**I**) to (**IV**) show the tentatively assigned isomers of asperphenamate and their positions in the chromatogram.



The other analogue IV with the same accurate mass as III had a similar fragmentation pattern with addition of the most prominent fragment ion 40 eV at *m*/*z* 121.0287 $[C_7H_5O_2]^+$. This fragment could match the presence of an extra oxygen atom in the benzoyl part instead of the phenylalanine part. The last two analogues (I and II) had $[M + H]^+$ *m*/*z* 508.2232 with similar fragmentation patterns to asperphenamate. Analogue II had *m*/*z* 239.1176 $[C_{15}H_{17}N_2O_2]^+$ and *m*/*z* 257.1283 $[C_{15}H_{15}N_2O]^+$ as major fragment ions not present in the asperphenamate MS/HRMS spectrum (Figure 6F), showing a replacement of a CH with an N atom, presumably in the phenylalanine moiety to the right of the ester bond, as a fragment ion corresponding to change in the benzoyl part was not observed. For Analogue I, the two ions differentiating it from asperphenemate were *m*/*z* 253.0964 $[C_{15}H_{13}N_2O_2]^+$ and 225.1010 $[C_{14}H_{13}N_2O]^+$ (Figure 6I), which also corresponded to the replacement of a CH with a nitrogen atom, in this case to the left of the ester bond (Figure 6I) and, as in the example with II, showed a lack of extra fragments.

As the MS/HRMS library only covers about 5% of the compounds reported from fungi in AntiMarine (2012), though with a higher coverage of *Pencilllium* and *Aspergillus* compounds (~20%), we compared the MS/HRMS-based results with those obtained with: (i) aggressive dereplication based on extracted ion chromatograms and isotope patterns, using a search list of all metabolites known from *Penicillium* [20]; and (ii) an unbiased approach based on the Agilent Molecular Feature Extraction (MFE) algorithm which finds all chromatographic peaks and collects adduct, dimeric, and trimeric ions into one feature [62]. The peaks and matching candidates that were identified by the aggressive dereplication approach were evaluated by manually assessing the fragmentation pattern and by using the MassHunter Molecular Structure Correlator program which uses a systematic bond disconnection approach [27]. Likewise, the retention time was compared to the calculated LogD [43], and if possible the UV/Vis data evaluated. This further identified the known compounds chrysogesides B (Figure 7), C, D and E (characteristic loss of glucose and other specific fragments) [63] and three preaustinoids (fragmentations not very specific). Xanthoepocin (Figure 7) [64] was identified and verified from the very specific UV/Vis spectrum and MS/HRMS fragmentations. In full scan positive mode only $[M + Na]^+$ and $[M + H - H_2O]^+$ were observed.

Figure 7. Structures of preaustinoid A, xanthoepocin, and chrysogeside B.



The aggressive dereplication approach also identified fellutamides and breviones which are expected from the species [55]; this could, however, not be supported by the MS/HRMS. Most false positive results originated from fragments or adducts of other compounds in the extract. Examples of these were: (i) the loss of acetate from the andrastin A in ESI^+ matching andibenin B; (ii) quinolactacin A producing $[2M + Na]^+$ and $[2M + H]^+$ ions matching the $[M + Na]^+$ and $[M + H]^+$ of fellutanine D, respectively. Close inspection of adduct pattern and retention times, however, showed that andibenin B and fellutanine D were false positives. This underlines the importance of the MS/HRMS dimension for improved confidence in dereplication. False positives are eliminated and compounds that are missed because they are not part of the library can still be verified based on the MS/HRMS data. The unbiased minimum free energy (MFE) algorithm did, as expected, find many more peaks (50%–100%) than the two targeted approaches (data not shown); however, all major peaks in the chromatograms were detected by the targeted approaches, and all major biological activities could be accounted for by compounds in the MS/HRMS library.

2.2.2. Ophiobolins from a Marine-Derived *Aspergillus insuetus*

The extract of a *Aspergillus insuetus* strain (IBT 28443) derived from a sea water sample collected near Greenland was found to have activity in an anticancer assay (unpublished assay). The most potent fractions were found to be enriched in compounds belonging to the ophiobolin family. They are fungal sesterterpenoids with more than 35 known, closely related analogues [1,65]. Of these analogues, eight were available as standards and included in the library. The ophiobolins are known to exhibit a broad spectrum of bioactivities including antifungal and anticancer [1,65].

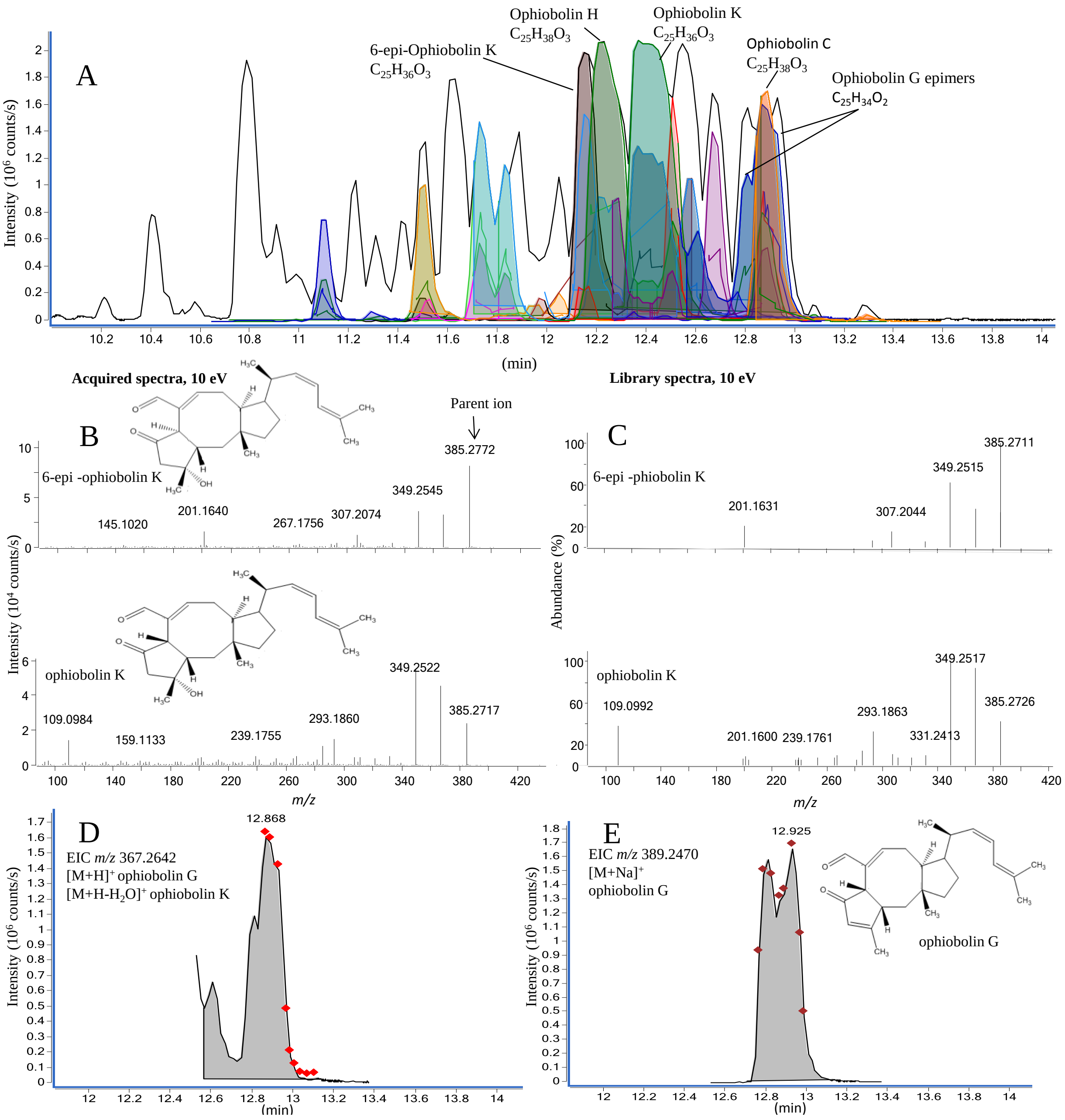
The analysis of a potent fraction is seen in Figure 8A, depicting MS/HRMS library-identified ophiobolins. The identification of four ophiobolins, namely 6-epi-ophiobolin K, ophiobolin H, ophiobolin K, and ophiobolin C, was further corroborated by matching HRMS, retention time, and UV/Vis. Several unidentified ophiobolin analogues seemed to be present in the fraction based on the HRMS and MS/HRMS data.

To illustrate the value of the MS/HRMS library approach, the MassHunter scoring and matching results for the two epimers, 6-epi-ophiobolin K and ophiobolin K (reference standards included in the LC-MS sequence) were compared to demonstrate if compounds varying at only one stereocenter would be unambiguously assigned by the library search. Both reverse and especially forward scoring showed that the epimers could be differentiated based on the intensity for the fragment ions as illustrated for 10 eV in Figure 8 B and C. The forward score for the MS/HRMS of the $[M + H]^+$ ion for the 6-epi-ophiobolin K peak was 71% 6-epi-ophiobolin K and 52% ophiobolin K, while it was 71% ophiobolin K and 58% 6-epi-ophiobolin K for the ophiobolin K peak.

For closely related analogues like the ophiobolins, the number of scans and the integration by auto MS/MS highly influence the outcome from the algorithm. This can be seen in Figure 8 for the series of overlapping peaks (between 12.8 and 13.0 min). From the EIC of the MS/HRMS scans of the *m*/*z* 367.2642 ion (Figure 8D) four peaks at 12.60, 12.78, 12.87, and 12.94 integrated as one peak and the average spectrum was matched to ophiobolin K (forward 81%) as the best match which is incorrect, while the likely correct match ophiobolin G (forward 53%) was the second best match. The reason for the incorrect match was both (i) poor peak integration mixing spectra from several compounds, and (ii) that the water loss ion of ophiobolin K was included in the library as it loses water in the ion source. Looking at the structures of ophiobolins K and G, it is apparent that ophiobolin K reacts into ophiobolin G losing water and forming a double bond. The subsequent MS/HRMS spectra of $[M + H]^+$ ophiobolin G and $[M + H - H_2O]^+$ ophiobolin K will thus be identical.

This underlines the difficulty of differentiation of isomers based on library matches. Fortunately investigating the $[M + Na]^+$ ions (EIC shown in Figure 8E) solves the problem and shows the likely ophiobolin G and 6-epi- ophiobolin G peaks at 12.81 and 12.93 min, respectively. Thus it would strengthen the validity of a compound identification if the matches from the different adducts could be combined and forced to include e.g., the match of the $[M + Na]^+$ ion.

Figure 8. (**A**) Active fraction enriched with compounds from the ophiobolin family in positive ESI mode. More than one color shading of the same peak is either due to different EIC and ECC for same match but different adducts or for other matches that had scored less; (**B**) MS/HRMS acquired spectra (10 eV) for library match of 6-epi-ophiobolin K and ophiobolin K; (**C**) MS/HRMS library spectra (10 eV) of 6-epi-ophiobolin K and ophiobolin K; (**D**) The EIC for the parent *m/z* 367.2642 with ophiobolin K as the best library match; (**E**) The EIC for the parent *m/z* 389.2470 with ophiobolin G as the library match. The diamond markers indicate number of scans across the peak.

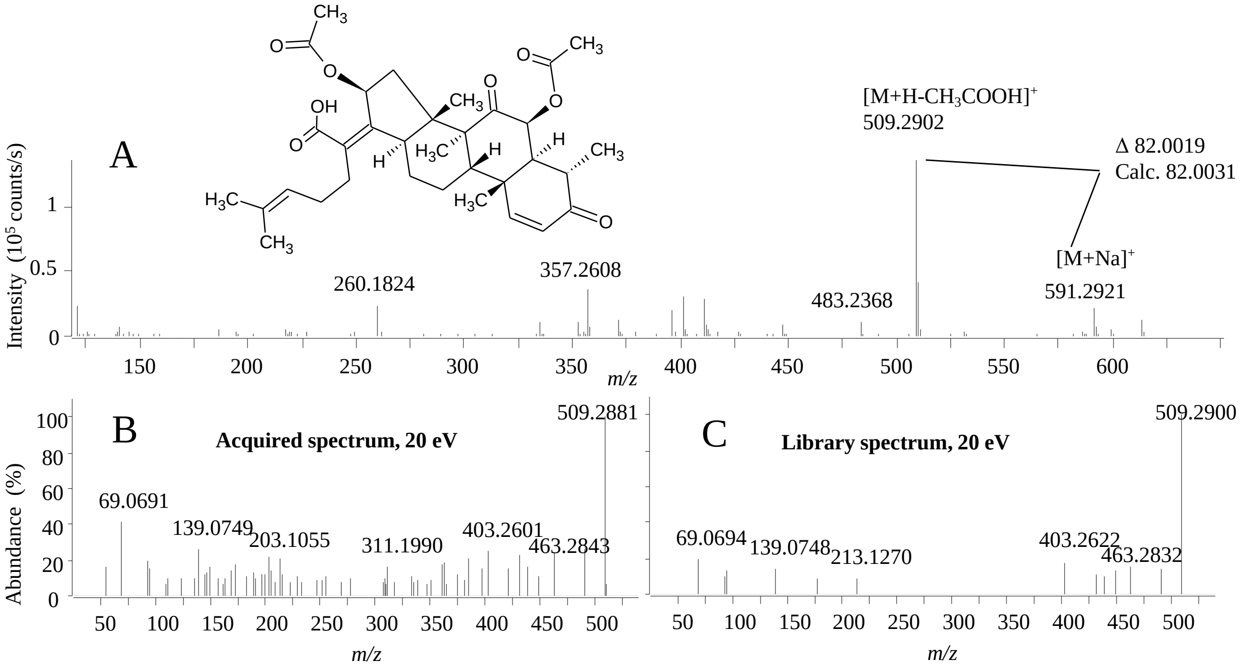


2.2.3. Helvolic Acid as the Anti-Microbial Compound in a Marine-Derived *Emericellopsis* sp.

Emericellopsis sp. strain (IBT 28361), a possibly new species, was isolated from a sea water sample collected off the coast of the Danish island Fanoe. *Emericellopsis* include both terrestrial and marine species with *E. maritima* being associated with the seaweed Fucus (*Phaeophyceae*) [66]. A potent fraction of the crude extract displayed antibacterial activity against methicillin-resistant *Staphylococcus aureus* (MRSA) [67].

Emericellopsis has not been extensively studied for its chemical potential and thus is it likely that it is not well represented by the compounds in the MS/HRMS library. This was the reason for the very few peaks identified by the MS/HRMS approach compared to the previous cases. Nonetheless, the known antibacterial nortriterpenoid, helvolic acid was identified by MS/HRMS (Figure 9) [68,69] consistent with the biological activity of the fraction. In full scan, ESI$^+$ identification was not based on the [M + H]$^+$ but rather the accuracy of the fragment at *m/z* 509.2902 [$C_{31}H_{41}O_6$]$^+$ which was the most abundant peak in the spectrum (Figure 9A). This fragment corresponds to the loss of an O-acetyl group that can be easily lost from the structure of helvolic acid which was verified by the presence of the [M + Na]$^+$ at *m/z* 591.2921 ion also showing that it was not a deacetyl-helvolic acid. As for the ophiobolins, automated use of both the MS/HRMS spectrum and [M + Na]$^+$ from full scan would increase validity of spectral matches. Helvolic acid has formerly been found in related fungal species such as *Emericellopsis terricola* [70] and *Sarocladium oryzae* [71].

Figure 9. (**A**) The structure of helvolic acid and the ESI$^+$ MS spectrum; (**B**) MS/HRMS spectrum at 20 eV from ESI$^+$ of helvolic acid and (**C**) the corresponding library spectrum (parent ion *m/z* 509).



The aggressive dereplication approach based on compounds known from *Emericellopsis* and five related genera (*Acremonium*, *Verticillium*, *Chaetomium*, *Sarocladium*, and *Cephalosporium*) likewise only returned few candidates. The total number of annotated peaks was roughly similar for the two methods. Helvolic acid was annotated by both methods, but apart from that there was almost no overlap between the candidates suggested, underlining the inherent bias of the compound library. The MS/HRMS library is biased by mainly containing metabolites from *Penicillium* and *Aspergillus*, while the targeted is generally biased towards the compounds from the most examined genera. The aggressive dereplication matched the anti-protozoal compound, antiamebin I [72], which was originally not included in the MS/HRMS library. However, this compound could be verified later from a reference standard added to the MS/HRMS library. Using the MFE, a series of another five peptaibiotics in the same mass range as antiamebin was detected. These were not detected by the aggressive dereplication approach, as they were not indexed in AntiMarin 2012. However, searching the monoisotopic masses in The Comprehensive Peptaibiotics Database [73] tentatively identified them as different antiamebins (XIII, XIV, XV, III/IV/IX/VII/VIII, and XVI).

3. Experimental Section

3.1. Strains and Cultivation

All fungal strains used were from the IBT culture collection at the Department of Systems Biology, DTU. The strains described here were *Penicillium antarticum* (IBT 20733 and IBT 27985), *Penicillium bialowiezense* (IBT 28294), *Aspergillus insuetus* (IBT 28443) and *Emericellopsis* sp. (IBT 28361). The marine-derived fungi were cultivated on Czapek yeast extract agar (CYA) and Yeast extract sucrose agar (YES) media for 9 days in the dark at 25 ℃ [43].

3.2. Sample Preparation

Eight plates in total (four CYA and four YES) were extracted with 150 mL ethyl acetate containing 1% formic acid. The crude extracts were fractionated on a reversed phase C_{18} flash column (Sepra ZT, Isolute, 10 g) using an Isolera One automated flash system (Biotage, Uppsala, Sweden). The gradient used was 15%–100% acetonitrile buffered with 20 mM formic acid over 28 min (12 mL/min). Fractions were automatically collected based on UV signal (210 nm and 254 nm). A total of 126 crude, fractions, and blanks were submitted for bioassays antifungal (*A. fumigatus*, *C. albicans*) [74,75], and antibacterial MRSA [67], anticancer (unpublished assay), CNS in zebra fish larvae [51], and anti-inflammatory activity [49].

3.3. Standard Metabolites

Secondary metabolite standards have been collected over the past 30 years, either from commercial sources, as gifts from other research groups, or purified from our own projects [43,76], hence their quantity and purity varies (micro- to milligram quantity, ≥50% purity). The collection contains approximately 1600 standards with 95% of them being of fungal origin (5% of bacterial origin). Commercial sources of purchased standards include Sigma-Aldrich (Steinheim, Germany), Axxora (Bingham, UK), Cayman (Ann Arbor, MI, USA), TebuBio (Le-Perray-en-Yvelines, France), Biopure (Tulln, Austria), Calbiochem (San Diego, CA, USA), ICN (Irvine, CA, USA), Bachem GmbH (Weil am Rhein, Germany), and AnalytiCon Discovery GmbH (Potsdam, Germany). All standards were kept dry at −20 °C and, unless stated otherwise, were dissolved in 140 µL acetonitrile prior to analysis. If not soluble in pure acetonitrile, 50% acetonitrile in MilliQ water was used. Prepared standard solutions were also preserved on −20 ℃.

3.4. UHPLC-DAD-QTOFMS Analysis

Ultra-high performance liquid chromatography-diode array detection-quadruple time of flight mass spectrometry (UHPLC-DAD-QTOFMS) was performed on an Agilent Infinity 1290 UHPLC system (Agilent Technologies, Santa Clara, CA, USA) equipped with a diode array detector. Separation was obtained on an Agilent Poroshell 120 phenyl-hexyl column (2.1 × 150 mm, 2.7 μm) with a linear gradient consisting of water (A) and acetonitrile (B) both buffered with 20 mM formic acid, starting at 10% B and increased to 100% in 15 min where it was held for 2 min, returned to 10% in 0.1 min and keeping it for 3 min (0.35 mL/min, 60 ℃). Injection volume, depending on sample concentration, typically varied between 0.1 and 1 μL. To avoid carry-over, the auto-sampler was operated in the flow-through-needle mode and further coupled to an Agilent Flex Cube which was used to back flush the needle seat for 15 s. at a flow of 4 mL/min with each of: (i) isopropanol: 0.2% ammonium hydroxide in water (1:1 *v/v*); (ii) acetonitrile with 2% formic acid; (iii) water with 2% formic acid.

MS detection was done on an Agilent 6550 iFunnel QTOF MS equipped with Agilent Dual Jet Stream electrospray ion source with the drying gas temperature of 160 ℃ and gas flow of 13 L/min and sheath gas temperature of 300 ℃ and flow of 16 L/min. Capillary voltage was set to 4000 V and nozzle voltage to 500 V. Ion-source parameters were the same for ESI^+ and ESI^- mode. Mass spectra were recorded as centroid data for *m*/*z* 85–1700 in MS mode and *m*/*z* 30–1700 in MS/MS mode, with an acquisition rate of 10 spectra/s. Automated data-dependent acquisition MS/HRMS (auto-MS/HRMS) analysis was commonly done for ions detected in the full scan above 50,000 counts (may be adjusted for low/high concentration samples) with a cycle time of 0.5 s, the quadrupole isolation width in narrow (*m*/*z* ±0.65), using fixed CID energies of 10, 20, and 40 eV and maximum three selected precursor ions per cycle. A narrow exclusion time of 0.04 min was used to get MS/MS of less abundant ions when compounds co-eluted.

Lock mass solution in 95% acetonitrile was infused in the second sprayer using an extra LC pump at a flow of 10–50 μL/min, the solution contained 1 μM tributyle amine (Sigma-Aldrich), 10 μM Hexakis(2,2,3,3-tetrafluoropropoxy)phosphazene (Apollo Scientific Ltd., Cheshire, UK), and 1 μM trifluoroacetic acid (Sigma-Aldrich) as lock masses. The $[M + H]^+$ ions of first two (*m*/*z* 186.2216 and 922.0098 respectively) were used in positive mode, while $[M + HCOO]^-$ and $[M - H]^-$ of the latter two were used in negative mode (*m*/*z* 966.0007 and 112.9856).

3.5. Library Setup and Auto-MS/MS Data Analysis

The MS/HRMS library was constructed from our internal ChemFolder library (Advanced Chemical Developments, Toronto, ON, Canada) of 7400 compounds of which 1600 were available as reference standards [20]. For reference standards and tentatively identified compounds, name, structure, and CAS no. were transferred to the Agilent Masshunter PCDL manager 4.00 (Service release 1), and linked to the retention time and MS/HRMS spectra of 10, 20, and 40 eV, either by manually pasting from MassHunter or imported via a cef file. All major pseudomolecular ions ($[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$, $[M - H]^-$, $[M + HCOO]^-$), and simple fragment ions (mainly $[M + H - (H_2O)_n]^+$, $[M + H - CH_3COOH]^+$, $[M – H - CO_2]^-$) which provided characteristic MS/MS spectra were included.

Data files were processed by the Find by Auto MS/MS function in Masshunter, usually without any intensity threshold but often with a limit to the 200 largest peaks, mass match tolerance *m*/*z* 0.05. Unless otherwise stated the MS/HRMS library was searched using precursor and product ion expansion of 50 ppm + 2 mDa as well as minimal reverse and forward scores of 50 each.

3.6. Aggressive Dereplication and Molecular Feature Extraction

For analysis of compounds described in the literature and not necessarily available as reference standards, Aggressive dereplication (Klitgaard *et al.* 2014 [20]) was performed on the ESI$^+$ and ESI$^-$ full scan data using the *Find by Formulae* function in Agilent Masshunter Qualitative analysis B06.00 software. The following adducts and common fragments were included: ESI$^+$, [M + H]$^+$ and [M + Na]$^+$; ESI$^-$, [M − H]$^-$, [M + HCOO]$^-$. All ions analyzed were treated as being singularly charged. The area cut-off was set to 10,000, and the mass spectrum was recorded below 10% of the height of the peak to avoid detector overload. A minimum score of 70 was used to ensure that only compounds with fitting isotope patterns were annotated.

The search lists were constructed from the AntiMarin2012 which was converted into an sdf-database and then imported into ChemFolder and from here to Excel (Klitgaard *et al.* 2014 [20]) where it was formatted to the Agilent search list format. All this work was made on an AntiMarin-licensed computer.

The MFE screening was performed in the Agilent Masshunter Qualitative analysis B06.00. The following adducts and common fragments were included: ESI$^+$, [M + H]$^+$ and [M + Na]$^+$; ESI$^-$, [M − H]$^-$, [M + HCOO]$^-$. All ions analyzed were treated as being singly charged. The area cut-off was set to 10,000, and the mass spectrum was recorded below 10% of the height of the peak to avoid detector overload. A minimum quality score of 99 was used to ensure that only compounds with fitting mass, isotope patterns, and peak shape were annotated.

4. Conclusions

In this work we demonstrate that MS/HRMS search in a library is a robust and reliable way of tentatively identifying known bioactive compounds on a single instrument. With spectra reproducibility across Agilent instruments [26,77] the library should be directly usable on these, while others instruments presumably need adjustment against collision energies (e.g., 10 eV on the Agilent may correspond to 15 eV on a Bruker QTOF). Furthermore MS/HRMS aided the tentative identification of novel isomers, e.g., to be used in bioactivity optimization. Many highly bioactive compounds are found across the fungal kingdom, and even when exploring specialized marine environments where it is likely to find novel bioactive compounds it is of outmost importance to identify known nuisance compounds in the first screen. To aid drug discovery dereplication we thus suggest that it is required to deposition MS/MS spectra of all novel published compounds in Massbank, MetLin and/or GNPS [31], although for all mentioned an easy interface for depositing spectra is needed.

Aggressive dereplication of full scan data supplemented by auto MS/HRMS to strengthen the correct match and elimination of false positives proved efficient and could in many cases be strengthened even further by UV/Vis data.

Both described strategies can handle extracts produced months in-between which is a problem for the unbiased peak picking and adduct pattern algorithms which in general requires samples to be run within a sequence and with replicated and blank samples to handle variations in chromatographic separation, mass spectra, sample preparation, and growth media. Nonetheless, an unbiased peak picking strategy was the only way to detect a series of non-data based compounds as demonstrated in the last case, proving the need to integrate many data-analysis strategies and tools to obtain comprehensive compound coverage.

Acknowledgments

Funding is acknowledged from PharmaSea (Grant Agreement No 312184). We are grateful to Agilent technologies for the Thought Leader Donation of the UHPLC-QTOF system and for financial support from the Villum foundation for I.D.

Author Contributions

S.K., I.D., and K.F.N. designed the research; S.K. and I.D. performed the experimental work; S.K., I.D., J.C.F. and A.K. analysed the data; T.O.L. and J.C.F. provided strains and samples as well as assisted in compound identification; S.K., M.M., I.D., and K.F.N. wrote the paper. All authors read and corrected the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bladt, T.T.; Frisvad, J.C.; Knudsen, P.B.; Larsen, T.O. Anticancer and Antifungal Compounds from *Aspergillus*, *Penicillium* and Other Filamentous Fungi. *Molecules* **2013**, *18*, 11338–11376.
2. El-Elimat, T.; Figueroa, M.; Ehrmann, B.M.; Cech, N.B.; Pearce, C.J.; Oberlies, N.H. High-Resolution MS, MS/MS, and UV Database of Fungal Secondary Metabolites as a Dereplication Protocol for Bioactive Natural Products. *J. Nat. Prod.* **2013**, *76*, 1709–1716.
3. Mouton, M.; Postma, F.; Wilsenach, J.; Botha, A. Diversity and Characterization of Culturable Fungi from Marine Sediment Collected from St. Helena Bay, South Africa. *Microb. Ecol.* **2012**, *64*, 311–319.
4. Richards, T.A.; Jones, M.D.; Leonard, G.; Bass, D. Marine Fungi: Their Ecology and Molecular Diversity. *Ann. Rev. Mar. Sci.* **2012**, *4*, 495–522.
5. Debbab, A.; Aly, A.H.; Lin, W.H.; Proksch, P. Bioactive Compounds from Marine Bacteria and Fungi. *Microb. Biotechnol.* **2010**, *3*, 544–563.
6. Jensen, P.A.; Mincer, T.J.; Williams, P.G.; Fenical, W. Marine actinomycete diversity and natural product discovery. *Antonie Van Leeuwenhoek* **2005**, *87*, 43–48.
7. Holmstrom, C.; Kjelleberg, S. Marine *Pseudoalteromonas* species are associated with higher organisms and produce biologically active extracellular agents. *FEMS Microbiol. Ecol.* **1999**, *30*, 285–293.

8. Bowman, J.P. Bioactive compound synthetic capacity and ecological significance of marine bacterial genus *Pseudoalteromonas*. *Mar. Drugs* **2007**, *5*, 220–241.
9. Mansson, M.; Gram, L.; Larsen, T.O. Production of Bioactive Secondary Metabolites by Marine Vibrionaceae. *Mar. Drugs* **2011**, *9*, 1440–1468.
10. Jones, E. Are there more marine fungi to be described? *Bot. Mar.* **2011**, *54*, 343–354.
11. Jones, E. Fifty years of marine mycology. *Fungal Divers* **2011**, *50*, 73–112.
12. Burgaud, G.; Woehlke, S.; Redou, V.; Orsi, W.; Beaudoin, D.; Barbier, G.; Biddle, J.F.; Edgcomb, V.P. Deciphering the presence and activity of fungal communities in marine sediments using a model estuarine system. *Aquat. Microb. Ecol.* **2013**, *70*, 45–62.
13. Bugni, T.S.; Janso, J.E.; Williamson, R.T.; Feng, X.; Bernan, V.S.; Greenstein, M.; Carter, G.T.; Maiese, W.M.; Ireland, C.M. Dictyosphaeric Acids A and B: New Decalactones from an Undescribed *Penicillium* sp. Obtained from the Alga Dictyosphaeria versluyii. *J. Nat. Prod.* **2004**, *67*, 1396–1399.
14. Abdel-Wahab, M.; Gareth Jones, E.B. Three new marine ascomycetes from driftwood in Australia sand dunes. *Mycoscience* **2000**, *41*, 379–388.
15. Loque, C.P.; Medeiros, A.O.; Pellizzari, F.M.; Oliveira, E.C.; Rosa, C.A.; Rosa, L.H. Fungal community associated with marine macroalgae from Antarctica. *Polar Biol.* **2010**, *33*, 641–648.
16. Burgaud, G.; Le Calvez, T.; Arzur, D.; Vandenkoornhuyse, P.; Barbier, G. Diversity of culturable marine filamentous fungi from deep-sea hydrothermal vents. *Environ. Microbiol.* **2009**, *11*, 1588–1600.
17. Khudyakova, Y.V.; Pivkin, M.V.; Kuznetsova, T.A.; Svetashev, V.I. Fungi in sediments of the sea of Japan and their biologically active metabolites. *Microbiology* **2000**, *69*, 608–611.
18. Alker, A.P.; Smith, G.W.; Kim, K. Characterization of *Aspergillus sydowii* (Thom et Church), a fungal pathogen of Caribbean sea fan corals. *Hydrobiologia* **2001**, *460*, 105–111.
19. Roy, K.; Mukhopadhyay, T.; Reddy, G.C.S.; Desikan, K.R.; Ganguli, B.N. Mulundocandin, A New Lipopeptide Antibiotic. 1. Taxonomy, Fermentation, Isolation and Characterization. *J. Antibiot.* **1987**, *40*, 275–280.
20. Klitgaard, A.; Iversen, A.; Andersen, M.R.; Larsen, T.O.; Frisvad, J.C.; Nielsen, K.F. Aggressive dereplication using UHPLC-DAD-QTOF—Screening extracts for up to 3000 fungal secondary metabolites. *Anal. Bioanal. Chem.* **2014**, *406*, 1933–1943.
21. Murray, K.K. Glossary of terms for separations coupled to mass spectrometry. *J. Chromatogr. A* **2010**, *1217*, 3922–3928.
22. Schymanski, E.L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H.P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48*, 2097–2098.
23. Hu, Q.; Noll, R.J.; Li, H.; Makarov, A.; Hardmac, M.; Cooksa, R.G. The Orbitrap: A new mass spectrometer. *J. Mass Spectrom.* **2005**, *40*, 430–443.
24. Lehner, S.M.; Neumann, N.K.N.; Sulyok, M.; Lemmens, M.; Krska, R.; Schuhmacher, R. Evaluation of LC-high-resolution FT-Orbitrap MS for the quantification of selected mycotoxins and the simultaneous screening of fungal metabolites in food. *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.* **2011**, *28*, 1457–1468.

25. Konishi, Y.; Kiyota, T.; Draghici, C.; Gao, J.M.; Yeboah, F.; Acoca, S.; Jarussophon, S.; Purisima, E. Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Anal. Chem.* **2007**, *79*, 1187–1197.
26. Broecker, S.; Herre, S.; Wust, B.; Zweigenbaum, J.; Pragst, F. Development and practical application of a library of CID accurate mass spectra of more than 2500 toxic compounds for systematic toxicological analysis by LC-QTOF-MS with data-dependent acquisition. *Anal. Bioanal. Chem.* **2011**, *400*, 101–117.
27. Hill, A.W.; Mortishire-Smith, R.J. Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111–3118.
28. Hufsky, F.; Scheubert, K.; Bocker, S. Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends Anal. Chem.* **2014**, *53*, 41–48.
29. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; *et al*. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714.
30. Smith, C.A.; O'Maille, G.; Want, E.J.; Qin, C.; Trauger, S.A.; Brandon, T.R.; Custodio, D.E.; Abagyan, R.; Siuzdak, G. METLIN: A metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747–751.
31. Global Natural Products Research and Mass spectrometry. Available online: http://gnps.ucsd.edu/ (accessed on 17 June 2014).
32. Champarnaud, E.; Hopley, C. Evaluation of the comparability of spectra generated using a tuning point protocol on twelve electrospray ionisation tandem-in-space mass spectrometers. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1001–1007.
33. Yates, J.R.; Cociorva, D.; Liao, L.J.; Zabrouskov, V. Performance of a linear ion trap-orbitrap hybrid for peptide analysis. *Anal. Chem.* **2006**, *78*, 493–500.
34. Kind, T.; Liu, K.H.; Lee, D.Y.; DeFelice, B.; Meissen, J.K.; Fiehn, O. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods* **2013**, *10*, 755–758.
35. Fredenhagen, A.; Derrien, C.; Gassmann, E. An MS/MS Library on an Ion-Trap Instrument for Efficient Dereplication of Natural Products. Different Fragmentation Patterns for $[M + H]^+$ and $[M + Na]^+$ Ions. *J. Nat. Prod.* **2005**, *68*, 385–391.
36. Yang, J.Y.; Sanchez, L.M.; Rath, C.M.; Liu, X.; Boudreau, P.D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; *et al*. Molecular Networking as a Dereplication Strategy. *J. Nat. Prod.* **2013**, *76*, 1686–1699.
37. Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B.S.; Yang, J.Y.; Kersten, R.D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J.M.; *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1743–E1752.
38. PharmaSea. Available online: http://www.pharma-sea.eu/ (accessed on 13 June 2014).
39. Hopfgartner, G.; Vilbois, F. The impact of accurate mass measurements using quadrupole/time-of-flight mass spectrometry on the characterisation and screening of drug metabolites. *Analusis* **2000**, *28*, 906–914.

40. Colombo, M.; Sirtori, F.R.; Rizzo, V. A fully automated method for accurate mass determination using high-performance liquid chromatography with a quadrupole/orthogonal acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 511–517.
41. Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.* **2007**, *8*, 105.
42. Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszar, E.; Kofeler, H.C. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *J. Mass Spectrom.* **2009**, *44*, 485–493.
43. Nielsen, K.F.; Månsson, M.; Rank, C.; Frisvad, J.C.; Larsen, T.O. Dereplication of microbial natural products by LC-DAD-TOFMS. *J. Nat. Prod.* **2011**, *74*, 2338–2348.
44. DTU Mycotoxin-Fungal Secondary Metabolite MS/HRMS library. Available online: http://www.bio.dtu.dk/english/Research/Platforms/Metabolom/MSMSLib (accessed on 13 June 2014).
45. Liu, B.-H.; Yu, F.; Wu, T.-S.; Li, W. Evaluation of genotoxic risk and oxidative DNA damage in mammalian cells exposed to mycotoxins, patulin and citrinin. *Toxicol. Appl. Pharmacol.* **2003**, *191*, 255–263.
46. Vansteelandt, M.; Kerzaon, I.; Blanchet, E.; Tankoua, O.F.; du Pont, T.R.; Joubert, Y.; Monteau, F.; Le Bizec, B.; Frisvad, J.C.; Pouchus, Y.F.; *et al*. Patulin and secondary metabolite production by marine-derived *Penicillium* strains. *Fungal Biol.* **2012**, *116*, 954–961.
47. Rasmussen, T.B.; Skindersoe, M.E.; Bjarnsholt, T.; Phipps, R.K.; Christensen, K.B.; Jensen, P.O.; Andersen, J.B.; Koch, B.; Larsen, T.O.; Hentzer, M.; *et al*. Identity and effects of quorum-sensing inhibitors produced by *Penicillium* species. *Microbiology* **2005**, *151*, 1325–1340.
48. Gram, L.; Melchiorsen, J.; Bruhn, J.B. Antibacterial Activity of Marine Culturable Bacteria Collected from a Global Sampling of Ocean Surface Waters and Surface Swabs of Marine Organisms. *Mar. Biotechnol.* **2010**, *12*, 439–451.
49. Lind, K.F.; Hansen, E.; Osterud, B.; Eilertsen, K.E.; Bayer, A.; Engqvist, M.; Leszczak, K.; Jorgensen, T.O.; Andersen, J.H. Antioxidant and Anti-Inflammatory Activities of Barettin. *Mar. Drugs* **2013**, *11*, 2655–2666.
50. Bohni, N.; Lorena Cordero-Maldonado, M.; Maes, J.; Siverio-Mota, D.; Marcourt, L.; Munck, S.; Kamuhabwa, A.R.; Moshi, M.J.; Esguerra, C.V.; de Witte, P.A.; *et al*. Integration of Microfractionation, qNMR and Zebrafish Screening for the In Vivo Bioassay-Guided Isolation and Quantitative Bioactivity Analysis of Natural Products. *PLoS One* **2013**, *8*, e64006.
51. Buenafe, O.E.; Orellana-Paucar, A.; Maes, J.; Huang, H.; Ying, X.; de Borggraeve, W.; Crawford, A.D.; Luyten, W.; Esguerra, C.V.; de Witte, P. Tanshinone IIA Exhibits Anticonvulsant Activity in Zebrafish and Mouse Seizure Models. *ACS Chem. Neurosci.* **2013**, *4*, 1479–1487.
52. Frisvad, J.C.; Samson, R.A. Polyphasic taxonomy of *Penicillium* subgenus *Penicillium*. A guide to identification of the food and air-borne terverticillate Penicillia and their mycotoxins. *Stud. Mycol.* **2004**, *49*, 1–173.
53. Overy, D.P.; Frisvad, J.C. Mycotoxin production and postharvest storage rot of ginger (Zingiber officinale) by Penicillium brevicompactum. *J. Food Prot.* **2005**, *68*, 607–609.
54. Andersen, B. Consistent production of phenolic compounds by *Penicillium brevicompactum* from chemotaxonomic characterization. *Antonie Van Leeuwenhoek* **1991**, *60*, 115–123.

55. Frisvad, J.C.; Smedsgaard, J.; Larsen, T.O.; Samson, R.A. Mycotoxins, drugs and other extrolites produced by species in *Penicillium* subgenus *Penicillium*. *Stud. Mycol.* **2004**, *49*, 201–241.
56. Bentley, R. Mycophenolic acid: A one hundred year Odyssey from antibiotic to immunosuppressant. *Chem. Rev.* **2000**, *100*, 3801–3825.
57. Williams, R.H.; Lively, D.H.; Delong, D.C.; Cline, J.C.; Sweeney, M.J.; Poore, G.A.; Larsen, S.H. Mycophenolic Acid—Antiviral and Antitumor Properties. *J. Antibiot.* **1968**, *21*, 463–464.
58. Kern, I.; Xu, R.; Julien, S.; Suter, D.; Preynat-Seauve, O.; Baquie, M.; Poncet, A.; Combescure, C.; Stoppini, L.; Thriel, C.V.; *et al*. Embryonic Stem Cell-Based Screen for Small Molecules: Cluster Analysis Reveals Four Response Patterns in Developing Neural Cells. *Curr. Med. Chem.* **2013**, *20*, 710–723.
59. Koshino, H.; Muroi, M.; Tajika, T.; Kimura, Y.; Takatsuki, A. F13459, a new derivative of mycophenolic acid: II. Physico-chemical properties and structural elucidation. *J. Antibiot.* **2001**, *54*, 494–500.
60. Muroi, M.; Sano, K.; Okada, G.; Koshino, H.; Oku, T.; Takatsuki, A. F13459, a new derivative of mycophenolic acid: I. Taxonomy, isolation, and biological properties. *J. Antibiot.* **2001**, *54*, 489–493.
61. Yuan, L.; Li, Y.; Zou, C.; Wang, C.; Gao, J.; Miao, C.; Ma, E.; Sun, T. Synthesis and *in vitro* antitumor activity of asperphenamate derivatives as autophagy inducer. *Bioorg. Med. Lett.* **2012**, *22*, 2216–2220.
62. Hu, Y.M.; Yu, Z.L.; Yang, Z.J.; Zhu, G.Y.; Fong, W.F. Comprehensive chemical analysis of Venenum Bufonis by using liquid chromatography/electrospray ionization tandem mass spectrometry. *J. Pharm. Biomed. Anal.* **2011**, *56*, 210–220.
63. Peng, X.; Wang, Y.; Sun, K.; Liu, P.; Yin, X.; Zhu, W. Cerebrosides and 2-Pyridone Alkaloids from the Halotolerant Fungus Penicillium chrysogenum Grown in a Hypersaline Medium. *J. Nat. Prod.* **2011**, *74*, 1298–1302.
64. Ando, T.; Igarashi, Y.; Kuwamori, Y.; Takagi, K.; Ando, T.; Fudou, R.; Furumai, T.; Oki, T. Xanthoepocin, a new antibiotic fom *Penicillium simplicissimum* IFO5762. *J. Antibiot.* **2000**, *53*, 928–933.
65. Bladt, T.T.; Duerr, C.; Knudsen, P.B.; Kildgaard, S.; Frisvad, J.C.; Gotfredsen, C.H.; Seiffert, M.; Larsen, T.O. Bio-Activity and Dereplication-Based Discovery of Ophiobolins and Other Fungal Secondary Metabolites Targeting Leukemia Cells. *Molecules* **2013**, *18*, 14629–14650.
66. Zuccaro, A.; Summerbell, R.C.; Gams, W.; Schroers, H.J.; Mitchell, J.I. A new *Acremonium* species associated with Fucus spp., and its affinity with a phylogenetically distinct marine Emericellopsis clade. *Stud. Mycol.* **2004**, *50*, 283–297.
67. Graca, A.P.; Bondoso, J.; Gaspar, H.; Xavier, J.R.; Monteiro, M.C.; de la Cruz, M.; Oves-Costales, D.; Vicente, F.; Lage, O.M. Antimicrobial Activity of Heterotrophic Bacterial Communities from the Marine Sponge *Erylus discophorus* (Astrophorida, Geodiidae). *PLoS One* **2013**, *8*, e78992.
68. Ratnaweera, P.B.; Williams, D.E.; de Silva, E.D.; Wijesundera, R.L.C.; Dalisay, D.S.; Andersen, R.J. Helvolic acid, an antibacterial nortriterpenoid from a fungal endophyte, *Xylaria* sp. of orchid *Anoectochilus setaceus* endemic to Sri Lanka. *Mycology* **2014**, *5*, 23–28.

69. Qin, L.; Li, B.; Guan, J.; Zhang, G. *In vitro* synergistic antibacterial activities of helvolic acid on multi-drug resistant *Staphylococcus aureus*. *Nat. Prod. Res.* **2009**, *23*, 309–318.
70. Pinheiro, A.; Dethoup, T.; Bessa, J.; Silva, A.M.; Kijjoa, A. A new bicyclic sesquiterpene from the marine sponge associated fungus *Emericellopsis minima*. *Phytochem. Lett.* **2012**, *5*, 68–70.
71. Bills, G.F.; Platas, G.; Gams, W. Conspecificity of the cerulenin and helvolic acid producing '*Cephalosporium caerulens*', and the hypocrealean fungus *Sarocladium oryzae*. *Mycol. Res.* **2004**, *108*, 1291–1300.
72. Thirumalachar, M.J. Antiamoebin Anti Parasit A New Anti Protozoal Anti Helminthic Antibiotic I Production and Biological Studies Emericellopsis-Poonensis Emericellopsis-Synnematicola Cephalosporium-Pimprina. *Hindustan Antibiot. Bull.* **1968**, *10*, 287–289.
73. Stoppacher, N.; Neumann, N.K.N.; Burgstaller, L.; Zeilinger, S.; Degenkolb, T.; Bruckner, H.; Schuhmacher, R. The Comprehensive Peptaibiotics Database. *Chem. Biodiv.* **2013**, *10*, 734–743.
74. Monteiro, M.C.; de la Cruz, M.; Cantizani, J.; Moreno, C.; Tormo, J.R.; Mellado, E.; de Lucas, J.R.; Asensio, F.; Valiante, V.; Brakhage, A.A.; *et al*. A New Approach to Drug Discovery: High-Throughput Screening of Microbial Natural Extracts against Aspergillus fumigatus Using Resazurin. *J. Biomol. Screen.* **2012**, *17*, 542–549.
75. De la Cruz, M.; Martin, J.; Gonzalez-Menendez, V.; Perez-Victoria, I.; Moreno, C.; Tormo, J.R.; El Aouad, N.; Guarro, J.; Vicente, F.; Reyes, F.; *et al*. Chemical and Physical Modulation of Antibiotic Activity in Emericella Species. *Chem. Biodiv.* **2012**, *9*, 1095–1113.
76. Frisvad, J.C.; Thrane, U. Standardised High-Performance Liquid Chromatography of 182 mycotoxins and other fungal metabolites based on alkylphenone retention indices and UV-VIS spectra (Diode Array Detection). *J. Chromatogr.* **1987**, *404*, 195–214.
77. Zhu, Z.J.; Schultz, A.W.; Wang, J.H.; Johnson, C.H.; Yannone, S.M.; Patti, G.J.; Siuzdak, G. Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nat. Protoc.* **2013**, *8*, 451–460.

6.3 Paper 3 – Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in *Aspergillus niger*

Holm, D. K., Petersen, L. M., **Klitgaard, A.**, Knudsen, P. B. Jarczynska, Z. D., Nielsen, K. F., Gotfredsen, C. H., Larsen, T. O., & Mortensen, U. H.

Paper accepted in Chemistry and Biology (2014)

Molecular and Chemical Characterization of the Biosynthesis of the 6-MSA-Derived Meroterpenoid Yanuthone D in *Aspergillus niger*

Dorte K. Holm,[1,6] Lene M. Petersen,[2,6] Andreas Klitgaard,[3] Peter B. Knudsen,[5] Zofia D. Jarczynska,[1] Kristian F. Nielsen,[3] Charlotte H. Gotfredsen,[4] Thomas O. Larsen,[2,*] and Uffe H. Mortensen[1,*]

[1]Eukaryotic Molecular Cell Biology Group, Department of Systems Biology, Center for Microbial Biotechnology, Soltofts Plads, Building 223, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[2]Chemodiversity Group, Department of Systems Biology, Center for Microbial Biotechnology, Soltofts Plads, Building 221, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[3]Metabolic Signaling and Regulation Group, Department of Systems Biology, Center for Microbial Biotechnology, Soltofts Plads, Building 221, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[4]Department of Chemistry, Kemitorvet, Building 201, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[5]Fungal Physiology and Biotechnology Group, Department of Systems Biology, Center for Microbial Biotechnology, Soltofts Plads, Building 223, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

[6]These authors contributed equally to this work

*Correspondence: tol@bio.dtu.dk (T.O.L.), um@bio.dtu.dk (U.H.M.)

http://dx.doi.org/10.1016/j.chembiol.2014.01.013

SUMMARY

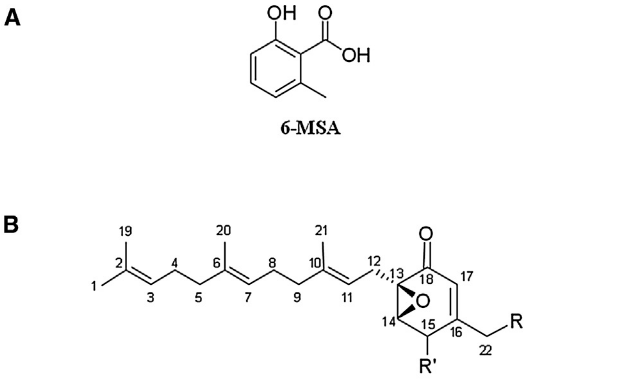
Secondary metabolites in filamentous fungi constitute a rich source of bioactive molecules. We have deduced the genetic and biosynthetic pathway of the antibiotic yanuthone D from *Aspergillus niger*. Our analyses show that yanuthone D is a meroterpenoid derived from the polyketide 6-methylsalicylic acid (6-MSA). Yanuthone D formation depends on a cluster composed of ten genes including *yanA* and *yanI*, which encode a 6-MSA polyketide synthase and a previously undescribed O-mevalon transferase, respectively. In addition, several branching points in the pathway were discovered, revealing five yanuthones (F, G, H, I, and J). Furthermore, we have identified another compound (yanuthone X_1) that defines a class of yanuthones that depend on several enzymatic activities encoded by genes in the *yan* cluster but that are not derived from 6-MSA.

INTRODUCTION

Fungal polyketides (PKs) comprise a large and complex group of metabolites with a wide range of bioactivities. Hence, the group includes compounds that are used by fungi as pigments for UV-light protection, in intra- and interspecies signaling, and in chemical warfare against competitors (Williams et al., 1989). Many PKs are mycotoxins that are harmful to human health, e.g., patulin and the highly carcinogenic aflatoxins (Olsen et al., 1988). On the other hand, several PKs have a great medical potential, e.g., cholesterol-lowering statins (Endo et al., 1976), the antimicrobial and immunosuppressive mycophenolic acid (Bentley, 2000), the acetyl-coenzyme A acetyltransferase-inhibiting pyripyropenes (Frisvad et al., 2009), and the farnesyltransferase inhibiting andrastins (Rho et al., 1998). Although more than 6,000 different PKs have been isolated and characterized (AntiBase 2012), these compounds are likely only the tip of the iceberg. For example, for each fungus analyzed, only a small part of its full repertoire of PKs genes appears to be produced under laboratory conditions (Pel et al., 2007; Andersen et al., 2013). In agreement with this view, genome sequencing of several fungal species have uncovered far more genes for PKs production than can be accounted for by the number of compounds that they are actually known to produce. Hence, the chemical space of PKs is far from fully known, and many new drugs and mycotoxins await discovery.

The fungal genome sequencing projects have demonstrated that genes necessary for production of individual PKs often cluster around the gene encoding the polyketide synthase (PKS), which delivers the first intermediate in a given PK pathway. Although this is helpful for pathway elucidation, compounds produced by orphan gene clusters (Gross, 2007) can still not be easily predicted by bioinformatic tools (for review, see Cox, 2007 and Hertweck, 2009). This is because most fungal PKs are produced by type I iterative PKSs whose products are notoriously difficult to predict. Moreover, the specificities and the order of actions of the tailoring enzymes that modify the PK released from the PKS further complicate prediction of the end products. To elucidate the biochemical pathway of an orphan gene cluster, it is therefore necessary to create gene cluster mutations and/or to genetically reconstitute the pathway in a heterologous host. Subsequent analytical and structural chemistry analyses of the compounds that are present in the reference strain but not in the mutant strains and of compounds that accumulate in the mutant strains but are absent or present in minute amounts in the reference strain may deliver insights that can be used for pathway elucidation.

Aspergillus niger is an industrially important filamentous fungus, which has obtained GRAS status for use in several industrial processes and is used for production of organic acids and enzymes. Importantly, when the full genome sequence of *A. niger* was examined, a gene cluster resembling the fumonisin gene



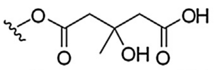
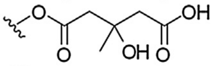
Compound	R	R'
Yanuthone A	OAc	◀OH, ····H
Yanuthone B	OAc	=O
Yanuthone C	OH	◀OAc , ····H
Yanuthone D (**1**)		=O
Yanuthone E (**2**)		◀OH, ····H
7-deacetoxyyanuthone A (**3**)	H	◀OH, ····H
22-deacetylyanuthone A (**6**)	OH	◀OH, ····H

Figure 1. Chemical Structures of 6-MSA and Previously Described Yanuthones

(A) Chemical structure of 6-MSA.

(B) Chemical structures of previously described yanuthones: yanuthones A–E, 7-deacetoxyyanuthone A, and 22-deacetylyanuthone A (Bugni et al., 2000; Li et al., 2003).

cluster from *Gibberella moniliformis* was surprisingly identified, suggesting that this well-characterized fungus has the genetic potential to produce the carcinogenic fumonisins (Baker, 2006). This possibility was later confirmed by genetic and chemical analyses (Pel et al., 2007; Frisvad et al., 2007). The fact that the *A. niger* genome contains several orphan gene clusters for production of secondary metabolites (Fisch et al., 2009) raises the question of whether it can produce other bioactive PKs that could be harmful, or perhaps beneficial, to human health. To this end, one silent cluster in *A. niger* was recently activated by expression of a transcription factor-encoding gene, which was embedded in the cluster. The resulting strain produced six azaphilone compounds, and further studies uncovered substantial new insights into the biosynthesis of this class of compounds (Zabala et al., 2012). It is interesting to note that among the 33 predicted PKS and PKS-like genes in *A. niger*, one encodes a putative PKS, which is phylogenetically close to fungal 6-methylsalicylic acid (6-MSA) synthases (Fisch et al., 2009). Importantly, the model PK 6-MSA (Wattanachaisaereekul et al., 2008) (Figure 1A) is known to be the precursor to, for example, the mycotoxin patulin (Beck et al., 1990) produced by many *Aspergillus* and *Penicillium* species, substantiating the possibility that this gene could be the source of yet another unknown bioactive PK in *A. niger*. Importantly, none of these 6-MSA-derived compounds have been observed in *A. niger* (Nielsen et al., 2009). We therefore investigated whether *A. niger* has the potential to produce 6-MSA or 6-MSA-derived compounds.

Known yanuthones constitute a group of compounds that are derived from a six-membered methylated ring (the C_7 core scaffold) with three side chains: one sesquiterpene and two varying side chains (-R and R') (Figure 1B). In this study we demonstrate that in *A. niger*, 6-MSA is the precursor for formation of yanuthone D, which is an antibiotic against *Candida albicans*, methicillin-resistant *Staphylococcus aureus* (MRSA), and vancomycin-resistant *Enterococcus* (Bugni et al., 2000). We also show that yanuthone D is in fact a complex meroterpenoid synthesized by a pathway where 6-MSA is decarboxylated, heavily oxidized, and fused to a sesquiterpene and a mevalon moiety (the di-acid of mevalonic acid). This is surprising, because yanuthones have been hypothesized to originate from the shikimate pathway (Bugni et al., 2000).

RESULTS

A. niger PKS48 Encodes a 6-MSA Synthase

To investigate the possibility that the *A. niger* gene PKS48/ASPNIDRAFT_44965 encodes a 6-MSA synthase, we transferred the gene to *A. nidulans*, which has not been shown to produce 6-MSA and which does not contain a close homolog to known 6-MSA PKSs. To ensure a high expression level on a defined medium, the PKS48 gene was integrated into a well characterized integration site, *IS1* (Hansen et al., 2011), under control of the strong constitutive promoter *PgpdA*. As expected, the metabolite profile obtained with an *Aspergillus nidulans* reference strain (IBT 29539) did not show any indications of 6-MSA when analyzed by ultra-high-performance liquid chromatography (UPHLC)-UV-visible diode array detector (DAD)-high-resolution time-of-flight mass spectrometry (TOFMS) (Figure 2A). In contrast, the metabolite profile of the strain expressing PKS48 showed the presence of a prominent new peak, which had the same retention time as an authentic 6-MSA standard and displayed the same adducts and monoisotopic mass for the pseudomolecular ion. We therefore conclude that PKS48 encodes a 6-MSA synthase.

Production of Yanuthones D and E Is Eliminated by Deletion of PKS48

The fact that 6-MSA has not previously been reported from *A. niger* prompted us to investigate whether this compound could be a precursor to a known secondary metabolite produced by this fungus. We therefore cultivated an *A. niger* reference strain (KB1001) and an *A. niger* PKS48Δ strain on four different solid media (minimal medium [MM], CYA, YES, and MEA) that are known to trigger the production of a wide range of metabolites (Nielsen et al., 2011). The resulting UHPLC-DAD-TOFMS metabolite profiles were almost identical (Figure S1 available online), showing that the PKS48Δ mutation did not induce a global response on the secondary metabolism. However, on YES and MM media, we identified two compounds that were produced by KB1001, but not by the PKS48Δ strain (Figures 2B and 2C; Table S1). UHPLC separation with UV-visible and

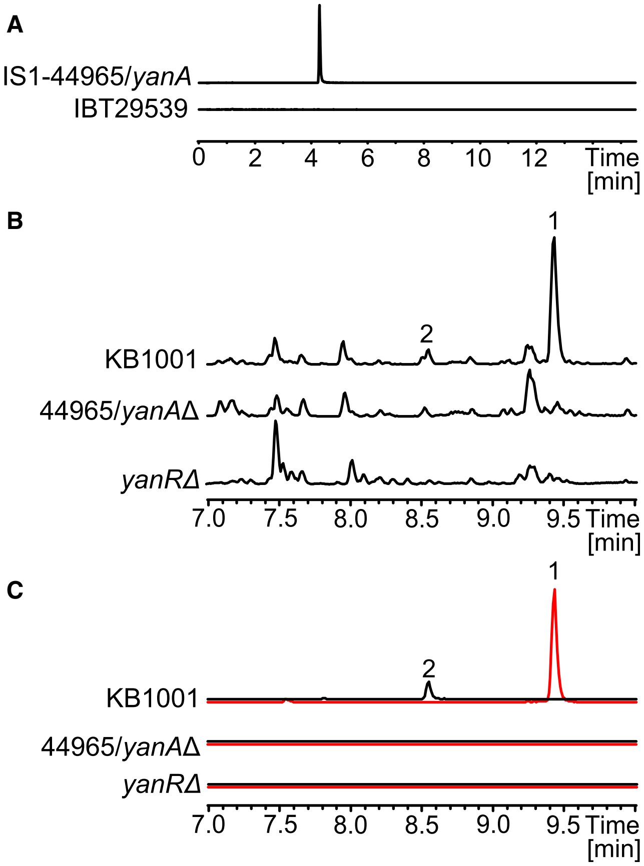


Figure 2. Extracted Ion Chromatograms

(A) Extracted ion chromatogram (EIC, *m/z* 153.0546 ± 0.005) of an *A. nidulans* reference strain (IBT 29539) and a 6-MSA producing strain (IS*1*-44965/*yanA*).

(B) Base peak chromatograms (BPC) *m/z* 100-1,000 of the *A. niger* reference (KB1001), *yanA*Δ, and *yanR*Δ strains.

(C) EICs of yanuthone D (**1**) 503.2640 ± 0.005 (red) and yanuthone E (**2**) 505.2791 ± 0.005 (black) for KB1001, *yanA*Δ, and *yanR*Δ.

All chromatograms are to scale.

high-resolution MS detection as well as MS/MS suggested that the two compounds were yanuthones D and E. This was confirmed by isolation of the compounds, nuclear magnetic resonance (NMR) spectroscopy, and circular dichroism (CD) (Tables S3 and S4). Hence, production of yanuthones D and E appears to be based on the use of 6-MSA as a key precursor. In this scenario, one carbon must be eliminated from C_8-based 6-MSA to form the C_7 core scaffold of yanuthones D and E.

Yanuthones Constitute a Complex Group of Compounds That Appear to Originate from Different Precursors

In addition to yanuthones D and E, *A. niger* has previously been reported to produce yanuthones A, B, and C, 1-hydroxyyanuthone A, 1-hydroxyyanuthone C, and 22-deacetylyanuthone A (Bugni et al., 2000), and 7-deacetoxyyanuthone A has been reported from the genus *Penicillium* (Li et al., 2003) (Figure 1B). We thus examined the extracted ion chromatograms from the UHPLC-DAD-TOFMS profiles obtained by KB1001 for the presence of these metabolites. In extracts obtained after cultivation on MM, YES, and CYA media, this analysis identified trace amounts of a compound (yanuthone X_1) with a mass and elemental composition corresponding to the yanuthone isomers A and C. The nature of this compound was further investigated by MS/MS, and its fragmentation pattern was similar to the pattern of other yanuthones, showing characteristics such as loss of a sesquiterpene chain. Moreover, the UV-visible spectrum of the compound was similar to spectra obtained for yanuthones D and E, substantiating that this compound was a yanuthone. Surprisingly, when the UHPLC-DAD-TOFMS metabolite profiles obtained with the PKS48Δ strain were examined for the presence of this yanuthone, it was still present. This observation strongly suggested that some yanuthones are produced independently of PKS48.

Fully Labeled $^{13}C_8$-6-MSA Is Incorporated into Yanuthones D and E In Vivo

The fact that some yanuthones could be produced independently of PKS48, combined with the fact that yanuthones have been proposed to originate from the shikimate pathway, raised the possibility that the absence of yanuthones D and E in the PKS48 deletion strain potentially could be the result of an indirect effect. To investigate this possibility, we fed fully labeled $^{13}C_8$-6-MSA to KB1001 and the PKS48Δ strain at different time points during growth (24, 48, and 72 hr; see Experimental Procedures). The addition of $^{13}C_8$-6-MSA did not seem to adversely affect the growth rate, and the morphologies of the colonies of the two strains were identical (Figure S2). This indicates that the amounts of $^{13}C_8$-6-MSA added (2-10 μg/ml) did not significantly influence strain fitness. Metabolites were then extracted from the plates and analyzed by UHPLC-DAD-TOFMS. For both strains, $^{13}C_8$-6-MSA was incorporated into yanuthones D and E, resulting in a mass shift of 7.023 Da. This is in agreement with the scenario described above, where one carbon atom must be eliminated from 6-MSA in the biosynthetic processing toward yanuthones D and E. Moreover, the MS-based metabolite profiles also showed that $^{13}C_8$-6-MSA was exclusively incorporated into compounds related to yanuthones. These compounds are only present in tiny amounts and are likely intermediates or analogs of yanuthone D or E, because they share the same UV chromophore and because their masses corresponded to water loss(es) or gain from yanuthone D or E. Based on these results, we named the 6-MSA synthase (encoded by PKS48/ASPNIDRAFT_44965) YanA (*yan*uthone) and the corresponding gene *yanA.* On the other hand, no labeled yanuthone X_1 was observed in KB1001 as well as in the PKS48 deletion strain after addition of $^{13}C_8$-6-MSA (mass spectra are shown in Figure S3), confirming our finding that yanuthone X_1 is formed in the absence of PKS48. Hence, we conclude that 6-MSA is not the precursor of yanuthone X_1.

The *yan* Gene Cluster Comprises Ten Genes

To determine whether *yanA* defines a gene cluster for a biosynthetic pathway toward yanuthones D and E, ten genes up- and downstream of *yanA* were annotated using FGeneSH (Softberry) and AUGUSTUS software (Stanke and Morgenstern, 2005). Subsequently, these twenty putative genes were examined using the NCBI Conserved Domain Database (Marchler-Bauer et al., 2011) for open reading frames (ORFs) encoding activities that are typically employed for the modification of PKs. Based on these

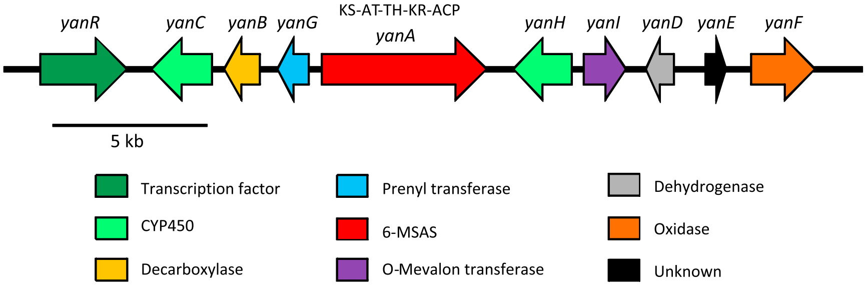


Figure 3. The Proposed *yan* Cluster

The *yanA* 6-MSA synthase-encoding gene is flanked by nine cluster genes (*yanB*, *yanC yanD*, *yanE*, *yanF*, *yanG*, *yanH*, *yanI*, and *yanR*) whose products contain all necessary activities for conversion of 6-MSA into yanuthone D.

analyses, eight additional genes could potentially belong to the *yanA* cluster, including genes encoding a transcription factor (TF), a prenyl transferase, an O-acyltransferase, a decarboxylase, two oxidases, two cytochrome P450s (CYP450s), and a dehydrogenase (Figure 3; Table S2). Together with *yanA* and 192604 (a gene with no known homologs), these eight genes form a cluster of ten genes that are not interrupted by any of the remaining eleven genes included in the analysis. The fact that one of the ten genes in this cluster (44961) putatively encodes a TF raised the possibility that expression of the genes involved in yanuthones D and E production is controlled by this TF. In agreement with this view, deletion of 44961 resulted in a strain that did not produce these two yanuthones (Figures 2B and 2C). To further delineate the *yanA* gene cluster, we determined the expression levels of the ten cluster genes as well as of four flanking genes by RT-quantitative PCR (qPCR) in a 44961Δ strain and KB1001. When the two data sets were compared, we found, as expected, that expression from 44961 is eliminated in the 44961Δ strain where the entire gene is deleted (Figure S4). More importantly, the analysis demonstrated that expression from the other nine genes in the cluster was significantly downregulated in the 44961Δ strain as compared to KB1001 (p value < 0.05). Specifically, the expression was reduced more than 10-fold for seven of the genes, including *yanA*. Expression of the remaining two genes, 54844 and 44964, was expressed at a level corresponding to 20% and 11%, respectively, of the level obtained with KB1001. In contrast, expression levels from the four flanking genes were not significantly different from KB1001 (Figure S4). Next, we individually deleted the remaining eight genes in the proposed *yan* gene cluster, which encode putative activities for PK modification. None of the resulting strains, including 192604Δ, produced yanuthone D, indicating that all genes belong to the *yan* cluster (Table S1). As a control, the four additional genes flanking this cluster were also individually deleted, but all these four strains produced yanuthone D. Based on these analyses and the results from the RT-qPCR, we propose that the *yan* gene cluster is composed by 10 genes, *yanA*, *yanB*, *yanC*, *yanD*, *yanE*, *yanF*, *yanG*, *yanH*, *yanI*, and *yanR*, where *yanR* encodes a TF that regulates the gene cluster (Figure 3; Table S2). Finally, all ten genes were simultaneously deleted in one strain. When $^{13}C_8$-6-MSA was fed to this strain, no labeled metabolites were detected, showing that all 6-MSA-derived yanuthones depend on this gene cluster (see above).

YanF Converts Yanuthone E into Yanuthone D

As the first step toward elucidating the order of reaction steps in the pathway toward yanuthones D and E, we asked whether yanuthones D and E are two different end products or whether one is an intermediate in the pathway toward production of the other. To this end, we note that individual deletion of genes in the *yan* gene cluster generally resulted in loss of production of both yanuthones D and E on YES medium. The only exception is the *yanF*Δ strain, which produced substantial amounts of yanuthone E (**2**), but no yanuthone D (**1**) (Figure 4). These findings suggest that YanF converts yanuthone E into yanuthone D, which is the true end product of the pathway. Interestingly, the *yanF*Δ strain produced a new and unknown compound, which was not detected in KB1001. Elucidation of its structure revealed a yanuthone E analog with a hydroxylation at C-2 at the expense of the first double bond (between C-2 and C-3) in the sesquiterpene moiety (Table S4). This compound was named yanuthone J (**9**).

m-Cresol and Toluquinol Are Intermediates of the Yanuthone D Biosynthesis

Deletion of *yanB*, *yanC*, *yanD*, *yanE*, and *yanG* did not produce any detectable intermediates, and the phenotype of these mutations therefore does not link any of the genes to specific reaction steps in the pathway toward formation of yanuthone D. However, one of the five putative enzymes, YanC, has a defined homolog, PatI, in the *Aspergillus clavatus* patulin biosynthesis pathway (Artigot et al., 2009) where it catalyzes the oxidation of *m*-cresol into toluquinol, suggesting that toluquinol and *m*-cresol are also likely intermediates in the yanuthone biosynthesis. To test this hypothesis, we fed *m*-cresol and toluquinol to the *yanA*Δ strain. Analysis of the metabolite profiles of the two strains indeed showed that addition of *m*-cresol or toluquinol restored production of yanuthones D and E in the *yanA*Δ strain (Figure 5).

In an attempt to further elucidate the role of the five enzymes, the corresponding genes were inserted into plasmid pDHX2 (Figure S5) and individually expressed in the *A. nidulans* strain harboring the *yanA* gene. No new compounds were produced in these *IS1-yanA* strains expressing *yanC*, *yanD*, *yanE*, and *yanG*, despite the fact that 6-MSA was produced in high amounts (Figure S6). Similarly, in the strain expressing *yanB*, no new product was observed, but in this case 6-MSA was absent, indicating that 6-MSA is a substrate for YanB.

Deletion of *yanI* and *yanH* Reveals Key Intermediates in the Biosynthesis of Yanuthone D

In contrast to the *yanB*Δ-*E*Δ and *yanG*Δ strains, new products were observed in the *yanH*Δ and *yanI*Δ strains. Deletion of *yanH* resulted in a strain where the most prominent compound accumulating is 7-deacetoxyyanuthone A (**3**) (NMR data in Table S4). Interestingly, we also identified two compounds in this strain (Figure 4). Isolation and structure elucidation revealed two C-1 oxidized yanuthone derivatives, which we named

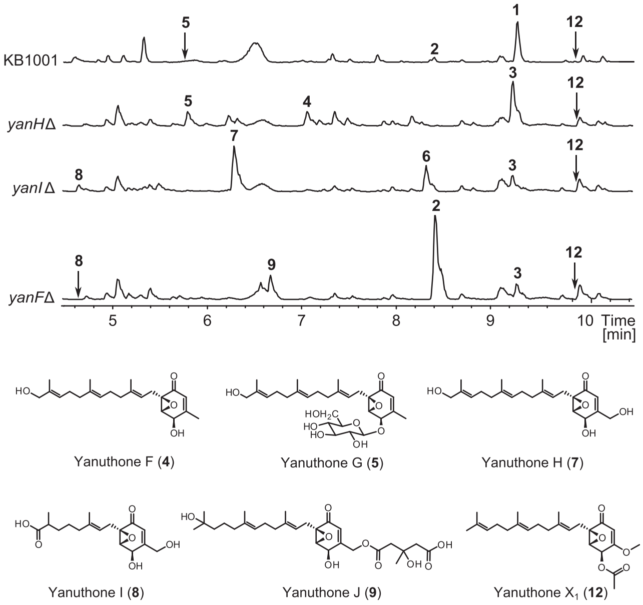


Figure 4. BPC *m/z* 100–1,000 of Reference Strain KB1001, *yanH*Δ, *yanI*Δ, and *yanF*Δ

All NMR-elucidated compounds are shown for comparison of intensity and relative retention times. Below are structures of the yanuthones identified in this study. The structures of yanuthone D (**1**), yanuthone E (**2**), 7-deacetoxyyanuthone A (**3**), and 22-deacetylyanuthone A (**6**) are shown in Figure 1.

yanuthone F (**4**) and yanuthone G (**5**) (NMR data in Table S4). Yanuthone G (**5**) is a glycosylated version of yanuthone F (**4**), which can also be detected in trace amounts in KB1001 (Table S1). Deletion of *yanI* resulted in a strain producing the known compounds 7-deacetoxyyanuthone A (**3**) and 22-deacetylyanuthone A (**6**) (NMR data in Table S4; Figure 1B). Importantly, the latter compound corresponds to yanuthone E (**2**) without the mevalon moiety. In addition, two compounds were produced. The structures were elucidated by NMR spectroscopy, revealing that one, which we named yanuthone H (**7**), is very similar to 22-deacetylyanuthone A (**6**), but with a hydroxyl group at C-1 (Figure 4; Table S4). The other compound, which we named yanuthone I (**8**), is a modification of 22-deacetylyanuthone A (**6**) with a shorter and oxidized terpene (NMR data in Table S4). We note that yanuthone I (**8**) was also detected in trace amounts in KB1001 (Table S1).

Determination of the Yanuthone X_1 Structure

As mentioned above, yanuthone X_1 (**12**) has an elemental composition corresponding to yanuthone A and C but was biosynthesized from another precursor than yanuthone D and E. We therefore isolated and elucidated the structure (Figure 4; Table S4). This analysis confirmed that yanuthone X_1 (**12**) does not have the same C_7 core scaffold but instead has a C_6 core with a methoxygroup directly attached to the six-membered ring at the expense of a methyl group (Figure 4). Despite the fact that yanuthone X_1 (**12**) and yanuthones D and E employ different precursors, they share common features like the epoxide and the sesquiterpene side chain, and we therefore hypothesized that they share common enzymatic steps during their biosynthesis. In agreement with this, examination of the metabolite profiles obtained with the *yan* gene deletion strains revealed that yanuthone X_1 (**12**) was absent in the *yanC*, *yanD*, *yanE*, and *yanG* deletion strains (Table S1). In contrast, yanuthone X_1 (**12**) is produced in larger amounts in the *yanA*Δ strain, which cannot produce 6-MSA.

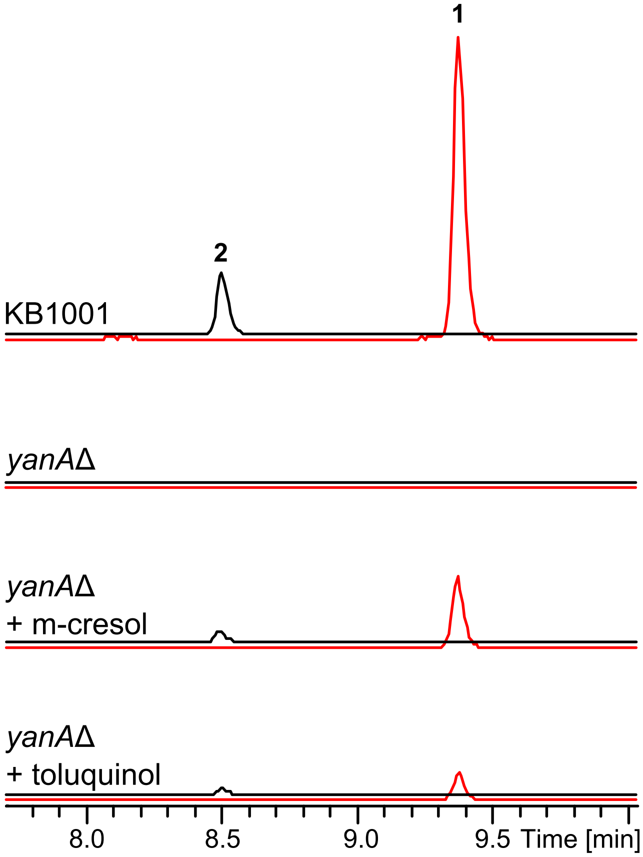


Figure 5. Feeding with Unlabeled *m*-cresol and Toluquinol

Shown are EICs of yanuthone D (**1**) 503.2640 ± 0.005 (red) and yanuthone E (**2**) 505.2791 ± 0.005 (black) for KB1001 and the *yanA*Δ strain with and without feeding. Chromatograms are to scale.

Antifungal Activity of Yanuthones

Yanuthones have earlier been reported to display antimicrobial activity (Bugni et al., 2000), and we therefore tested all ten yanuthones presented in this study for antifungal activity toward *C. albicans* (Table 1). Among these compounds, our analysis identified yanuthone D as the most toxic species in agreement with the fact that it represents the most likely end point of the pathway. Among the remaining yanuthones, three other species, yanuthone G, yanuthone H, and 22-deacetylyanuthone A, exhibited antimicrobial activity. In these cases, IC_{50} values were ~5- to 10-fold higher than the IC_{50} value determined for yanuthone D.

DISCUSSION

Elucidation of the Biosynthetic Route from 6-MSA toward Yanuthone D

We have used a combination of bioinformatics, genetic tools, chemical analyses, and feeding experiments to investigate

Table 1. The Half-Maximal Inhibitory Concentration for *C. albicans* Treated with a Small Library of Yanuthones

Compound	Origin	Isolate	IC_{50} (μM)
Yanuthone D	*A. niger*	KB1001	3.3 ± 0.5
Yanuthone E	*A. niger*	KB1001	>100
Yanuthone F	*A. niger*	*yanH⊿*	>100
Yanuthone G	*A. niger*	*yanH⊿*	38.8 ± 5.1
Yanuthone H	*A. niger*	*yanI⊿*	24.5 ± 1.1
Yanuthone I	*A. niger*	*yanI⊿*	>100
Yanuthone J	*A. niger*	*yanF⊿*	>100
7-deacetoxyyanuthone A	*A. niger*	KB1001	>100
22-deacetylyanuthone A	*A. niger*	KB1001	19.4 ± 1.8
Yanuthone X_1	*A. niger*	KB1001	>100

The IC_{50} values were calculated based on duplicate experiments carried out in three independent trials and annotated with their respective SD.

whether 6-MSA is produced and whether it is used for production of toxic secondary metabolites in *A. niger*. Our work demonstrates that 6-MSA is synthesized by the YanA PKS and then subsequently modified into the antimicrobial end product yanuthone D. This is intriguing because yanuthones have previously been suggested to originate from shikimic acid (Bugni et al., 2000). Yanuthones have previously been observed on YES agar (Klitgaard et al., 2014; Nielsen et al., 2009) and a mixture of yeast, beef, and casein extract (Bugni et al., 2000). In this study yanuthones were detected on solid YES and MM medium, but not on solid CYA or MEA medium, and yanuthone synthesis is therefore conditionally induced. To this end, we find that yanuthone D is not produced in liquid YES and MM medium, in agreement with the fact that secondary metabolism is generally turned off in submerged cultures (González, 2012; Schachtschabel et al., 2013).

We have also shown that *yanA* defines a gene cluster of ten members: *yanA*, *yanB*, *yanC*, *yanD*, *yanE*, *yanF*, *yanG*, *yanH*, *yanI*, and *yanR*, which is regulated by YanR. In agreement with this, YanR is homologous to Zn_2Cys_6 transcription factors that are commonly involved in regulation of secondary metabolite production. The fact that deletion of *yanR* completely abolished production of yanuthone D suggests that YanR acts as an activator of the *yan* cluster. Additionally, analyses of strains where the remaining genes in the *yan* cluster were individually deleted have allowed us to isolate and characterize the full structures of three intermediates. Based on these compounds, we propose the entire pathway for yanuthone D formation including addition of a sesquiterpene and a mevalon to the core polyketide moiety at different stages of the biosynthesis (Figure 6).

In our model, the last intermediate in the pathway is yanuthone E (**2**), which is converted into the end product yanuthone D (**1**) by oxidation of the hydroxyl group at C-15 in a process catalyzed by YanF. The fact that yanuthone E (**2**) is present in KB1001 indicates that it may act as a reservoir for rapid conversion into the more potent antibiotic compound yanuthone D. Yanuthone E (**2**) is likely formed from 22-deacetylyanuthone A (**6**) by attachment of mevalon to the hydroxyl group at C-22. Because 22-deacetylyanuthone A (**6**), but not yanuthone E (**2**), accumulates in the *yanI*Δ strain, we propose that YanI, a putative O-acyltransferase, catalyzes this step. Intriguingly, YanI therefore appears to be an O-mevalon transferase, an activity, which, to the best of our knowledge, has not previously been described in the literature. Next, we propose that 22-deacetylyanuthone A (**6**) is formed by hydroxylation of C-22 of 7-deacetylyanuthone A (**3**). In agreement with this view, 7-deacetylyanuthone A (**3**), but not 22-deacetoxyyanuthone A (**6**), accumulates in the absence of YanH.

Unfortunately we did not detect any intermediates leading from 6-MSA to 7-deacetoxyyanuthone A (**3**) in any of the deletion strains in *A. niger*. The remaining tentative steps in the pathway were therefore deduced from bioinformatics and feeding experiments. First, analyses of patulin formation in *Aspergillus floccosus* (previously identified as *Aspergillus terreus*; Jens C. Frisvad, personal communication) and in *A. clavatus* have shown that it requires decarboxylation of 6-MSA into *m*-cresol (Artigot et al., 2009; Puel et al., 2010). This step is catalyzed by 6-MSA decarboxylase (Light, 1969), which has been proposed to be encoded by *patG* (Puel et al., 2010). *m*-Cresol is then converted into gentisyl alcohol in two consecutive hydroxylation steps catalyzed by the two cytochrome P450s CYP619C3 (PatH) and CYP619C2 (PatI). However, CYP619C2 may also act directly on *m*-cresol to form the co-metabolite toluquinol, which is not an intermediate toward patulin. When we inspected the *yan* gene cluster for similar activities, we found a putative 6-MSA decarboxylase (YanB) and CYP619C2 (YanC), but not CYP619C3. These observations suggest that *m*-cresol and toluquinol are intermediates in yanuthone D formation. We present two lines of evidence in support of this view. First, our feeding experiments demonstrate that both compounds can be converted into yanuthone D. Second, heterologous expression of *yanA* in *A. nidulans* leads to production of 6-MSA. This compound disappears if the strain also expresses *yanB*, indicating that 6-MSA is a substrate for the putative 6-MSA decarboxylase YanB. Together these results strongly suggest that *m*-cresol is formed directly from 6-MSA by a decarboxylation reaction, which is most likely catalyzed by YanB. This reaction explains how C_8-based 6-MSA can serve as the building block for the C_7-based core unit of yanuthones. Moreover, the analyses show that toluquinol is an intermediate in the production of yanuthone D and that it is formed from *m*-cresol in a process most likely catalyzed by the putative cytochrome P450 encoded by *yanC.* Conversion of toluquinol into 7-deacetylyanuthone A (**3**) requires epoxidation and prenylation. Based on the fact that prenylated toluquinol is never observed in KB1001 or mutant strains, we propose that epoxidation precedes prenylation. In this scenario, toluquinol is epoxidated into (**10**), which is in equilibrium with the tautomer (**11**). This compound (**11**) is then prenylated to form 7-deacetylyanuthone A (**3**) as a sesquiterpene moiety is attached to C-13 of (**11**). The latter reaction is likely catalyzed by YanG, a putative prenyltransferase. This is supported by the observation that yanuthone D (**1**) and all detectable intermediates, including 7-deacetoxyyanuthone A (**3**), were absent in the *yanG*Δ strain. The identity of the gene products(s) responsible for epoxidation of toluquinol is less clear. Among the putative activities encoded by the genes in the *yan* cluster, which have not been assigned to any reaction step during the analyses above, we note the presence of a putative dehydrogenase (YanD) and one with an unknown activity and with no obvious homologs (YanE) as judged by BLAST

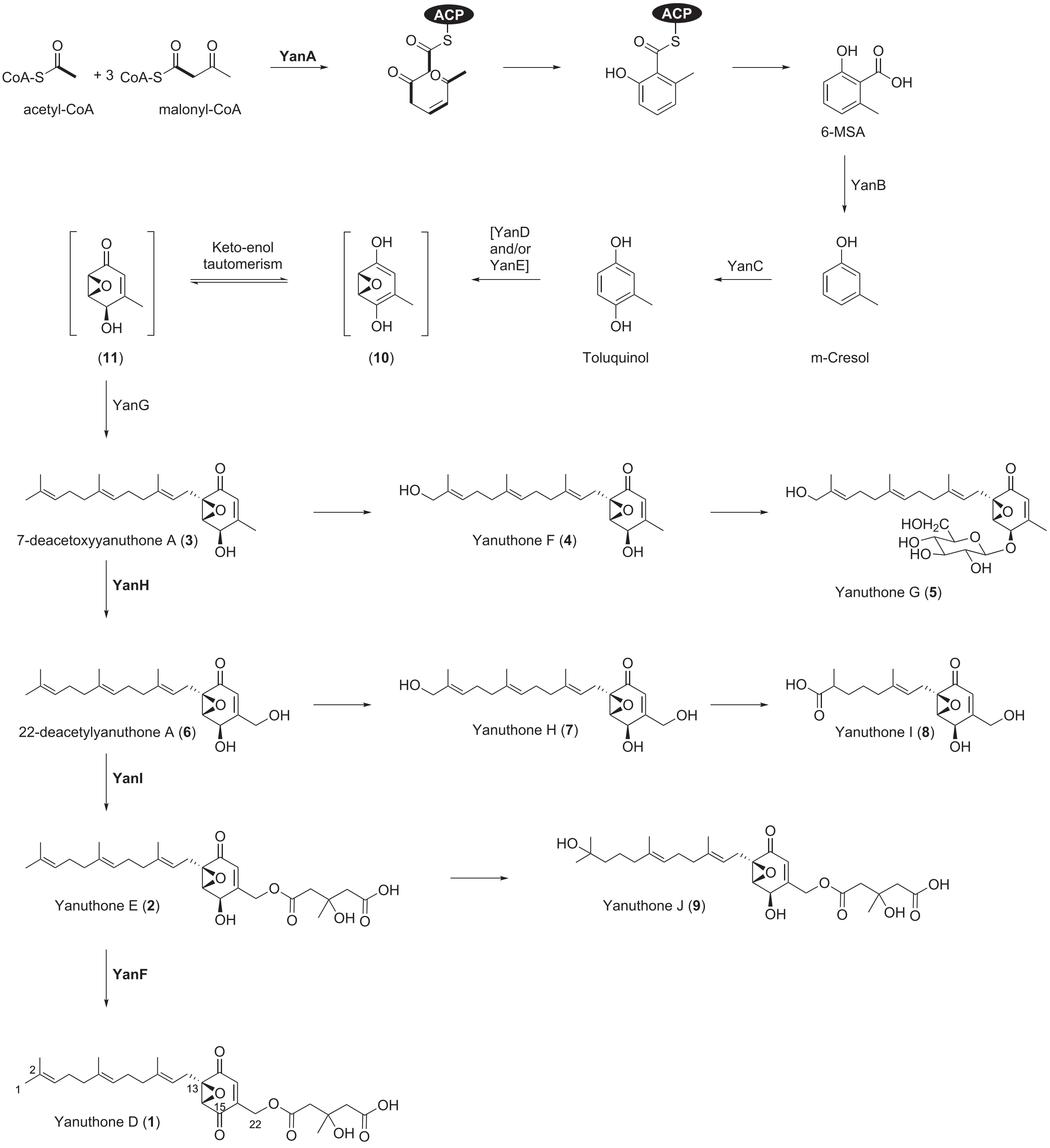


Figure 6. Proposed Biosynthesis of yanuthone D

Structures and enzymatic activities in brackets are hypothesized, activities in plain text have been proposed from bioinformatics, and activities in bold have been experimentally verified.

analysis of the GenBank database (Altschul et al., 1990). We hypothesize that one or both of these enzymes catalyze epoxidation. The fact that neither 6-MSA, *m*-cresol, toluquinol, nor any other intermediates were detected in the *yanB*Δ, *yanC*Δ, *yanD*Δ, and *yanE*Δ strains suggests that these small, aromatic compounds must be rapidly degraded or converted into other compound(s), or they may be incorporated into insoluble material, e.g., the cell wall.

Accumulation of Intermediates in the Yanuthone D Pathway Triggers Formation of Novel Yanuthones

Disruption of the biosynthetic pathway toward yanuthone D results in formation of three branch points in the pathway toward yanuthone D: at yanuthone E (**2**), at 7-deacetoxyyanuthone A (**3**), and at 22-deacetylyanuthone A (**6**). In addition to yanuthone E (**2**), yanuthone J (**9**) accumulates in the *yanF*Δ strain. Similarly, yanuthone F (**4**) accumulates in addition to 7-deacetoxyyanuthone A (**3**) in the *yanH*Δ strain, and yanuthone H (**7**) accumulates in addition to 22-deacetylyanuthone A (**6**) in the *yanI*Δ strain. In all cases, the sesquiterpenes of the accumulated intermediates in the main pathway are oxidized at C-1 or C-2. Because hydroxylation is a known detoxification mode, we speculate that the abnormally high amount of potentially toxic intermediates 7-deacetoxyyanuthone A (**3**), 22-deacetylyanuthone A (**6**), and yanuthone E (**2**) triggers the cell to initiate phase I type of detoxification processes in which the toxic intermediates are hydroxylated. This hypothesis is supported by the fact that there is no obvious assignment of an enzyme with this activity, encoded by the *yan* gene cluster, and by the fact that one of the intermediates, 22-deacetylyanuthone, is toxic to *C. albicans*.

An additional variant of yanuthone F (**4**) was identified in the *yanH*Δ strain, in which yanuthone F (**4**) is glycosylated at the hydroxyl group at C-15 to form yanuthone G (**5**). The glucose moiety of yanuthone G (**5**) is intriguing because sugar moieties are rare in fungal secondary metabolites, and the fact that yanuthone G (**5**) is detected in KB1001 shows that it is a naturally occurring compound (Figure 4; Table S1). Because yanuthone G (**5**) production is upregulated in *yanH*Δ, we suggest that glycosylation poses a second (phase II conjugation) type of mechanism for further detoxification of possible toxic intermediates.

The branch point at 22-deacetylyanuthone A (**6**) revealed a novel compound yanuthone I (**8**), which is identical to 22-deacetylyanuthone A (**6**) and yanuthone H (**7**) but with a shorter and oxidized sesquiterpene chain. A similar modification has been observed in the biosynthetic pathway for production of mycophenolic acid (Regueira et al., 2011). Here it was proposed to occur by oxidative cleavage between C-4 and C-5 of the sesquiterpene chain. Alternatively, it could occur by terminal oxidation of a geranyl side chain.

Yanuthone X_1 Defines a Novel Class of Yanuthones

Because yanuthones are based on a C_7 scaffold, they were previously proposed to originate from shikimic acid (Bugni et al., 2000). However, in our study we demonstrate that yanuthones D and E originate from the C_8 polyketide precursor 6-MSA, which is decarboxylated to form the C_7 core of the yanuthone structure. In contrast, the novel yanuthone X_1 (**12**) has a C_6 core scaffold that does not originate from 6-MSA and does not require decarboxylation by YanB. Based on this we define two classes of yanuthones: those that are based on the polyketide 6-MSA, class I, and those that are based on the yet unknown precursor leading to the formation of yanuthone X_1 (**12**), class II. The two classes of yanuthones share several enzymatic steps. First we note that the sesquiterpene side chain in yanuthone X_1 (**12**) is likely attached by YanG, as is the case for yanuthone D. Second, it depends on enzyme activities of YanC, YanD, and YanE, but not of YanB. Together this suggests that the precursor is a small aromatic compound similar to 6-MSA but lacking the carboxylic acid. Importantly, the main difference between yanuthone D and yanuthone X_1 (**12**) are the groups attached to C-16. In the case of yanuthone X_1 (**12**), this position is oxidized, whereas in yanuthones D and E there is a carbon-carbon bond that originates from the methyl group of 6-MSA. Consequently, yanuthone X_1 (**12**) cannot be mevalonated by YanI.

SIGNIFICANCE

This study has identified a cluster of 10 genes, which is responsible for production of antimicrobial yanuthone D in *A. niger*. We show that yanuthone D is based on the polyketide 6-MSA and not on shikimic acid as previously suggested, and we have proposed a detailed genetic and biochemical pathway for converting 6-MSA into yanuthone D. Interestingly, we have revealed that yanuthone X_1, although similar in structure, is not derived from 6-MSA, but the yet unknown precursor to yanuthone X_1 does employ several enzymes encoded by the *yan* cluster. An important finding in the elucidation of the biosynthesis is the identification of *yanI* encoding an O-mevalon transferase, which represents a different enzymatic activity. We have discovered that the pathway toward yanuthone D branches when intermediates accumulate, because three intermediates are hydroxylated. Two of the hydroxylated compounds are further modified by oxidative cleavage of the sesquiterpene and glycosylation, respectively, resulting in five yanuthones. The discovery of a glycosylated compound, yanuthone G, is intriguing because glycosylated compounds are very rare in fungal secondary metabolism. We successfully employed an interdisciplinary approach for solving the biosynthetic pathway: applying gene deletions, heterologous gene expression, UHPLC-DAD-MS, MS/MS, structural elucidation by NMR spectroscopy and CD, and feeding experiments with ^{13}C-labeled and unlabeled metabolites. Together, our analyses have cast insights into understanding the complexity of fungal secondary metabolism.

EXPERIMENTAL PROCEDURES

Strains and Media

The strain IBT 29539 was used for strain constructions in *A. nidulans*. ATCC1015-derived strain KB1001 was used for strain constructions in *A. niger*. All fungal strains prepared in the present work (Table S5) have been deposited in the IBT Culture Collection at the Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark. *Escherichia coli* strain DH5α was used for propagating plasmids, except *E. coli* ccdB survival2 cells (Invitrogen), which were used for plasmids carrying the *ccdB* gene.

MM for *A. nidulans* was made as described by Cove (1966), but with 1% glucose, 10 mM $NaNO_3$, and 2% agar. MM for *A. niger* was prepared as described by Chiang et al. (2011). YES, MEA, and CYA were prepared as described by Frisvad and Samson (Samson et al., 2010). When necessary, media were supplemented with 4 mM L-arginine, 10 mM uridine, 10 mM uracil, and/or 100 μg/ml hygromycin B (InvivoGen). Luria-Bertani (LB) medium was used for cultivation of *E. coli* strains and consisted of 10 g/l tryptone (Bacto), 5 g/l yeast extract (Bacto), and 10 g/l NaCl (pH 7.0). When necessary, LB was supplemented with 100 μg/ml ampicillin.

For batch cultivation the medium contained 20 g/l D-glucose-$^{13}C_6$ (99 atom % ^{13}C; Sigma-Aldrich) or D-glucose, 7.3 g/l $(NH_4)_2SO_4$, 1.5 g/l KH_2PO_4, 1.0 g/l $MgSO_4 \cdot 7\ H_2O$, 1.0 g/l NaCl, 0.1 g/l $CaCl_2$, 0.1 ml of Antifoam 204

(Sigma), and 1 ml/l trace element solution. Trace element solution contained 0.4 g/l $CuSO_4 \cdot 5\ H_2O$, 0.04 g/l $Na_2B_2O_7 \cdot 10\ H_2O$, 0.8 g/l $FeSO_4 \cdot 7\ H_2O$, 0.8 g/l $MnSO_4 \cdot H_2O$, 0.8 g/l $Na_2MoO_4 \cdot 2\ H_2O$, and 8.0 g/l $ZnSO_4 \cdot 7\ H_2O$.

Construction of Basic Vectors for Strain Construction

All primers are listed in Table S6. The PfuX7 polymerase (Nørholm, 2010) was used in all PCRs. Fragments were assembled via uracil-excision fusion (Geu-Flores et al., 2007) into a compatible vector.

pDH56 and pDH57 are designed for integration of novel genes into the *IS1* site of *A. nidulans*. In pDH56, the AsiSI/Nb.BtsI uracil-excision cassette of CMBU1111 (Hansen et al., 2011) is modified into a *ccdB-cm*^{R} AsiSI/Nb.BtsI uracil-excision cassette. Unlike with CMBU1111, new fragments can be introduced into this cassette and cloned in *ccdA*-deficient *E. coli* strains like DH5α without generating background because false positives resulting from incomplete digestion of the USER cassette are eliminated (Bernard and Couturier, 1992). Specifically, the suicide gene *ccdB* and the chloramphenicol resistance gene *cm*^{R} (*ccdB-cm*^{R}) construct was PCR amplified (using primers 84 and 85) from pDONR (Invitrogen) and inserted into CMBU1111 by uracil-excision cloning in a manner that reconstituted the original uracil excision cassette on either side of the insert. pDH57 was constructed from pDH56 by removing an undesirable Nb.BtsI nicking site located in the *amp*^{R} gene. pDH56 was PCR amplified in two pieces (81 + 76 and 75 + 80). 75 and 76 were designed to introduce a silent mutation into the Nb.BtsI recognition site. Fragments were assembled via uracil-excision cloning, and correct clones were verified by sequencing. The gene targeting substrate for insertion of the 6-MSA synthase gene *yanA* was made by amplifying the synthase gene *yanA* (PKS48/ASPNIDRAFT_44965) from IBT 29539 genomic DNA (primers 1 and 2) and inserted into pDH57, yielding pDH57-*yanA*.

The pDHX2 vector is AMA1-based and designed for episomal gene expression. pDHX2 was constructed by USER fusion of five fragments: (1) *E. coli* origin of replication (*ori*^{R}) and the *E. coli* ampicillin resistance gene (*amp*^{R}); (2) the 5′ half of AMA1; (3) the 3′ half of AMA1; (4) 0.5 kb of the P*gpdA* promoter, an AsiSI/Nb.BtsI USER cassette containing *ccdB* and *cm*^{R}, and the TtrpC terminator; and (5) the *A. fumigatus pyrG* selection marker (Figure S5). Fragment 1 was amplified from pDH57 (primers 77 + 78); fragments 2, 3, and 5 were amplified from pDEL2 (primers 86 + 89, 87 + 88, and 82 + 83) (Nielsen et al., 2008); and fragment 4 was amplified from pDH57 (primers 79 + 80). Fragments were assembled as described by Geu-Flores et al. (2007) using equal molar amounts of purified PCR product, and correct clones were verified by restriction digestion. Plasmids for episomal heterologous expression of cluster genes were constructed by PCR amplification of ORFs using primers 3–12 pairwise. Genes were inserted into AsiSI/Nb.BtsI-digested pDHX2 as described by Nour-Eldin et al. (2006), resulting in pDHX2-*yanB*, pDHX2-*yanC*, pDHX2-*yanD*, and pDHX2-*yanE*. Plasmids were verified by sequencing.

Plasmids carrying gene targeting substrates for gene deletion in *A. niger* were constructed by PCR amplification of upstream (US) and downstream (DS) targeting sequences along with the *hph* marker, conferring resistance to hygromycin B. US and DS targeting sequences were generated using the primers 17–72, and *hph* was amplified from pCB1003 (McCluskey et al., 2010) using primers 13 + 14. The three fragments were assembled into the CMBU0020 vector (Hansen et al., 2011).

Strain Construction

Protoplasting and gene-targeting procedures were performed as described previously for *A. nidulans* (Johnstone et al., 1985; Nielsen et al., 2006) and *A. niger* (Chiang et al., 2011). NotI-linearized pDH57-*yanA* was transformed into IBT 29539. Transformants were verified by diagnostic PCR as described by Hansen et al. (2011).

Strains for episomal expression of cluster genes were constructed by transforming the IS1-yanA strain with circular plasmids (pDHX2-*yanB*, pDHX2-*yanC*, pDHX2-*yanD*, and pDHX2-*yanE*) using *pyrG* as a selectable marker.

A. niger deletion strains were constructed by transforming KB1001 with bipartite gene targeting substrates. The substrates were generated by PCR amplification of the US::*hph*::DS cassettes of the CMBU0020-based plasmids using primers GENE_US-FW+73 and 74+GENE_DS-RV. Deletion strains were selected on 100 μg/ml hygromycin B and verified by diagnostic PCR.

RNA Extraction and RT-qPCR

RNA isolation from the *A. niger* strains and subsequent quantitative RT-PCRs were done as previously described by Hansen et al. (2011) except that biomass for RNA isolation was prepared with a Tissue-Lyser LT (QIAGEN) by treating samples for 1 min at 45 MHz. The *A. niger* histone 3-encoding gene, *hhtA* (ASPNIDRAFT_52637) and gamma-actin-encoding gene *actA* (ASPNIDRAFT_200483) were used as internal standards for normalization of expression levels. All primers used for quantitative RT-PCR are shown in Table S6 (primers 90–121). The relative expression levels were approximated based on $2^{\Delta\Delta c(t)}$, with $\Delta\Delta c(t) = \Delta c(t)_{(normalized)} - \Delta c(t)_{(calibrator)}$, where $\Delta c(t)_{(normalized)} = \Delta c(t)_{(target\ gene)} - \Delta c(t)_{(actA\ or\ hhtA)}$.The calibrator c(t) values are those from the *A. niger* reference strain KB1001. Statistical analysis of RT-qPCR results was performed as a Student's t test, and the error bars indicate the SD.

Chemical Analysis of Strains

Unless otherwise stated, strains were cultivated on solid MM media and incubated at 37°C for 5 days. Extraction of metabolites was performed as described by Smedsgaard (1997). 6-MSA was purchased from (Apin Chemicals). Analysis was performed using UPHLC-DAD-TOFMS on a maXis 3G orthogonal acceleration quadrupole time-of-flight mass spectrometer (Bruker Daltonics) equipped with an electrospray ionization (ESI) source and connected to an Ultimate 3000 UHPLC system (Dionex). The column used was a reverse-phase Kinetex 2.6 μm C_{18}, 100 mm × 2.1 mm (Phenomenex), and the column temperature was maintained at 40°C. A linear water-acetonitrile (liquid chromatography-mass spectrometry grade) gradient was used (both solvents were buffered with 20 mM formic acid) starting from 10% (v/v) acetonitrile and increased to 100% in 10 min, maintaining this rate for 3 min before returning to the starting conditions in 0.1 min and staying there for 2.4 min before the following run. A flow rate of 0.4 $ml \cdot min^{-1}$ was used. TOFMS was performed in ESI+ with a data acquisition range of 10 scans per second at *m/z* 100–1,000. The TOFMS was calibrated using Bruker Daltonics high precision calibration algorithm by means of the use of the internal standard sodium formate, which was automatically infused before each run. This provided a mass accuracy of better than 1.5 ppm in MS mode. UV-visible spectra were collected at wavelengths from 200 to 700 nm. Data processing was performed using DataAnalysis 4.0 and Target Analysis 1.2 software (Bruker Daltonics) (Klitgaard et al., 2014). Tandem MS was performed with fragmentation energies from 18 to 55 eV.

Preparative Isolation of Selected Metabolites

The fungal strains were cultivated on 10-200 YES plates at 30°C for 5 days. For details about each extraction, see Table S3. Extracts were filtered and concentrated in vacuo. The combined extract was dissolved in 9:1 methanol (MeOH):H_2O, and 1:1 heptane was added, resulting in two phases. To the MeOH/H_2O phase H_2O was added to a ratio of 1:1, and metabolites were extracted with dichlormethane (DCM). The phases were concentrated separately in vacuo. The DCM phase was adsorbed onto diol column material and dried before packing into a SNAP column (Biotage) with diol material. The extract was fractionated on an Isolera flash purification system (Biotage) using seven steps of heptane-DCM-EtOAc-MeOH. Solvents were of HPLC grade, and H_2O was purified and deionized by a Millipore system through a 0.22 μm membrane filter.

The Isolera fractions were subjected to further purification on a semipreparative high-performance liquid chromatography (HPLC), which was either a Waters 600 controller with a 996 photodiode array detector (Waters) or a Gilson 322 controller connected to a 215 Liquid Handler, 819 Injection Module, and a 172 DAD (Gilson). This was achieved using a Luna II C_{18} column (250 × 10 mm, 5 μm; Phenomenex) or a Gemini C6-Phenyl 110A column (250 × 10.00 mm, 5 μm; Phenomenex). 50 ppm TFA was added to acetonitrile of HPLC grade and Milli-Q water. For choice of system, flow rate, column, gradients, and yields, see Table S3.

NMR and Structural Elucidation

The 1D and 2D spectra were recorded on a Unity Inova-500 MHz spectrometer (Varian). Spectra were acquired using standard pulse sequences, and ^1H, double quantum filtered-correlated spectroscopy, nuclear Overhauser effect spectroscopy, heteronuclear single quantum coherence, and heteronuclear multiple bond correlation spectra were acquired. The deuterated solvent was acetonitrile-d_3, and signals were referenced by solvent signals for acetonitrile-d_3 at δ_H = 1.94 ppm and δ_C = 1.32/118.26 ppm. The NMR data was processed in Bruker Topspin 3.1 or ACD NMR Workbook. Chemical shifts are reported in ppm (δ) and scalar couplings in hertz. The sizes of the *J* coupling constants reported in the tables are experimentally measured values from the spectra. There are minor variations in the measurements that may be explained by the uncertainty of *J*. Descriptions of the structural elucidations are shown in Table S4.

CD spectra were obtained from a J-710 spectropolarimeter (Jasco). The methanol dissolved samples (1 mg/3 ml) were analyzed in 0.2 cm optical path length cells at 20°C, and the spectra were recorded from 200 to 500 nm. Optical rotation was measured on a PerkinElmer 321 Polarimeter.

Production and Purification of Fully Labeled ^{13}C-6-MSA

Because fully labeled $^{13}C_8$-6-MSA was not commercially available, it was produced in-house from the 6-MSA-producing strain by batch cultivation. Spores propagated on CYA media plates for 7 days at 30°C were harvested with 10 ml of 0.9% NaCl through Mira cloth. The spores were washed twice with 0.9% NaCl. The batch fermentation was initiated by inoculation of $2 \cdot 10^9$ spores/l. A Sartorious 1 l bioreactor (Satorious) with a working volume of 0.8 l equipped with two Rushton six-blade disc turbines was used. The pH electrode (Mettler) was calibrated according to manufacturer standard procedures. The bioreactor was sparged with sterile atmospheric air, and off-gas concentrations of oxygen and carbon dioxide were measured with a Prima Pro Process Mass Spectrometer (Thermo-Fischer Scientific). Temperature was maintained at 30°C, and pH was controlled by addition of 2 M NaOH and H_2SO_4. Start conditions were pH: 3.0, stir rate: 100 rpm, and air flow: 0.1 volume of air per volume of liquid per minute (vvm). These conditions were changed linearly in 720 min to pH: 5.0, stir rate: 800 rpm, and air flow: 1 vvm. The strain was cultivated until glucose was depleted, as measured by glucose test strips (Machery-Nagel), and the culture had entered stationary phase as monitored by off-gas CO_2 concentration.

The entire volume of the reactor was harvested, and the biomass was removed by filtration through a Whatman 1 qualitative paper filter followed by centrifugation at 8,000 × *g* for 20 min to remove fine sediments. The 6-MSA was then recovered from the supernatant by liquid-liquid extraction using ethyl acetate with 0.5% formic acid.

The organic extract then dried in vacuo to give a crude extract that was redissolved in 20 ml of ethyl acetate and dry loaded onto 3 g of Sepra ZT C18 (Phenomenex) resin prior to packing into a 25 g SNAP column (Biotage) with 22 g of pure resin in the base. The crude extract was fractionated on an Isolera flask purification system (Biotage) using an water-acetonitrile gradient starting at 15:85 going to 100% acetonitrile in 23 min at a flow rate of 25 ml min^{-1} and kept at that level for 4 min. Fractions were collected using UV detection at 210 and 254 nm, resulting in a total of 15 fractions, of which 3 were pooled and analyzed. 6-MSA concentration was assessed using a Dionex Ultimate 3000 UHPLC coupled with a ultimate 3000 RS diode array detector (Dionex) equipped with a Poroshell 120 phenyl hexyl 2.1 × 100 mm, 2.7 μm (Agilent) column. Finally, purity (98.7%) was analyzed by UHPLC-TOFMS (Figure S3A).

Feeding Experiments

Solid YES plates were prepared using a 6 mm plug drill to make a well in the middle of the agar. 25-100 μl of spore suspension was added to the well, and plates were incubated at 30°C for 5 days. 100 μg of ^{13}C-6-MSA, *m*-cresol, and toluquinol (ortoluquinol) was added to the plates after 24, 48, and 72 hr, respectively. Agar plugs were taken both as reported previously (Smedsgaard, 1997) and also separately from the center, the middle, and the rim of the colony, respectively, to verify diffusion and absorption of the 6-MSA and the location of yanuthone production. Samples were analyzed as described in "Chemical Analysis of Strains."

Antifungal Susceptibility Testing

All compounds were screened for antifungal activity toward *C. albicans* in accordance with the CLSI standards using RPMI 1640 medium adjusted to pH 7 with 0.165 M MOPS buffer (CLSI, 2012). Inoculum was prepared to a final concentration of approximately 2.5×10^3 cells per ml. Inoculated media were seeded into 96-well microtiter plates in aliquots of 200 μl using a Hamilton STAR liquid handling workstation with an integrated Thermo Cytomat shaking incubator and Biotek Synergy Mx microplate reader. The pure compounds were dissolved in Me_2SO and applied at 100 to 5 μM (1% Me_2SO per well). The plates were incubated at 35°C at 1,200 rpm shaking with a 2 mm amplitude. Optical density was recorded every hour for 20 hr. Endpoint optical densities from compound screens were normalized to the negative controls, and susceptibility was evaluated as the percentage of reduction in optical density. All bioactive compounds were tested in duplicate in three independent trials to ensure reproducibility and to evaluate potency of the compound toward the target organism. The half-maximal inhibitory concentration (IC_{50}) was extrapolated from compound specific dilution sequences and annotated as the average concentration for which 50% inhibition plus minus the SD was observed.

SUPPLEMENTAL INFORMATION

Supplemental information includes six figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.chembiol.2014.01.013.

ACKNOWLEDGMENTS

We kindly acknowledge Kenneth S. Bruno (Pacific Northwest National Laboratory, Richland, WA) for providing the KB1001 strain and Solveig Kallesøe (University of Copenhagen) for obtaining CD data. We also acknowledge Diana C. Anyaogu for assisting with the RT-qPCR analysis. The study was supported by grant 09-064967 from the Danish Council for Independent Research, Technology, and Production Sciences.

Received: October 18, 2013
Revised: January 14, 2014
Accepted: January 29, 2014
Published: March 27, 2014

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Andersen, M.R., Nielsen, J.B., Klitgaard, A., Petersen, L.M., Zachariasen, M., Hansen, T.J., Blicher, L.H., Gotfredsen, C.H., Larsen, T.O., Nielsen, K.F., and Mortensen, U.H. (2013). Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proc. Natl. Acad. Sci. USA *110*, E99–E107.

Artigot, M.P., Loiseau, N., Laffitte, J., Mas-Reguieg, L., Tadrist, S., Oswald, I.P., and Puel, O. (2009). Molecular cloning and functional characterization of two CYP619 cytochrome P450s involved in biosynthesis of patulin in Aspergillus clavatus. Microbiology *155*, 1738–1747.

Baker, S.E. (2006). Aspergillus niger genomics: past, present and into the future. Med. Mycol. *44* (*Suppl 1*), S17–S21.

Beck, J., Ripka, S., Siegner, A., Schiltz, E., and Schweizer, E. (1990). The multifunctional 6-methylsalicylic acid synthase gene of *Penicillium patulum*. Its gene structure relative to that of other polyketide synthases. Eur. J. Biochem. *192*, 487–498.

Bentley, R. (2000). Mycophenolic acid: a one hundred year odyssey from antibiotic to immunosuppressant. Chem. Rev. *100*, 3801–3826.

Bernard, P., and Couturier, M. (1992). Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. J. Mol. Biol. *226*, 735–745.

Bugni, T.S., Abbanat, D., Bernan, V.S., Maiese, W.M., Greenstein, M., Van Wagoner, R.M., and Ireland, C.M. (2000). Yanuthones: novel metabolites from a marine isolate of Aspergillus niger. J. Org. Chem. *65*, 7195–7200.

Chiang, Y.-M., Meyer, K.M., Praseuth, M., Baker, S.E., Bruno, K.S., and Wang, C.C.C. (2011). Characterization of a polyketide synthase in Aspergillus niger whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-γ-pyrone. Fungal Genet. Biol. *48*, 430–437.

CLSI (Clinical and Laboratory Standards Institute) (2012). Reference methods for broth dilution antifungal susceptibility testing of yeasts. CLSI document M27-S4, Fourth International Supplement. (Wayne, PA: Clinical and Laboratory Standards Institute).

Cove, D.J. (1966). The induction and repression of nitrate reductase in the fungus Aspergillus nidulans. Biochim. Biophys. Acta *113*, 51–56.

Cox, R.J. (2007). Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets. Org. Biomol. Chem. *5*, 2010–2026.

Endo, A., Kuroda, M., and Tsujita, Y. (1976). ML-236A, ML-236B, and ML-236C, new inhibitors of cholesterogenesis produced by *Penicillium citrinium*. J. Antibiot. (Tokyo) *29*, 1346–1348.

Fisch, K.M., Gillaspy, A.F., Gipson, M., Henrikson, J.C., Hoover, A.R., Jackson, L., Najar, F.Z., Wägele, H., and Cichewicz, R.H. (2009). Chemical induction of silent biosynthetic pathway transcription in Aspergillus niger. J. Ind. Microbiol. Biotechnol. *36*, 1199–1213.

Frisvad, J.C., Smedsgaard, J., Samson, R.A., Larsen, T.O., and Thrane, U. (2007). Fumonisin B2 production by Aspergillus niger. J. Agric. Food Chem. *55*, 9727–9732.

Frisvad, J.C., Rank, C., Nielsen, K.F., and Larsen, T.O. (2009). Metabolomics of Aspergillus fumigatus. Med. Mycol. *47* (*Suppl 1*), S53–S71.

Geu-Flores, F., Nour-Eldin, H.H., Nielsen, M.T., and Halkier, B.A. (2007). USER fusion: a rapid and efficient method for simultaneous fusion and cloning of multiple PCR products. Nucleic Acids Res. *35*, e55.

González, J.B. (2012). Solid-state fermentation: physiology of solid medium, its molecular basis and applications. Process Biochem. *47*, 175–185.

Gross, H. (2007). Strategies to unravel the function of orphan biosynthesis pathways: recent examples and future prospects. Appl. Microbiol. Biotechnol. *75*, 267–277.

Hansen, B.G., Salomonsen, B., Nielsen, M.T., Nielsen, J.B., Hansen, N.B., Nielsen, K.F., Regueira, T.B., Nielsen, J., Patil, K.R., and Mortensen, U.H. (2011). Versatile enzyme expression and characterization system for Aspergillus nidulans, with the Penicillium brevicompactum polyketide synthase gene from the mycophenolic acid gene cluster as a test case. Appl. Environ. Microbiol. *77*, 3044–3051.

Hertweck, C. (2009). The biosynthetic logic of polyketide diversity. Angew. Chem. Int. Ed. Engl. *48*, 4688–4716.

Johnstone, I.L., Hughes, S.G., and Clutterbuck, A.J. (1985). Cloning an Aspergillus nidulans developmental gene by transformation. EMBO J. *4*, 1307–1311.

Klitgaard, A., Iversen, A., Andersen, M.R., Larsen, T.O., Frisvad, J.C., and Nielsen, K.F. (2014). Aggressive dereplication using UHPLC-DAD-QTOF: screening extracts for up to 3000 fungal secondary metabolites. Anal. Bioanal. Chem. Published online January 18, 2014. http://dx.doi.org/10.1007/s00216-013-7582-x.

Li, X., Choi, H.D., Kang, J.S., Lee, C.O., and Son, B.W. (2003). New polyoxygenated farnesylcyclohexenones, deacetoxyyanuthone A and its hydro derivative from the marine-derived fungus *Penicillium sp*. J. Nat. Prod. *66*, 1499–1500.

Light, R.J. (1969). 6-methylsalicylic acid decarboxylase from *Penicillium patulum*. Biochim. Biophys. Acta *191*, 430–438.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. *39* (Database issue), D225–D229.

McCluskey, K., Wiest, A., and Plamann, M. (2010). The Fungal Genetics Stock Center: a repository for 50 years of fungal genetics research. J. Biosci. *35*, 119–126.

Nielsen, M.L., Albertsen, L., Lettier, G., Nielsen, J.B., and Mortensen, U.H. (2006). Efficient PCR-based gene targeting with a recyclable marker for Aspergillus nidulans. Fungal Genet. Biol. *43*, 54–64.

Nielsen, J.B., Nielsen, M.L., and Mortensen, U.H. (2008). Transient disruption of non-homologous end-joining facilitates targeted genome manipulations in the filamentous fungus Aspergillus nidulans. Fungal Genet. Biol. *45*, 165–170.

Nielsen, K.F., Mogensen, J.M., Johansen, M., Larsen, T.O., and Frisvad, J.C. (2009). Review of secondary metabolites and mycotoxins from the Aspergillus niger group. Anal. Bioanal. Chem. *395*, 1225–1242.

Nielsen, M.L., Nielsen, J.B., Rank, C., Klejnstrup, M.L., Holm, D.K., Brogaard, K.H., Hansen, B.G., Frisvad, J.C., Larsen, T.O., and Mortensen, U.H. (2011). A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in Aspergillus nidulans. FEMS Microbiol. Lett. *321*, 157–166.

Nørholm, M.H.H. (2010). A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. BMC Biotechnol. *10*, 21.

Nour-Eldin, H.H., Hansen, B.G., Nørholm, M.H.H., Jensen, J.K., and Halkier, B.A. (2006). Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. Nucleic Acids Res. *34*, e122.

Olsen, J.H., Dragsted, L., and Autrup, H. (1988). Cancer risk and occupational exposure to aflatoxins in Denmark. Br. J. Cancer *58*, 392–396.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., Schaap, P.J., Turner, G., de Vries, R.P., Albang, R., Albermann, K., et al. (2007). Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. Nat. Biotechnol. *25*, 221–231.

Puel, O., Galtier, P., and Oswald, I.P. (2010). Biosynthesis and toxicological effects of patulin. Toxins (Basel) *2*, 613–631.

Regueira, T.B., Kildegaard, K.R., Hansen, B.G., Mortensen, U.H., Hertweck, C., and Nielsen, J. (2011). Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum. Appl. Environ. Microbiol. *77*, 3035–3043.

Rho, M.C., Toyoshima, M., Hayashi, M., Uchida, R., Shiomi, K., Komiyama, K., and Omura, S. (1998). Enhancement of drug accumulation by andrastin A produced by Penicillium sp. FO-3929 in vincristine-resistant KB cells. J. Antibiot. (Tokyo) *51*, 68–72.

Samson, R.A., Houbraken, J., Thrane, U., Frisvad, J.C., and Andersen, B. (2010). Food and Indoor Fungi. CBS Laboratory Manual Series 2. (Utrecht, The Netherlands: CBS KNAW Fungal Biodiversity Centre).

Schachtschabel, D., Arentshorst, M., Nitsche, B.M., Morris, S., Nielsen, K.F., van den Hondel, C.A.M., Klis, F.M., and Ram, A.F.J. (2013). The transcriptional repressor TupA in Aspergillus niger is involved in controlling gene expression related to cell wall biosynthesis, development, and nitrogen source availability. PLoS One *8*, e78102.

Smedsgaard, J. (1997). Micro-scale extraction procedure for standardized screening of fungal metabolite production in cultures. J. Chromatogr. A *760*, 264–270.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33* (Web Server issue), W465–W467.

Wattanachaisaereekul, S., Lantz, A.E., Nielsen, M.L., and Nielsen, J. (2008). Production of the polyketide 6-MSA in yeast engineered for increased malonyl-CoA supply. Metab. Eng. *10*, 246–254.

Williams, D.H., Stone, M.J., Hauck, P.R., and Rahman, S.K. (1989). Why are secondary metabolites (natural products) biosynthesized? J. Nat. Prod. *52*, 1189–1208.

Zabala, A.O., Xu, W., Chooi, Y.-H., and Tang, Y. (2012). Characterization of a silent azaphilone gene cluster from Aspergillus niger ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. Chem. Biol. *19*, 1049–1059.

SUPPORTING INFORMATION

Figures:

Tables:

Figure S1, related to Figure 2. BPC of *yanA*Δ strain relative to the reference KB1001, cultivated on MM (**A**), YES (**B**), CYA (**C**), and MEA (**D**) for 5 days at 30°C.

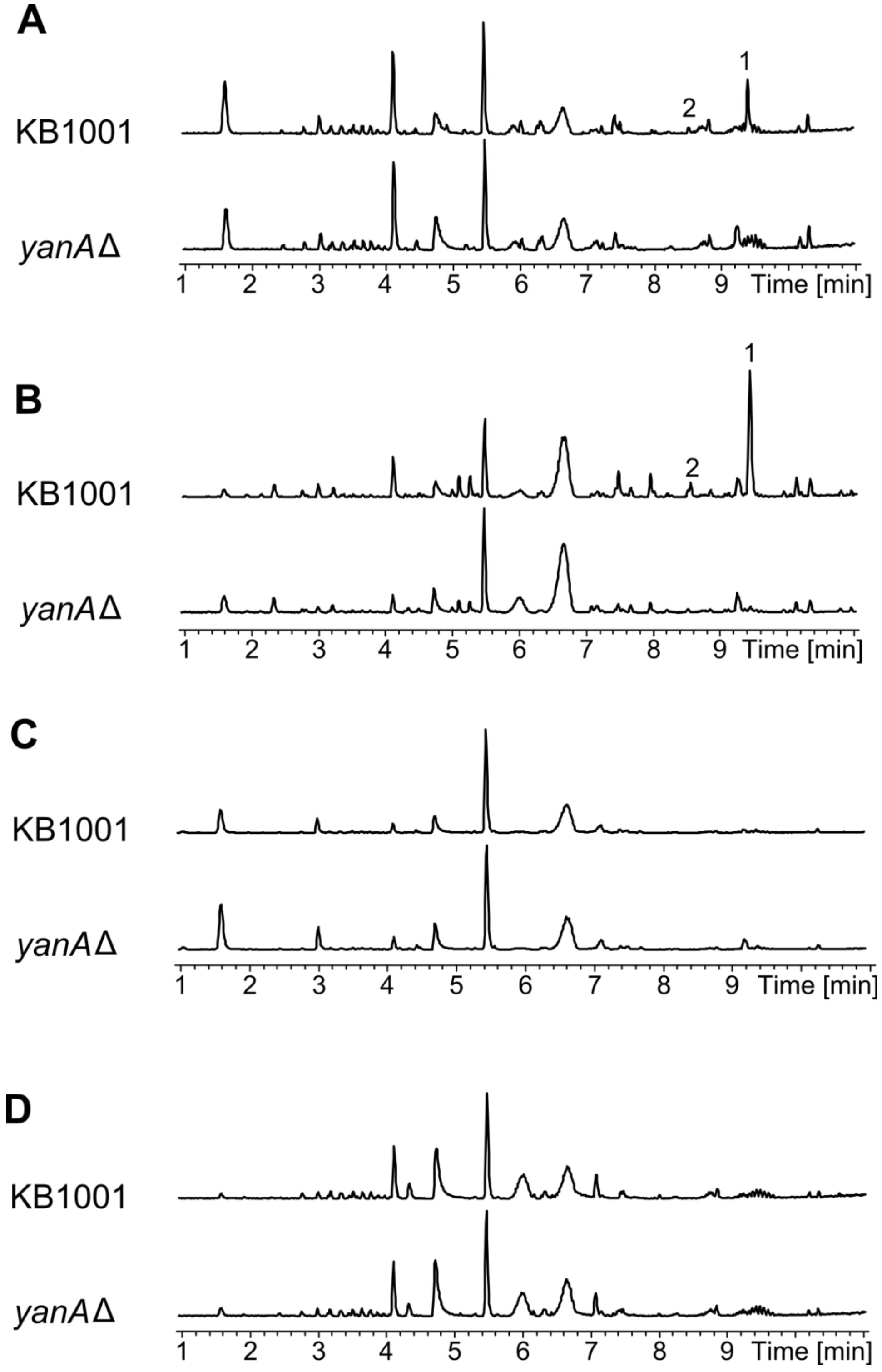


Figure S2, related to Figure 2. The morphology of *A. niger* KB1001 is identical with and without addition of $^{13}C_8$-6-MSA. The figure shows the top and bottom of KB1001 cultivated on YES medium for 5 days at 30 °C.

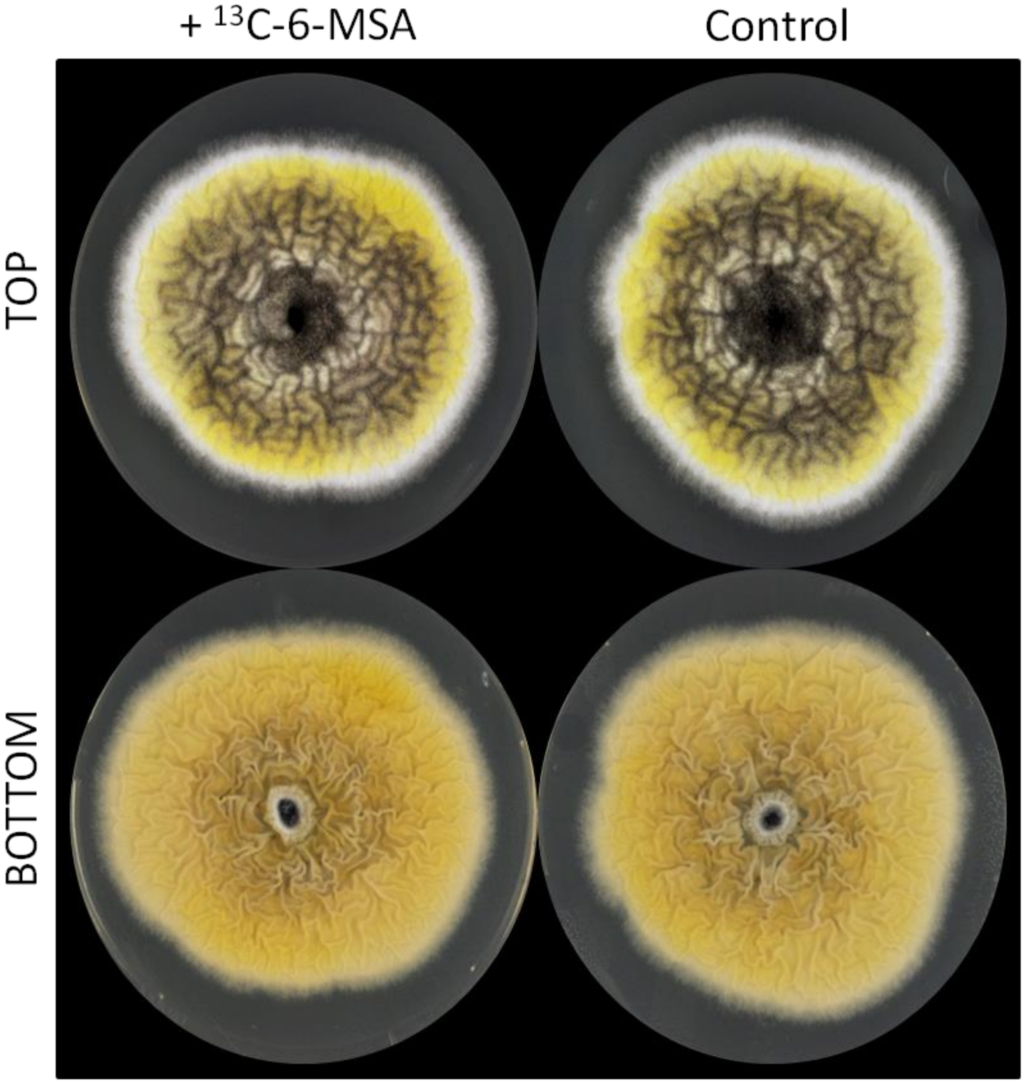


Figure S3, related to Figure 2. Positive electrospray (ESI+) mass spectrum of: **(A)** uniformly labeled $^{13}C_8$-6-methylsalicylic acid, **(B)** unlabeled and labeled yanuthone D**, (C)** unlabeled and labeled yanuthone E, and **(D)** unlabelled yanuthone X_1. The calculated shift from $^{12}C_7$ to $^{13}C_7$ is 7.0234 Da.

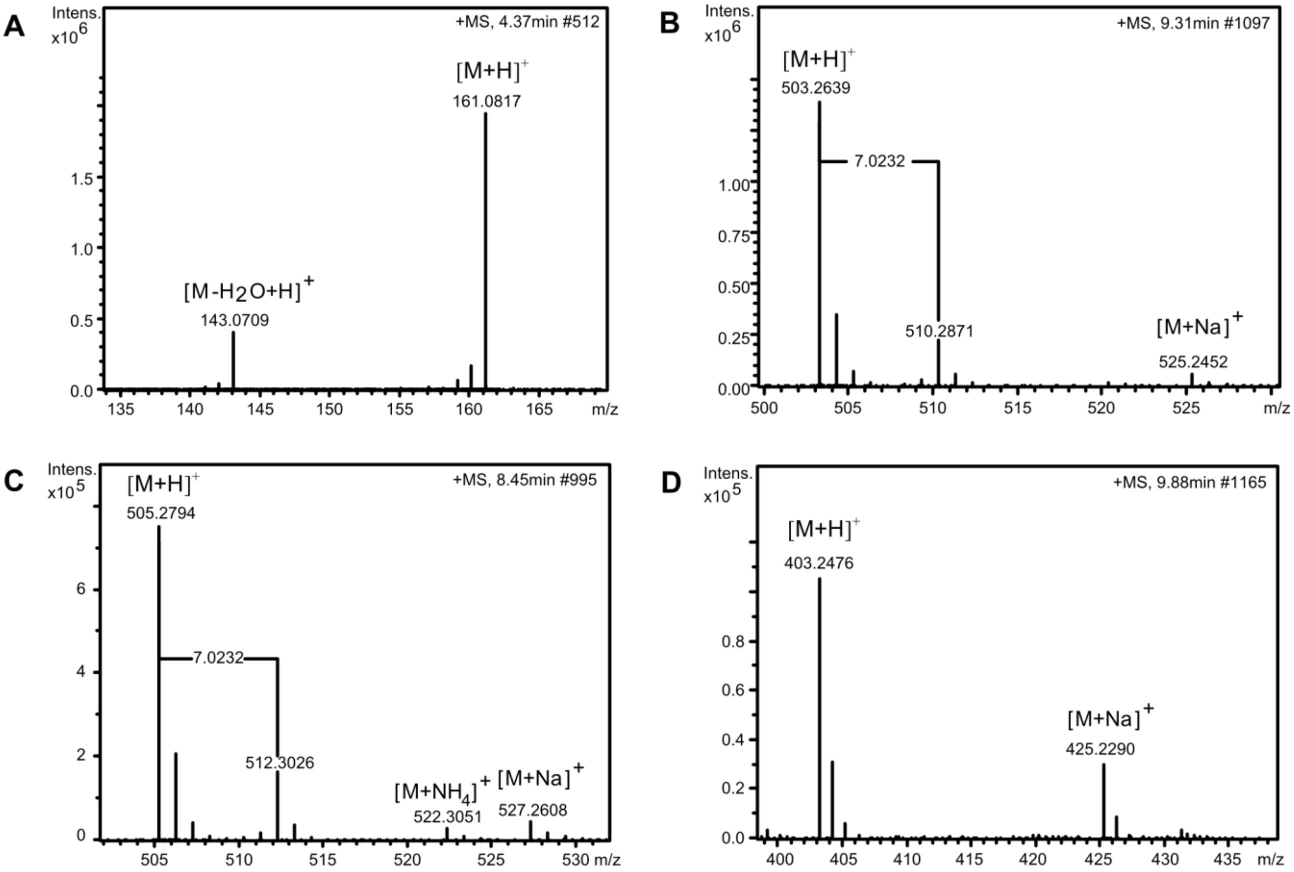


Figure S4, related to Figure 3. RT-qPCR expression analysis of *yanA* and 13 flanking genes in a *yanR*Δ (44961Δ) strain and a reference strain, KB1001. **A**. Absolute expression levels ($(\Delta c(t))^{-1}$ values) in KB1001 (grey bar) compared to the TFΔ strain (blue/red bar). Red bars indicate significant down-regulation (p-value <0.05), blue bars indicate non-significant changes (p-value >0.05). Error bars indicate the standard deviation (SD). Values are normalized to expression of *hhtA*. **B**. Relative expression levels (fold change, $2^{\Delta\Delta c(t)}$ values) of *yan* cluster genes (except *yanR*, which is deleted) in the TFΔ strain relative to KB1001. Error bars indicate the standard deviation (SD). Expression levels are individually normalized to expression of *actA* and *hhtA*, as indicated.

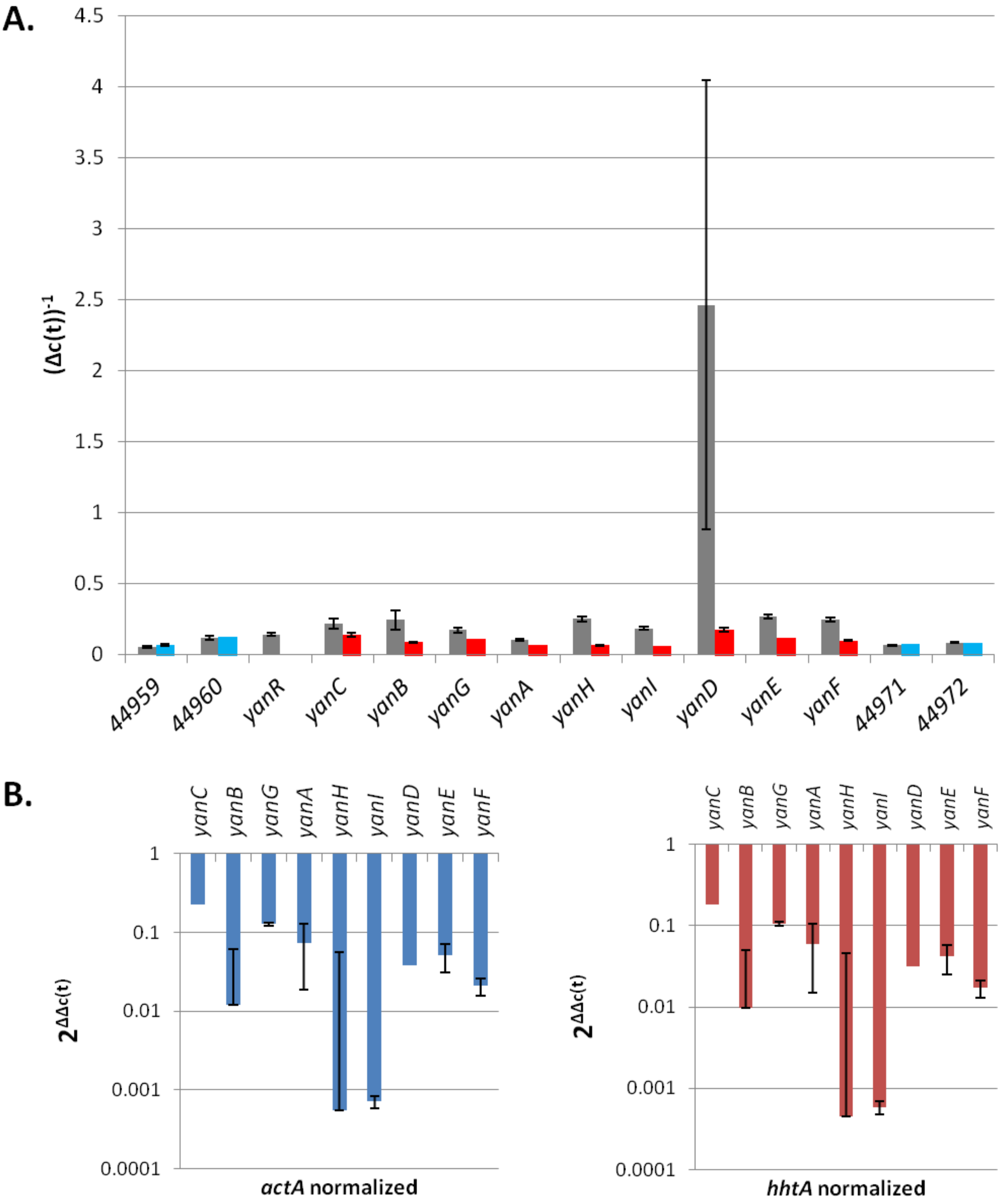

A.
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
$(\Delta c(t))^{-1}$
44959
44960
yanR
yanC
yanB
yanG
yanA
yanH
yanI
yanD
yanE
yanF
44971
44972
B.
yanC
yanB
yanG
yanA
yanH
yanI
yanD
yanE
yanF
1
0.1
0.01
0.001
0.0001
$2^{\Delta\Delta c(t)}$
actA normalized
yanC
yanB
yanG
yanA
yanH
yanI
yanD
yanE
yanF
1
0.1
0.01
0.001
0.0001
$2^{\Delta\Delta c(t)}$
hhtA normalized

Figure S5, related to Figure 2. (**A)** Gene deletion in *A. niger*. The ORF is replaced by the selectable hygromycin B phosphatase (*hph*) gene. Not to scale. (**B)** pDHX2 vector used for expression of all cluster genes in the *A. nidulans* IS*1-yanA* strain. The vector contains the AMA1 sequence for autonomous replication in *Aspergillus* and the *pyrG* gene for selection.

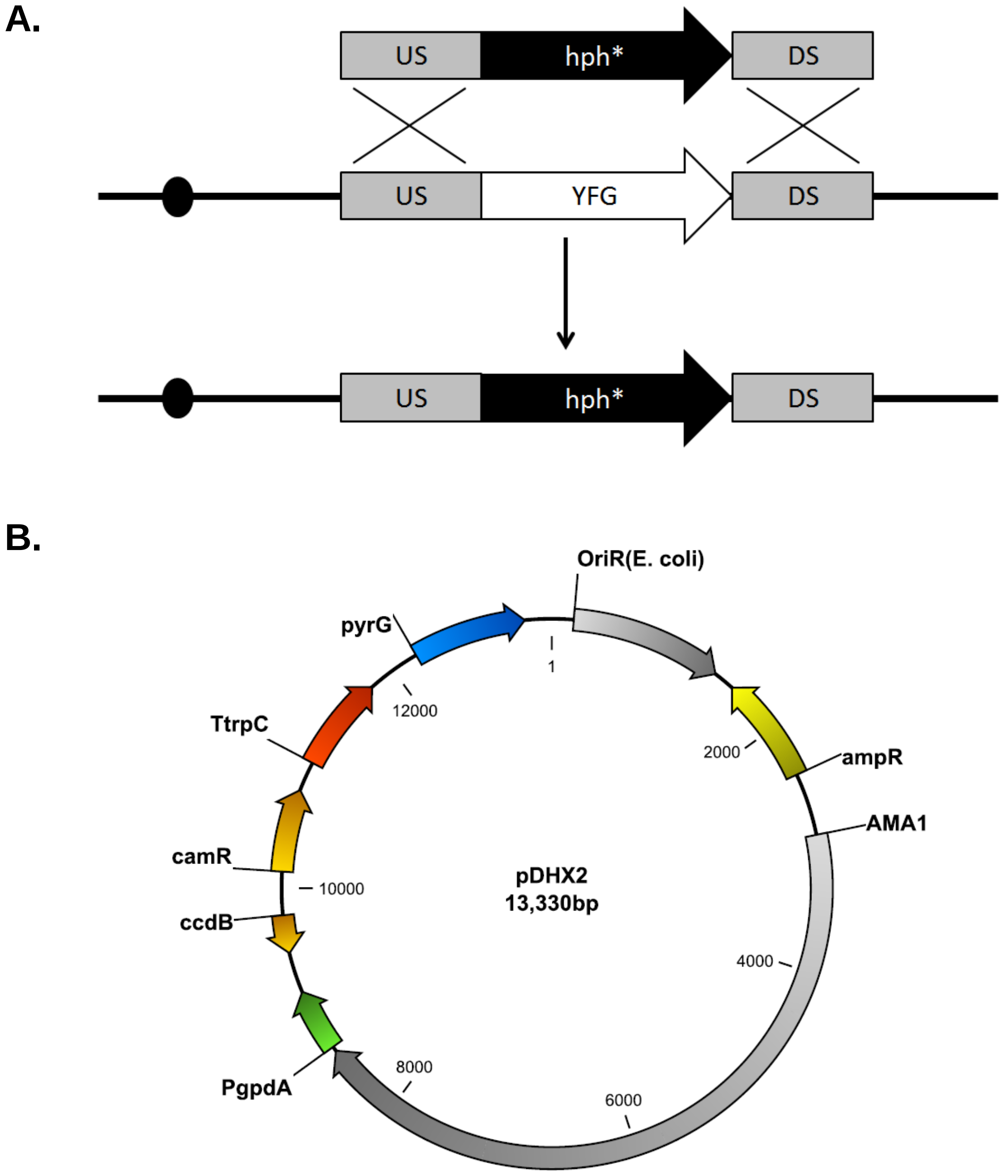


Figure S6, related to Figure 5. Base peak chromatogram (ESI+) of the five strains that express putative cluster genes *yanB*, *yanC*, *yanD*, *yanE*, and *yanG* in the *A. nidulans* IS*1*-*yanA* strain (6-MSA producing reference strain). All strains, except Oex-*yanB* still produce 6-MSA. IS = internal standard: chloroamphenicol, A = austinol, DHA = dehydroaustinol. Chromatograms are to scale.

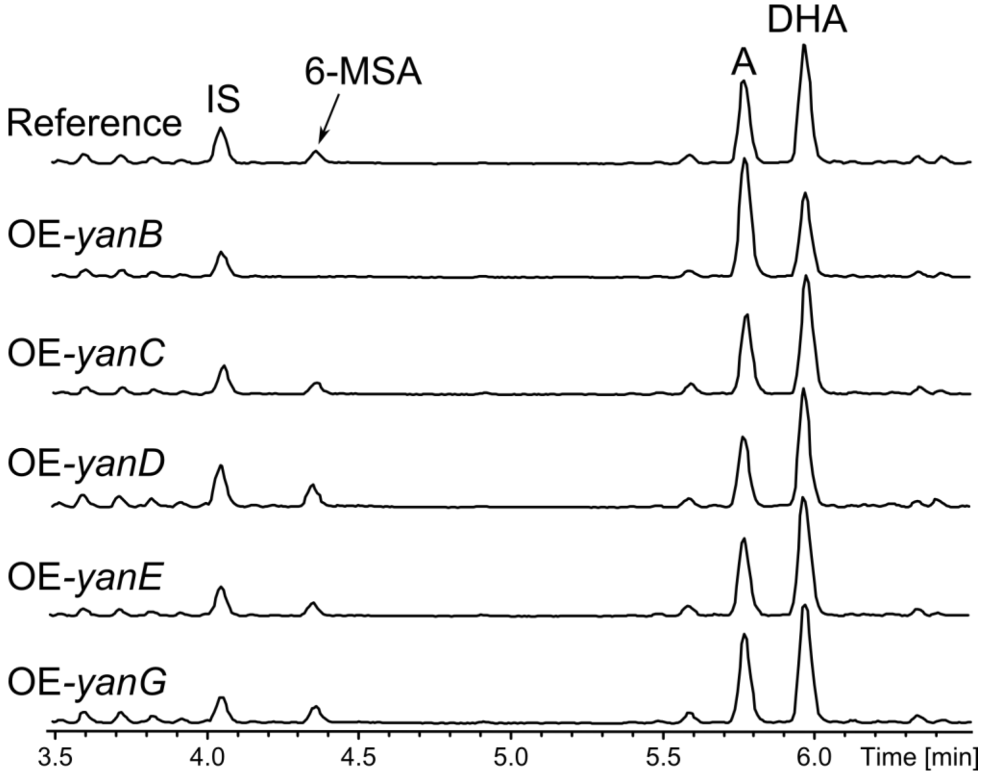


Table S1, related to Figures 2 and 4. Detection of metabolites in the deletion and overexpression strains using extracted ion chromatograms of $[M+H]^+$ ± 0.005.

	6-MSA	Yanuthone D (**1**)	Yanuthone E (**2**)	7-deacetoxyyanuthone A (**3**)	Yanuthone F (**4**)	Yanuthone G (**5**)	22-deacetylyanutone A (**6**)	Yanuthone H (**7**)	Yanuthone I (**8**)	Yanuthone J (**9**)	Yanuthone X_1 (**12**)
KB1001	-	+	+	-	-	+	-	-	+	-	+
OE-*yanA*	+	-	-	-	-	-	-	-	-	-	-
*yanA*Δ	-	-	-	-	-	-	-	-	-	-	+
*yanB*Δ	-	-	-	-	-	+	-	-	+	-	+
*yanC*Δ	-	-	-	-	-	-	-	-	-	-	-
*yanD*Δ	-	-	-	-	-	-	-	-	-	-	-
*yanE*Δ	-	-	-	-	-	-	-	-	-	-	-
*yanF*Δ	-	-	+	+	-	-	-	-	+	+	+
*yanG*Δ	-	-	-	-	-	-	-	-	-	-	-
*yanH*Δ	-	-	-	+	+	+	-	-	-	-	+
*yanI*Δ	-	-	-	+	-	-	+	+	+	-	+
*yanR*Δ	-	-	-	+	-	+	-	-	-	-	+

Table S2, related to Figure 3. Overview of genes and proposed activities of the *yan* cluster.

Locus (ASPNIDRAFT_)	**Gene name**	**Predicted functional domains (CDD)**	**Proposed activity**
44959	-	No conservation	-
44960	-	Glyoxalase	-
44961	*yanR*	Fungal transcription factor	Transcription factor
54844	*yanC*	CYP450	*m*-Cresol hydroxylase
44963	*yanB*	Amidohydrolase, decarboxylase	6-MSA decarboxylase
44964	*yanG*	UbiA-like prenyltransferase	Prenyltransferase
44965	*yanA*	Polyketide synthase	6-MSA synthase
193092	*yanH*	CYP450	Cytochrome P450
44967	*yanI*	Membrane bound O-acyl transferase	O-Mevalon transferase
127904	*yanD*	Short-chain dehydrogenase	Dehydrogenase
192604	*yanE*	No conservation	-
44970	*yanF*	FAD/FMN-containing dehydrogenases	Oxidase
44971	-	No conservation	-
44972	-	No conservation	-

Table S3, related to Figure 4. Purification of metabolites

Compound name	**Strain**	**Number of plates**	**Extraction**	**Isolera fractionation**	**Purification**	**Yield**
Yanuthone D (**1**)	*A. niger* KB1001	200	EtOAc + 1 % FA	50g diol, 40 mL·min^{-1}, CV = 66 mL, auto fractionation, 15 fractions	Waters, Luna II C_{18}, 4 mL·min^{-1}, 40-100% ACN over 20 min.	4.5 mg
Yanuthone E (**2**)	*A. niger* KB1001	200	EtOAc + 1 % FA	50g diol, 40 mL·min^{-1}, CV = 66 mL, auto fractionation, 15 fractions	Gilson, Luna II C_{18}, 5 mL·min^{-1}, 40-100% ACN over 20 min.	2.9 mg
7-deacetoxy Yanuthone A (**3**)	*A. niger* KB1001	200	EtOAc + 1 % FA	50g diol, 40 mL·min^{-1}, CV = 66 mL, auto fractionation, 15 fractions	Waters, Luna II C_{18}, 4 mL·min^{-1}, 40-100% ACN over 20 min.	9.3 mg
Yanuthone F (**4**)	*yanH*Δ	100	EtOAc	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 10 fractions	Gilson, Luna II C_{18}, 5 mL·min^{-1}, 30-80% ACN over 18 min. Waters, gemini, 5 mL·min^{-1}, 40-65% ACN over 20 min	1.8 mg
Yanuthone G (**5**)	*yanH*Δ	100	EtOAc	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 10 fractions	Gilson, Luna II C_{18}, 5 mL·min^{-1}, 30-80% ACN over 18 min. Waters, gemini, 5 mL·min^{-1}, 30-40% ACN over 20 min, to 45% for 2 min.	4.0 mg
22-deacetylyanuthone A (**6**)	*yanI*Δ	100	EtOAc	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 12 fractions	Gilson, Luna II C_{18}, 5 mL·min^{-1}, 20-90% ACN over 17 min. Waters, Luna II C_{18}, 5 mL·min^{-1}, 30-100% ACN over 20 min.	7.3 mg
Yanuthone H (**7**)	*yanI*Δ	100	EtOAc	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 12 fractions	Gilson, Luna II C_{18}, 5 mL·min^{-1}, 30-60% ACN over 16 min.	9.8 mg
Yanuthone I **(8)**	*yanI*Δ	100	EtOAc	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 12 fractions	Waters, Luna II C_{18}, 5 mL·min^{-1}, 30-60% ACN over 16 min.	4.7 mg
Yanuthone J **(9)**	*yanF*Δ	150	EtOAc + 1 % FA	25g diol, 25 mL·min^{-1}, CV = 33 mL, auto fractionation, 15 fractions	Waters, Luna II C_{18}, 4 mL·min^{-1}, 20-100% ACN over 20 min.	1.5 mg
Yanuthone X_1 (**12**)	*A. niger* KB1001	200	EtOAc + 1 % FA	50g diol, 40 mL·min^{-1}, CV = 66 mL, auto fractionation, 15 fractions	Waters, gemini, 4 mL·min^{-1}, 90 % ACN isocratic for 15 min, then to 100 % ACN for 5 min.	1.5 mg

Table S4, related to Figure 6. Spectroscopic data

The structural elucidation of the compounds showed several similar features in the ^{1}H as well as 2D spectra comparable to those reported for the known yanuthones(Bugni *et al.*, 2000; Li *et al.*, 2003). All compounds except yanuthone I displayed 8H overlapping resonances at δ_H 1.93-2.11 ppm in the ^{1}H spectrum corresponding to the four methylene groups H4, H5, H8 and H9 in the sesquiterpene moiety. Other common resonances were from the diastereotopic pair H-12/H-12' and 3 methyl groups (H-19, H-20 and H-21) around δ_H 1.60 ppm, whereof H-20 and H-21 were overlapping. In the HMBC spectrum a correlation to the quaternary C-18 around δ_C 194 ppm was seen and all compounds also had two carbons around 60 ppm (one quaternary, one methine) being the carbons in the epoxide ring.

The compounds however differed greatly in the moiety attached to C-16. Yanuthone D, yanuthone E and yanuthone J all displayed two methylene groups around δ_C 45 ppm, a methyl group around δ_C 28 ppm, two carbonyls around δ_C 171 ppm and another quaternary carbon around δ_C 70 ppm for the mevalonic acid part. 7-deacetoxyanuthone A, yanuthone F and yanuthone G all had a methyl group attached at C-16 while yanuthone H, yanuthone I and 22-deacetylyanuthone A had a further hydroxy group at C-22. The hydroxylation in this position was indicated by a significant shift downfield of H-22/C-22.

Some structures had a further modification being a hydroxy group at either C-1 or C-2. The compounds yanuthone F, G and H all had a hydroxy group at C-1, shifting C-1 and H-1 significantly downfield. Yanuthone J had the hydroxy group attached at C-2 which shifted the resonances for C-2 and H-2 downfield, and due to the lack of the double bond in those structures, the resonance for H-3 was no longer observed in the double bond area but at δ_H 1.35 ppm. Yanuthone I differed in this part of the structure with fewer resonances due to the shorter terpene chain.

The ^{1}H NMR spectrum for yanuthone G stood out from the rest due to several resonances between 3-5 ppm. Elucidation of the structure revealed a sugar moiety attached to the hydroxy group at C-15. The presence of this hexose unit gave rise to the additional resonances observed.

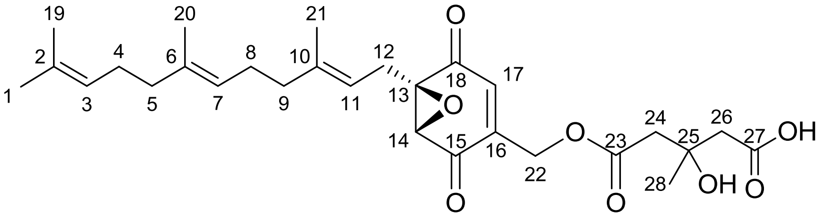
The NMR data for yanuthone X_1 displayed the same resonances for the sesquiterpene part of the molecule, but the methoxy group attached to C-16 is different for all other reported yanuthones, and was obvious from the chemical shift of C-16 which gave rise to a resonance at δ_C 168.3 ppm, which is considerable further downfield than in the other structures. Furthermore C-17 was affected shifting upfield to δ_C 100.3 ppm. NMR data for all compounds can be found in invidual tabs in this file.

The stereochemistry of the compounds was investigated by circular dichroism (CD) and optical rotation. The CD data for yanuthone D, E and 7-deacetoxyyanuthone A showed that the positive and negative cotton effects were identical to those previously reported for these compounds (Bugni et al., 2000).

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone D

HRESIMS: *m/z* = 503.2640 [M + H]$^+$, calculated for [$C_{28}H_{38}O_8$+H]$^+$: 503.2639. $[\alpha]_{587}^{20} = 32.4°$
CD$_{MeOH}$: $[\theta]_{213}$ = 12.4263, $[\theta]_{230}$ = -7.8722, $[\theta]_{240}$ = 1.1769, $[\theta]_{260}$ = -4.0282, $[\theta]_{298}$ = 0.7690



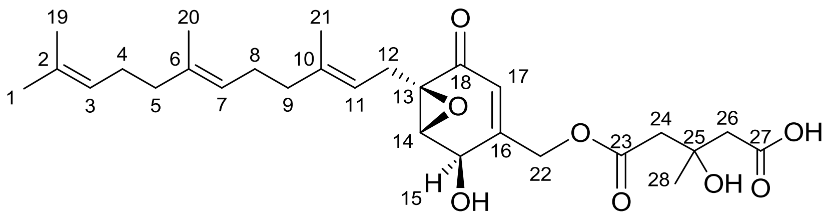
NMR data for yanuthone D

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]	^{13}C-chemical shift [ppm]	HMBC correlations	NOESY connectivities
1	1.59 (3H, s)	17.1	2, 19	-
2	-	132.1	-	-
3	5.10 (1H, m)	125.0	1, 19	19
4	2.06 (2H, m)	27.3	2, 3, 5/9	20
5	1.97 (2H, m)	40.2	3/7, 6, 4/8,20	20
6	-	136.4	-	-
7	5.09 (1H, m)	124.9	5/9, 8, 20	9
8	2.09 (2H, m)	27.0	5/9, 6, 7, 10	20, 21
9	1.99 (2H, m)	40.1	7, 8, 10, 11, 21	7, 11
10	-	141.0	-	-
11	5.03 (1H, t, 1.5)	116.4	9, 12, 21	9, 12, 12’, 14
12	2.66 (1H, m)	25.6	10, 11, 13, 14, 15/18	11, 12’, 14, 21
12’	2.76 (1H, m)	25.6	10, 11, 13, 14, 15/18	11, 12, 14, 21
13	-	63.0	-	-
14	3.68 (1H, s)	58.6	12, 13, 15/18, 16, 22	11, 12, 12’, 21
15	-	193.2	-	-
16	-	143.9	-	-
17	6.58 (1H, t, 1.5)	133.1	13, 15/18, 22	22, 22’, 24, 28
18	-	193.2	-	-
19	1.66 (3H, s)	25.7	1, 2, 3	3, 4
20	1.591 (3H, s)	16.1	5, 6, 7	5, 8
21	1.63 (3H, s)	16.2	9, 10, 11	8, 12, 12’, 14
22	4.84 (1H, dd, 16.1, 1.5)	60.2	15/18, 16, 17, 23	17
22’	4.89 (1H, dd, 16.1, 1.5)	60.2	15/18, 16, 17, 23	17
23	-	171.2	-	-
24	2.69 (2H, m)	45.6	23, 26	17, 28
25	-	70.0	-	-
26	2.64 (1H, m)	45.1	24, 25, 28	28
26’	2.57 (1H, m)	45.1	27	-
27	-	173.1	-	-
28	1.33 (3H, s)	27.6	24, 25	17, 24, 26

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone E

HRESIMS: *m/z* = 505.2791 $[M + H]^+$, calculated for $[C_{28}H_{40}O_8+H]^+$: 505.2789. $[\alpha]^{20}_{587} = 11.7°$
CD_{MeOH}: $[\theta]_{209} = 2.1815$, $[\theta]_{241} = -12.3941$, $[\theta]_{340} = 7.4622$



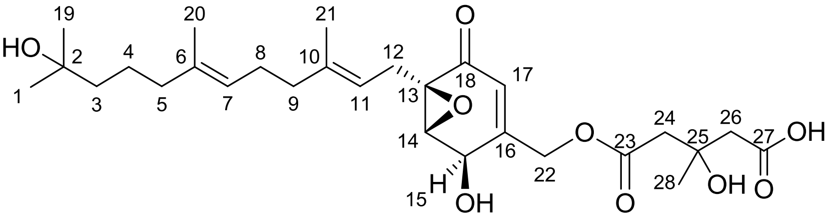
NMR data for yanuthone E

Atom	^{1}H-chemical shift [ppm]/	^{13}C-chemical	HMBC	NOESY
assignment	J coupling constants [Hz]	shift [ppm]	correlations	connectivities
1	1.58 (3H, s)	17.7	2, 3, 19	-
2	-	131.8	-	-
3	5.07 (1H, m)	124.9	1, 4, 5, 19	4, 5, 19
4	2.02 (2H, m)	27.2	2, 6	3
5	1.96 (2H, m)	40.3	3/7, 4, 6, 20,	3, 7
6	-	135.8	-	-
7	5.06 (1H, t, 6.4)	124.9	5/9, 8, 20	5, 9, 20
8	2.06 (2H, m)	27.2	5/9, 6, 7, 10	-
9	1.99 (2H, m)	40.3	6, 7, 8, 11, 21	7, 11
10	-	139.9	-	-
11	5.04 (1H, t, 7.2)	117.7	9, 12, 13, 21	9, 12, 12’, 14
12	2.67 (1H, m)	26.6	10, 11, 13, 14, 18	11, 12’, 14
12’	2.42 (1H, m)	26.6	10, 11, 13, 14, 18	11, 12, 14
13	-	61.3	-	-
14	3.64 (1H, d, 2.8)	60.1	12/12’, 13, 15, 16, 22/22’,	11, 12, 12’ 15
15	4.68 (1H, br. s)	65.8	16, 17	14
16	-	154.6	-	-
17	5.86 (1H, q, 1.4)	121.6	13, 15, 16, 22, 22’	22, 22’, 24
18	-	194.8	-	-
19	1.66 (3H, s)	25.6	1, 2, 3	3
20	1.581 (3H, s)	16.0	5, 6, 7	7
21	1.62 (3H, s)	16.3	9, 10, 11	-
22	4.82 (1H, d, 16.1)	63.6	15, 16, 17, 18, 23	17, 22’
22’	4.76 (1H, d, 16.1)	63.6	15, 16, 17, 18, 23	17, 22
23	-	171.4	-	-
24	2.69 (2H, m)	45.8	23, 25, 26, 28	17
25	-	70.2	-	-
26	2.59 (2H, m)	45.4	24, 25, 27, 28	-
27	-	173.5	-	-
28	1.33 (3H, s)	27.7	25, 26	-

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone J

HRESIMS: *m/z* = 523.2907 [M + H]$^+$, calculated for [$C_{28}H_{42}O_9$+H]$^+$: 523.2901 $[\alpha]_{587}^{20} = 3.6°$



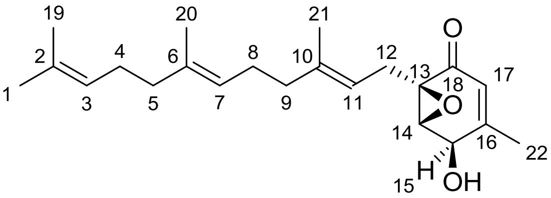
NMR data for yanuthone J

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]	^{13}C-chemical shift [ppm]	HMBC correlations
1	1.28 (3H, m)	29.7	-
2	-	70.9	-
3	1.35 (2H, m)	44.0	4, 19
4	1.43 (2H, m)	23.1	-
5	1.93 (2H, m)	40.5	3, 4, 6, 7, 11, 20
6	-	136.3	-
7	5.08 (1H, tq, 6.3, 1.0)	124.4	-
8	2.09 (2H, m)	26.7	6, 7, 9, 10
9	2.02 (2H, m)	40.3	7, 8, 10, 11, 21
10	-	139.6	-
11	5.03 (1H, tq, 6.3, 1.0)	117.5	-
12	2.45 (1H, m)	26.5	10, 11, 13, (18)
12’	2.68 (1H, m)	26.5	10, 11, 13
13	-	61.1	-
14	3.63 (1H, d, 2.8)	59.9	13, 15, 16
15	4.69 (1H, m)	65.7	-
16	-	154.7	-
17	5.86 (1H, q, 1.5)	121.5	13, 15, 22
18	-	(194.2)	-
19	1.12 (3H, s)	29.2	1, 2, 3
20	1.58 (3H, s)	15.9	5, 6, 7
21	1.62 (3H, s)	16.2	9, 10, 11
22	4.83 (1H, d, 16.0)	63.6	16, 17, 23
22’	4.76 (1H, d, 16.0)	63.6	16, 17, 23
23	-	171.2	-
24	2.681 (2H, m)	43.7	23, 25, 26, 28
25	-	70.9	-
26	2.61 (2H, m)	45.1	24, 25, 27, 28
27	-	173.6	-
28	1.33 (3H, s)	27.5	25, 26

Table S4, related to Figure 6. Spectroscopic data. *Continued*

7-deacetoxyyanuthone A

HRESIMS: *m/z* = 345.2434 [M + H]$^+$, calculated for [$C_{22}H_{32}O_3$+H]$^+$: 345.2424 $[\alpha]^{20}_{587} = 20.0°$
CD_{MeOH}: $[\theta]_{206}$ = -13.6324, $[\theta]_{242}$ = -25.0218, $[\theta]_{335}$ = 16.7472



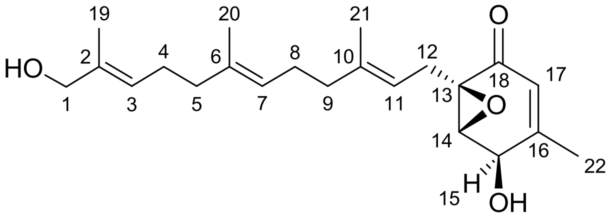
NMR data for 7-deacetoxyyanuthone A

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]
1	1.67 (3H, d 1.0)
2	-
3	5.10 (1H, m)
4	2.07-1.90 (2H, m)
5	2.07-1.90 (2H, m)
6	-
7	5.09 (1H, m)
8	2.07-1.90 (2H, m)
9	2.07-1.90 (2H, m)
10	-
11	5.05 (1H, m)
12	2.72 (1H, dd, 15.2, 7.9)
12’	2.39 (1H, dd, 15.2, 6.7)
13	-
14	3.62 (1H, d, 2.89)
15	4.49 (1H, br. s)
16	-
17	5.70 (1H, p, 1.5)
18	-
19	1.60 (3H, s)
20	1.59 (3H, s)
21	1.64 (3H, s)
22	2.07-1.90 (3H, m)

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone F

HRESIMS: *m/z* = 361.2373 [M + H]$^+$, calculated for [$C_{22}H_{32}O_4$+H]$^+$: 361.2373 $[\alpha]^{20}_{587} = 14.2°$



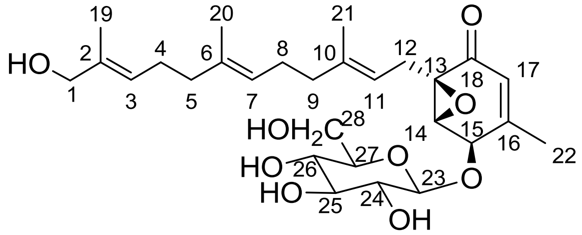
NMR data for yanuthone F

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]	^{13}C-chemical shift [ppm]	HMBC correlations	NOESY connectivities
1	3.84 (2H, br. s)	68.3	2, 3, 19	3, 5
2	-	136.2	-	-
3	5.33 (1H, tq, 7.4, 1.5)	125.3	1, 19	1, 4, 5
4	2.11 (2H, m)	26.8	2, 5, 6	3, 7
5	1.99 (2H, m)	40.1	2/6, 4/8, 20	1, 3, 7
6	-	136.2	-	-
7	5.09 (1H, tq, 6.9, 1.1)	124.9	4/8, 5/9, 20	4, 5, 8, 9
8	2.07 (2H, m)	26.8	6, 5/9, 10	7
9	2.01 (2H, m)	40.1	6, 8/12, 10, 21	7, 11, 14
10	-	139.7	-	-
11	5.04 (1H, tq, 7.3, 1.2)	118.0	8/12, 9, 21	9, 12, 12', 14
12	2.70 (1H, dd, 15.1, 8.1)	27.0	10, 11, 13	11, 12', 14
12'	2.37 (1H, m)	27.0	10, 11, 13, 18	11, 12, 14
13	-	61.4	-	-
14	3.60 (1H, d, 2.7)	60.2	13, 15, 16	9, 11, 12, 12', 15
15	4.48 (1H, m)	67.6	16, 17	14, 22
16	-	158.8	-	-
17	5.68 (1H, m)	123.2	13, 15, 22	22
18	-	194.9	-	-
19	1.59 (3H, s)	13.5	1, 2, 3, 4	-
20	1.591 (3H, s)	16.1	5, 6, 7	-
21	1.62 (3H, s)	16.1	6, 10, 11, 13	-
22	1.92 (2H, br. s)	20.0	15, 16	15, 17

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone G

HRESIMS: *m/z* = 523.2917 $[M + H]^+$, calculated for $[C_{28}H_{42}O_9+H]^+$: 523.2901. $[\alpha]_{587}^{20} = 19.6°$



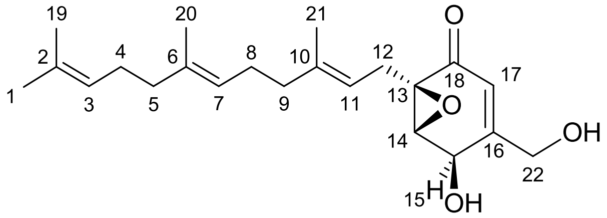
NMR data for yanuthone G

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]	^{13}C-chemical shift [ppm]	HMBC correlations	NOESY connectivities
1	3.86 (2H, s)	68.2	2, 3, 19	3, 19
2	-	135.9	-	-
3	5.34 (1H, t, 6.3)	125.2	1, 4, 5, 19	1
4	2.10 (2H, m)	26.8	2/6, 3, 5	5, 20
5	1.99 (2H, m)	40.0	2/6, 4/8, 7, 20	4, 7
6	-	135.9	-	-
7	5.10 (1H, t, 6.3)	124.9	4/8, 5/9, 20	5, 9
8	2.06 (2H, m)	26.9	5/9, 6, 10	9, 20, 21
9	2.02 (2H, m)	40.0	6, 8/12, 10, 11, 21	7, 8, 11, 14
10	-	139.7	-	-
11	5.04 (1H, t, 6.7)	117.9	8/12, 9, 21	9, 12, 12’, 14, 15
12	2.67 (1H, m)	26.9	10, 11, 13, 18	11, 14, 21
12'	2.37 (1H, dd, 14.8, 6.2)	26.9	10, 11, 13, 18	11, 14, 21
13	-	61.3	-	-
14	3.82 (1H, br. s)	59.7	15, 16	9, 11, 12, 12’, 15, 17, 21
15	4.59 (1H, m)	76.2	23	11, 14, 17
16	-	157.0	-	-
17	5.71 (1H, br. s)	123.9	13, 15, 22	14, 15, 27
18	-	194.4	-	-
19	1.591 (3H, s)	13.7	1, 2, 3	1
20	1.59 (3H, s)	15.9	5, 6, 7	4, 8
21	1.62 (3H, s)	16.3	9, 10, 11	8, 12, 12’, 14
22	1.97 (3H, m)	20.1	15, 16	26, 27
23	4.56 (1H, m)	105.7	16	24, 25, 26, 27, 28, 28’
24	3.25 (1H, br. s)	74.6	23, 26/27	23, 25, 26, 27
25	3.37 (1H, m)	77.2	-	23, 24, 26, 27, 28, 28’
26	3.35 (1H, m)	71.2	-	22, 23, 24, 25, 27, 28, 28’
27	3.35 (1H, m)	77.2	-	17, 22, 23, 24, 25, 26, 28, 28’
28	6.77 (1H, d, 10.7)	62.6	-	23, 25, 26, 27, 28’
28’	3.66 (1H, d, 10.7)	62.6	-	23, 25, 26, 27, 28

Table S4, related to Figure 6. Spectroscopic data. *Continued*

22-deacetylyanuthone A

HRESIMS: *m/z* = 361.2372 [M + H]$^+$, calculated for [$C_{22}H_{32}O_4$+H]$^+$: 361.2373 $[\alpha]_{587}^{20} = 30.6°$



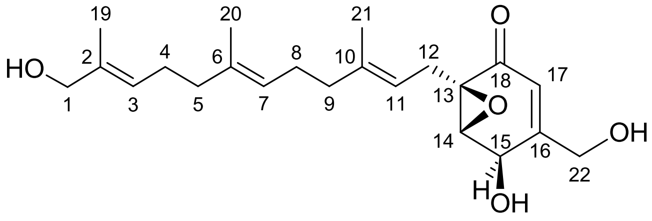
NMR data for 22-deacetylyanuthone A

Atom assignment	**^{1}H-chemical shift [ppm]/ J coupling constants [Hz]**	**^{13}C-chemical shift [ppm]**	**HMBC correlations**	**NOESY connectivities**
1	1.66 (3H, s)	25.8	2, 3, 19	-
2	-	132.0	-	-
3	5.10 (1H, m)	124.9	1, 4, 5, 19	4, 5, 19
4	2.07 (2H, m)	26.8	2, 5, 6	3, 5, 19, 20
5	1.98 (2H, m)	40.1	4/8, 6, 3/7, 20	3, 4, 7
6	-	136.0	-	
7	5.10 (1H, m)	124.9	4/8, 5/9, 20	8, 5/9, 20
8	2.09 (2H, m)	26.8	5/9, 6, 7, 10	7, 9, 20, 21
9	2.03 (2H, m)	40.1	7, 10, 11, 13, 8/12, 21	7, 8, 11
10	-	140.2	-	-
11	5.05 (1H, ddd, 7.93, 6.71, 1.22)	118.0	9, 8/12, 13, 21	9, 12, 12', 14, 21
12	2.70 (1H, dd, 15.3, 7.9)	26.8	10, 13, 17, 18	11, 12', 14, 21
12'	2.42 (1H, dd, 15.3, 6.7)	26.8	10, 13, 17, 18	11, 12, 14, 21
13	-	61.6	-	-
14	3.62 (1H, d, 2.8)	60.2	12,15, 16, 22	11, 12, 12', 15
15	4.64 (1H, m)	66.2	13, 16, 17, 18	14, 22, 22'
16	-	160.8	-	-
17	5.87 (1H, q, 1.73)	119.7	15, 16, 22	22, 22'
18	-	195.1	-	-
19	1.601 (3H, s)	17.7	1, 2, 3	3, 4
20	1.60 (3H, s)	15.9	5, 6, 7	4, 7, 8
21	1.63 (3H, s)	16.2	8/12, 9, 10, 11, 13	8, 11, 12, 12'
22	4.27 (1H, m)	62.0	15	15, 17
22'	4.21 (1H, m)	62.0	-	15, 17

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone H

HRESIMS: *m/z* = 377.2332 [M + H]$^+$, calculated for [$C_{22}H_{32}O_5$+H]$^+$: 377.2322. $[\alpha]_{587}^{20} = 27.7°$



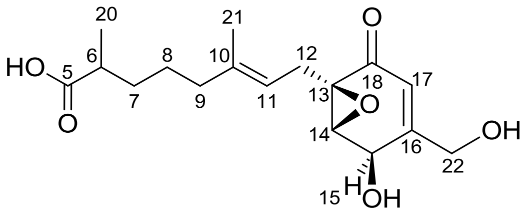
NMR data for yanuthone H

Atom assignment	**^{1}H-chemical shift [ppm]/ J coupling constants [Hz]**	**^{13}C-chemical shift [ppm]**	**HMBC correlations**	**NOESY connectivities**
1	3.85 (2H, s)	68.3	2, 3, 4, 19	3
2	-	136.0	-	-
3	5.34 (1H, m)	125.4	1, 4, 5, 19	1, 5
4	2.11 (2H, m)	26.8	3, 5, 6	-
5	1.99 (2H, m)	40.1	4/8, 6, 7, 12, 20	3
6	-	135.9	-	-
7	5.10 (1H, t, 6.4)	124.9	4/8, 5/9, 20	9
8	2.07 (2H, m)	26.8	5/9, 7, 10	-
9	2.02 (2H, m)	40.1	7, 10, 11, 8/12, 21	7, 11
10	-	139.8	-	-
11	5.05 (1H, t, 6.8)	118.0	8/12, 21	9, 12, 12’, 14
12	2.71 (1H, dd, 15.1, 8.3)	26.8	10, 11, 13, 18	11, 12’, 14
12'	2.40 (1H, dd, 15.1, 6.8)	26.8	20, 11, 13, 18	11, 12, 14
13	-	60.9	-	-
14	3.62 (1H, d, 2.9)	60.4	12, 13, 15, 16	11, 12, 12’ 15
15	4.64 (1H, br. s)	66.1	16, 17	14, 16, 17
16	-	160.6	-	15
17	5.87 (1H, d, 1.5)	119.7	13, 15, 16	15, 22, 22’
18	-	194.9	-	-
19	1.601 (3H, s)	13.5	1, 2	-
20	1.60 (3H, s)	16.1	5/8, 7	-
21	1.63 (3H, s)	16.2	8/12, 10, 11	-
22	4.27 (1H, d, 17.6)	62.0	15, 16, 17, 18	17, 22’
22'	4.21 (1H, m)	62.0	15, 16, 17, 18	17, 22

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone I

HRESIMS: *m/z* = 325.1635 [M + H]$^+$, calculated for [$C_{17}H_{24}O_6$+H]$^+$: 325.1646. $[\alpha]^{20}_{587} = 21.1°$



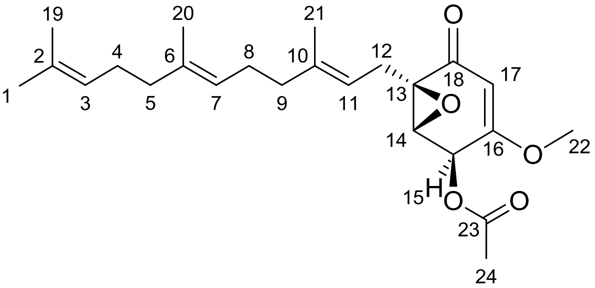
NMR data for yanuthone I

Atom assignment	**^{1}H-chemical shift [ppm]/ J coupling constants [Hz]**	**^{13}C-chemical shift [ppm]**	**HMBC correlations**	**NOESY connectivities**
5	-	177.4	-	-
6	2.28 (1H, m)	38.2	5, 7/7', 8, 20	7, 7', 8, 20
7	1.48 (1H, m)	32.5	5, 6, 8, 20	6, 7', 8, 9, 20
7'	1.26 (1H, m)	32.5	5, 6, 8, 20	6, 7, 8, 9, 21
8	1.32 (2H, m)	24.5	7/7', 9, 10	6, 7, 7', 9, 11, 20
9	1.92 (2H, t, 7.1)	38.7	7/7', 8, 10, 11, 21	7, 7', 8, 11, 14
10	-	137.9	-	-
11	5.00 (1H, t, 7.1)	116.9	9, 12/12', 13, 21	8, 9, 12, 12', 14
12	2.63 (1H, dd, 15.1, 7.9)	25.7	10, 11, 13, 14, 17, 18	12', 14, 17, 21
12'	2.32 (1H, m)	25.7	9, 10, 11, 13, 14, 17, 18	12, 14, 17, 21
13	-	60.0	-	-
14	3.59 (1H, d, 2.4)	59.1	12/12', 13/22/22', 15, 16	9, 11, 12, 12', 15, 17, 21
15	4.61 (1H, br. S)	64.3	16, 17	14, 22'
16	-	162.2	-	-
17	5.82 (1H, d, 1.6)	117.4	13/22/22', 14, 15, 16, 18	22'
18	-	193.7	-	-
20	1.02 (3H, d, 6.7)	16.7	5, 6, 7	6, 7, 7', 8
21	1.56 (3H, s)	15.6	9, 10, 11, 12/12', 13	7', 9, 12, 12', 14
22	4.21 (1H, d, 18.5)	60.0	15, 16, 17, 18	22'
22'	4.09 (1H, d, 18.5)	60.0	15, 16, 17, 18	15, 17, 22
-COOH	12.01 (1H, br. s)	-		

Table S4, related to Figure 6. Spectroscopic data. *Continued*

Yanuthone X_1

HRESIMS: *m/z* = 403.2482 $[M + H]^+$, calculated for $[C_{17}H_{24}O_6+H]^+$: 403.2479 $[\alpha]_{587}^{20} = 2.5°$



NMR data for yanuthone X_1

Atom assignment	^{1}H-chemical shift [ppm]/ J coupling constants [Hz]	^{13}C-chemical shift [ppm]	HMBC correlations	NOESY correlations
1	1.66 (3H, s)	25.6	2, 3, 19	3
2	-	132.2	-	-
3	5.08 (1H, m)	125.0	5	1, 4/5
4	2.05 (2H, m)	27.6	3, 5, 6	3, 5
5	2.00 (2H, m)	40.3	6, 4/8, 7, 11	3, 4
6	-	135.6	-	-
7	5.08 (1H, m)	125.0	5/9, 20	-
8	2.07 (2H, m)	27.6	5/9, 6, 7	9
9	1.95 (2H, m)	40.3	7, 8, 10, 11, 21	8, 11
10	-	139.8	-	-
11	5.01 (1H, t, 7.3)	117.5	9, 21	9, 12, 12’
12	2.78 (1H, dd, 15.3, 8.2)	26.2	10, 11, 13, 14	11, 12’
12’	2.45 (1H, dd, 15.3, 6.7)	26.2	10, 11, 13, 14, 18	11, 12
13	-	60.5	-	-
14	3.59 (1H, d, 3.1)	56.3	15, 16	15
15	5.95 (1H, d, 3.1)	66.6	16	14
16	-	168.3	-	-
17	5.30 (1H, s)	100.3	13, 15, 16	22
18	-	193.7	-	-
19	1.57 (3H, s)	17.6	1, 2, 3	-
20	1.571 (3H, s)	15.8	5, 6	-
21	1.61 (3H, s)	16.3	9, 10, 11	-
22	3.67 (3H, s)	57.4	16	17
23	-	170.7	-	-
24	2.14 (3H, s)	20.6	23	-

Table S5, related to Figures 2, 3, and 4. Fungal strains.

Name	Organism	Genotype
IBT 29539	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ
OE-*yanA*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ S1::P*gpdA*::*yanA*::T*trpC*::*argB*
OE-*yanG*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ IS1::P*gpdA*::*yanA*::T*trpC*::*argB*, pDHX2::*yanG*
OE-*yanC*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ IS1::P*gpdA*::*yanA*::T*trpC*::*argB*, pDHX2::*yanC*
OE-*yanB*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ IS1::P*gpdA*::*yanA*::T*trpC*::*argB*, pDHX2::*yanB*
OE-*yanD*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ IS1::P*gpdA*::*yanA*::T*trpC*::*argB*, pDHX2::*yanD*
OE-*yanE*	*A. nidulans*	*argB2*, *pyrG89*, *veA1*, *nkuA*Δ IS1::P*gpdA*::*yanA*::T*trpC*::*argB*, pDHX2::*yanE*
KB1001	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG*
*yanA*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanA*Δ
*yanG*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanG*Δ
*yanI*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanI*Δ
*yanC*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanC*Δ
*yanB*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanB*Δ
*yanH*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanH*Δ
*yanD*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanD*Δ
*yanE*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanE*Δ
*yanF*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanF*Δ
*yanR*Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG yanR*Δ
44959Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG* ASPNI_DRAFT44959Δ
44960Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG* ASPNI_DRAFT44960Δ
44971Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG* ASPNI_DRAFT44971Δ
44972Δ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG* ASPNI_DRAFT44972Δ
ClusterΔ	*A. niger*	*pyrG*Δ *kusA*::*AFpyrG* ASPNI_DRAFT44958-44972Δ

Table S6, related to Figures 2, 3, and 4. Primers used in the study. See details in experimental section.

#	Primer name	Sequence 5’→ 3’
1	44965-fw	AGAGCGAUATGCCAGGCCTTGTACAC
2	44965-rv	TCTGCGAUTTAAGCATCCAGCTCCTTTGT
3	44963_ORF_FW	AGAGCGAUATGGACCGTATCGACGTACACC
4	44963_ORF_RV	TCTGCGAUCTAGGTACTATAAGTATGAACACGAGACTG
5	44964_ORF_FW	AGAGCGAUATGTCTACTACTAAGCGCTCGGTAAC
6	44964_ORF_RV	TCTGCGAUCTAGTATACTTTCATGGGTGCGTGA
7	54844_ORF_FW	AGAGCGAUCGGGCTAGACTTTCTCTTCCTAAG
8	54844_ORF_RV	TCTGCGAUATGGCGCTTGTTCATCTGACT
9	127904_ORF_FW	AGAGCGAUATGGTCAAGTTTTTTCAGCCCA
10	127904_ORF_RV	TCTGCGAUCTAACGGAACTGGGGAGGAA
11	192604_ORF_FW	AGAGCGAUTACTTTGCGACTACCTGCCATG
12	192604_ORF_RV	TCTGCGAUCTACTCCGACTTTTCACCTTTGG
13	hph-1003-Fw	AGCCCAATAUGCTAGTGGAGGTCAACACATCA
14	hph-1003-Rv	ATTACCTAGUCGGTCGGCATCTACTCTATT
15	44965-chk-usF	CAGTTGACTAGACTAGGAACGGTCA
16	44965-chk-dsR	AACGACCATGATGCTTGTTCAG
17	44965_US-FW	GGGTTTAAUATGACTCCACATCATCTTCCACAC
18	44965_US-RV	ATATTGGGCUGATGGTGTGTACAAGGCCTGG
19	44965_DS-FW	ACTAGGTAAUGACTGTTATGCATTGAATTTGAGC
20	44965_DS-RV	GGTCTTAAUAGATCCTGACGCTCATATCTGCT
21	44964_US-FW	GGGTTTAAUGGTCTTTCCGACACGTAAGTCTG
22	44964_US-RV	ATATTGGGCUTGGACCTCAATGGCCGCT
23	44964_DS-FW	ACTAGGTAAUGCGAGTATGAAGAAGGTGGATGA
24	44964_DS-RV	GGTCTTAAUATTCAGGGTCTTGAGATTGGC
25	44963_US-FW	GGGTTTAAUAAGTCCTCCCACGTCGGAG
26	44963_US-RV	ATATTGGGCUAGAATCTAAACCTTGTCTCTTCGCT
27	44963_DS-FW	ACTAGGTAAUGAACGTTTGATTGGTAATGGATGT
28	44963_DS-RV	GGTCTTAAUTCATCCACCTTCTTCATACTCGC
29	54844_US-FW	GGGTTTAAUGTAGAATAACAGCTACCTCGAATTTGA
30	54844_US-RV	ATATTGGGCUACGTGGTGCGTAAGCAGACAT
31	54844_DS-FW	ACTAGGTAAUCCTGCTGAATAAACACGAAGG
32	54844_DS-RV	GGTCTTAAUATGGACCGTATCGACGTACACC
33	44960_US-FW	GGGTTTAAUTGAGTACCTATCCACTCTTCCTGG
34	44960_US-RV	ATATTGGGCUGATGGAGTGTGAAGCCAATGAG
35	44960_DS-FW	ACTAGGTAAUTCATTCTAAAATTGGCGTCTTCA
36	44960_DS-RV	GGTCTTAAUCTACTGCCGCCGTCACTATCTA
37	193092_US-FW	GGGTTTAAUCATCGACATCTCTCTGCCCAT
38	193092_US-RV	ATATTGGGCUGAAAGCTGGTTGGAAGTATAAGTGG
39	193092_DS-FW	ACTAGGTAAUTGTGCAGCGGTATTGACTTCA

40	193092_DS-RV	GGTCTTAAUCACGGAGTTATTTTCCACGCT
41	44967_US-FW	GGGTTTAAUCGTTGGCATGACAGTCTTCAA
42	44967_US-RV	ATATTGGGCUGTCTGCCATCACAACCAGTTTG
43	44967_DS-FW	ACTAGGTAAUAGCCATGTTGCCAGACACAGT
44	44967_DS-RV	GGTCTTAAUACTACCATCTCGTAACCGTCCTAG
45	127904_US-FW	GGGTTTAAUGACCGACTCTACACTACCGTTCC
46	127904_US-RV	ATATTGGGCUATTGAACTGGTAAACATGCCATG
47	127904_DS-FW	ACTAGGTAAUTAGCCCTAGGACGGTTACGAG
48	127904_DS-RV	GGTCTTAAUAACCAACTTTGTTCCATTCTATCG
49	192604_US-FW	GGGTTTAAUGACACATCGTATTGATGACGACC
50	192604_US-RV	ATATTGGGCUCATGGCAGGTAGTCGCAAAG
51	192604_DS-FW	ACTAGGTAAUAAGAGAATACGGAACACATTGACC
52	192604_DS-RV	GGTCTTAAUCGGTCCAACAGTGAGGGTCT
53	44970_US-FW	GGGTTTAAUCGTTGATAATTCCAATTCCAATTC
54	44970_US-RV	ATATTGGGCUCGTCGAAGATGACCTGATTTG
55	44970_DS-FW	ACTAGGTAAUCGGGTTATCACTGTATCAATATCG
56	44970_DS-RV	GGTCTTAAUGCTACTACTATGCCGACTGCGT
57	44971_US-FW	GGGTTTAAUGGCCACACCTCAAGTTTGTATG
58	44971_US-RV	ATATTGGGCUCGGGATTGGAGTGCTCTAGTT
59	44971_DS-FW	ACTAGGTAAUGTTGGCTGAGAGTCAGGGTTAG
60	44971_DS-RV	GGTCTTAAUCCATTAGCTTCGGAACACTGG
61	44959_US-FW	GGGTTTAAUCCTTGTATTCATATCAATTGCGA
62	44959_US-RV	ATATTGGGCUATGTGACAATGAAGAATGGTACG
63	44959_DS-FW	ACTAGGTAAUGGAAAGGATGTTCCAAACAGTT
64	44959_DS-RV	GGTCTTAAUCTTTGTTGATTACTAGTCGTAATCATATG
65	44961_US-FW	GGGTTTAAUTGTCATGTTGTATCGGAGTGTTTAG
66	44961_US-RV	ATATTGGGCUTGTAGCACAAGTGTCTCACTAGTAAATAG
67	44961_DS-FW	ACTAGGTAAUGATTGGAAGTATCCCACAGTCTG
68	44961_DS-RV	GGTCTTAAUGAGAACACCGATCTCCGACGTGGGA
69	44972_US-FW	GGGTTTAAUGACGCAGTCGGCATAGTAGTAG
70	44972_US-RV	ATATTGGGCUGGAGAAGTGGTCAAACTTGTTTCA
71	44972_DS-FW	ACTAGGTAAUACAGGTGATTAAGATGCAAGGCT
72	44972_DS-RV	GGTCTTAAUCTTGCATCATCCGTAATTATGCT
73	Upst-HygR-N	CTGCTGCTCCATACAAGCCAACC
74	Dwst-1003HygF-N	GACATTGGGGAGTTCAGCGAGAG
75	ampR_PM_FW	AGCGCTACAUAATTCTCTTACTGTCATGCCATCC
76	ampR_PM_RV	ATGTAGCGCUGCCATAACCATGAGTGATAACACTG
77	ori_coli_FW	ATCCCCACUACCGCATTAAGACCTCAGCG
78	ampR_RV	AGCTGCTUCGTCGATTAAACCCTCAGCG
79	p71_prom-ter_short_usF	AGCCCAATAUTAAGCTCCCTAATTGGCCC
80	p71_prom-ter_dsR	ATTACCTAGUGGGCGCTTACACAGTACA
81	argB_FW	ACTAGGTAAUATCGCGTGCATTCCGCGGT

82	pyrG_FW_C1	ACTAGGTAAUATGACATGATTACGAATTCGAGCT
83	pyrG_RV_G16	AGTGGGGAUGCCTCAATTGTGCTAGCTGC
84	ccdB-camR-fw	AGAGCGAUCGCAGAAGCCTACTCGCTATTGTCCTCA
85	ccdB-camR-rv	TCTGCGAUCGCTCTTGCGCCGAATAAATACCTGT
86	AMA1_FW	AAGCAGCUGACGGCCAGTGCCAAGCT
87	AMA1_RV	ATATTGGGCUGGAAACAGCTATGACCATGAGATCT
88	AMA-3'-Fw	ACCCCAAUGGAAACGGTGAGAGTCCAGTG
89	AMA-5'-Rv	ATTGGGGUACTAACATAGCCATCAAATGCC
90	ANIG-actA-qFw	GTATGCAGAAGGAGATCACTGCTCT
91	ANIG-actA-qRv	GAGGGACCGCTCTCGTCGT
92	ANIG-hhtA-qFw	CTTCCAGCGTCTTGTCCGTG
93	ANIG-hhtA-qRv	GCTGGATGTCCTTGGACTGGAT
94	44959-qFw	GGCAAAGTTCTAGTCATCGACGA
95	44959-qRv	CATATATCCCAGAGGCGGACAC
96	44960-qFw	GATAGAGGAGATGAGGAAGAGAGGCT
97	44960-qRv	CCTTGGGTACCATTCACAGTCAG
98	44961-qFw	ACATGGACCACCGAGTAGCGT
99	44961-qRv	TAGGGTGTGCGAGAATATCACTTG
100	54844-qFw	CGATGAAGATGGCAATCCCAT
101	54844-qRv	CTATGGCATCGCATACTGAGAAAGA
102	44963-qFw	GCGAGGAGGTAGAAAAGGCAAT
103	44963-qRv	TGAACACGAGACTGGAGTACGGA
104	44964-qFw	TCTTCTGGATACTGGGAATTGGAG
105	44964-qRv	GCGTGATGCACCCTCAACA
106	44965-qFw	TTGTCTGTCAAGGAGGACGAGATT
107	44965-qRv	CTTCACCAAATGCTGCACAGTC
108	193092-qFw	ATCACGGCAAAGAGAGCCAAGT
109	193092-qRv	GAACTGTGGCACGACCATGTC
110	44967-qFw	TGCTGGCTGCTAAGGATTGATG
111	44967-qRv	ATGTTCCAACGCAATGAACAAC
112	127904-qFw	ATGAGCAACATGCTCCCACTACAT
113	127904-qRv	CAGTGATGGTCTTATCCGCCAG
114	192604-qFw	GCCTAATCCTGGGCATCGTG
115	192604-qRv	CTGTGCTCCCCGATCTGCA
116	44970-qFw	TCTCAGGGTGTCCATCTTCCGT
117	44970-qRv	CGACACGAAATAGGCATCATTCT
118	44971-qFw	ATCTACTCCGGCTCCTGCGAT
119	44971-qRv	ACTCGCAAACAACTTCATTGCTC
120	44972-qFw	AGGTGACTCGAACTGGTATGCTG
121	44972-qRv	CAGAATATACTCGATATGATCGCCTC

6.4 Paper 4 – Combining UHPLC-High Resolution MS and Feeding of Stable Isotope Labeled Polyketide Intermediates for Linking Precursors to End Products

Klitgaard, A., Frandsen, R. J. N., Holm, D. K., Knudsen, P. B., Frisvad, J. C., & Nielsen, K. F.

Paper accepted in Journal of Natural Products (2015)

JOURNAL OF NATURAL PRODUCTS

Article

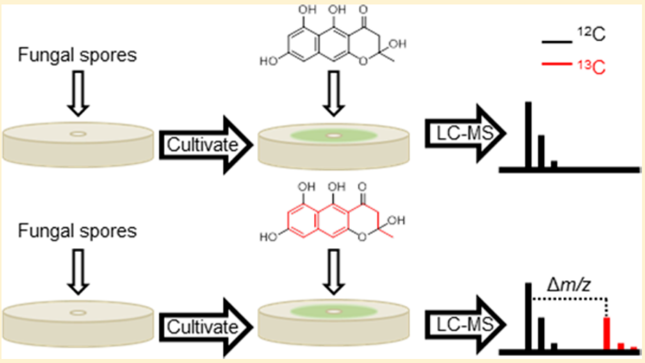
pubs.acs.org/jnp

Combining UHPLC-High Resolution MS and Feeding of Stable Isotope Labeled Polyketide Intermediates for Linking Precursors to End Products

Andreas Klitgaard, Rasmus J. N. Frandsen, Dorte K. Holm, Peter B. Knudsen, Jens C. Frisvad, and Kristian F. Nielsen*

Department of Systems Biology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

Supporting Information



ABSTRACT: We present the results from stable isotope labeled precursor feeding studies combined with ultrahigh performance liquid chromatography-high resolution mass spectrometry for the identification of labeled polyketide (PK) end-products. Feeding experiments were performed with $^{13}C_8$-6-methylsalicylic acid (6-MSA) and $^{13}C_{14}$-YWA1, both produced in-house, as well as commercial $^{13}C_7$-benzoic acid and $^2H_7$-cinnamic acid, in species of *Fusarium, Byssochlamys, Aspergillus,* and *Penicillium*. Incorporation of 6-MSA into terreic acid or patulin was not observed in any of six evaluated species covering three genera, because the 6-MSA was shunted into (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid. This indicates that patulin and terreic acid may be produced in a closed compartment of the cell and that (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid is a detoxification product toward terreic acid and patulin. In *Fusarium* spp., YWA1 was shown to be incorporated into aurofusarin, rubrofusarin, and antibiotic Y. In *A. niger*, benzoic acid was shown to be incorporated into asperrubrol. Incorporation levels of 0.7−20% into the end-products were detected in wild-type strains. Thus, stable isotope labeling is a promising technique for investigation of polyketide biosynthesis and possible compartmentalization of toxic metabolites.

Filamentous fungi are a rich source of bioactive metabolites, including the polyketides (PKs), which constitute one of the largest groups of natural products. PKs include important pharmaceutics such as lovastatin, mycophenolic acid, and griseofulvin.[1] Three of the five major economically important mycotoxins are also of PK origin: aflatoxins, zearalenones, and fumonisins, aflatoxins being the most carcinogenic natural compounds currently known and zearalenones being highly estrogenic.[2]

With the rapid decrease in the cost of fungal genome sequencing, a much more efficient foundation for elucidation of biosynthetic pathways is now available.[3] This can be used for direct studies of biosyntheses, for improving cell factories via metabolic engineering, or for product yield optimization.[4] Alternatively, biosynthetic clusters can be transferred to a heterologous host for higher yields, which is often vital for producing sufficient amounts of a new drug candidate for toxicological and pharmacological evaluation. However, linking of fungal biosynthetic genes to their products by genetic engineering approaches is still very time-consuming. This is mainly due to the difficulties with bioinformatic prediction of the products being synthesized by iterative polyketide synthases (PKSs).[5] In a recent study we have used feeding experiments and ultrahigh performance liquid chromatography-high resolution mass spectrometry (UHPLC-HRMS) to show that $^{13}C_8$-labeled 6-methylsalicylic acid (6-MSA, **1**; Chart 1) and not the previously hypothesized precursor shikimic acid was a central building block in formation of yanuthone D in *Aspergillus niger.*[6]

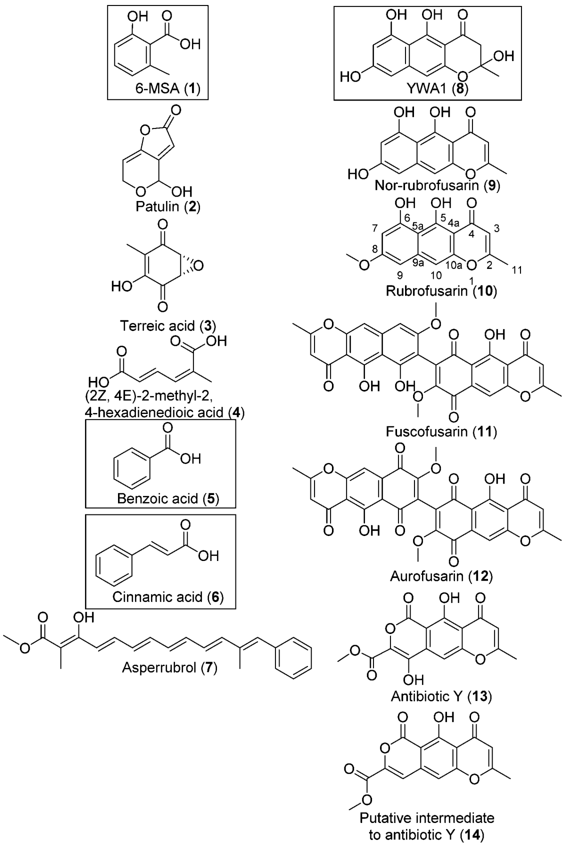
The earliest biosynthetic studies using labeled precursors were based on ^{14}C and other radioactive isotopes to enable detection.[7] However, this has been overtaken by NMR spectroscopy using stable isotope labeled (SIL) compounds, where a (usually) ^{13}C-, ^{15}N-, ^{2}H-, or ^{34}S-labeled precursor is used. NMR data can also reveal labeling positions in the final products.[8] The downside of NMR spectroscopy is the poorer sensitivity compared with liquid chromatography mass spectrometry (LC-MS), requiring time-consuming isolation of SIL-labeled product(s) as well as much higher consumption of SIL precursors. However, MS may not yield information on the position of the labeling unless MS/MS can be used to form assignable labeled fragments of the compound of interest.

SIL precursor feeding has been used in several studies of the aflatoxin pathway,[9] the asticolorin pathway (both NMR based),[10] and as noted above the yanuthone D pathway in *A. niger* (MS based).[6,11]

Received: December 5, 2014

DOI: 10.1021/np500979d

Chart 1. Chemical Structures of Compounds Investigated[a]



[a]Compounds are arranged according to biosynthetic origin. The boxed compounds correspond to the SIL compounds used in the study.

To ease interpretation of LC-MS results from these labeling experiments, it is advantageous to use a 100% labeled precursor that will result in formation of one distinct product isotopomer. Furthermore, the mass shift induced should preferably be large enough to be free of interference from the natural isotopomers of the target.

For MS investigation of pathways where SIL precursors are not available, the organism could be cultivated using fully isotope labeled media leading to nearly complete isotope enrichment in a so-called reciprocal or inverse labeling experiment.[12−14] This approach requires a minimal medium where all C, N, H, or S sources can be labeled, which is not available for complex media containing components that are often required to induce expression of fungal secondary metabolite pathways.

In recent studies, we were able to achieve close to 20% labeling of PK end-products in *A. niger* using $^{13}C_8$-6-MSA produced by heterologous expression of a 6-MSAS gene (*yanA*) in *Aspergillus nidulans.*[6] Based on these results, we speculated that it would be of scientific value to produce numerous SIL precursors this way and use them for examination of various biosynthetic pathways. To test the applicability of this strategy, we used two commercially available precursors [benzoic acid (**5**) and cinnamic acid (**6**)] and two in-house produced precursors [6-MSA and YWA1 (**8**)] to investigate a number of pathways where these four compounds are known or suspected to be precursors to other compounds.

Since labeled 6-MSA was already available, it seemed obvious to examine other known compounds biosynthesized using 6-MSA as precursor. A well-known compound is the mycotoxin patulin (**2**), for which the biosynthesis has already been elucidated.[15] Patulin is found in many species throughout three different genera (*Byssochlamys, Penicillium,* and *Aspergillus*), making it an excellent case for testing for broad versatility of labeling across organisms. The compound terreic acid (**3**,[16] Figure 1), produced by *A. terreus* ATCC 20542 (the original mevinolin producer)[17,18] is related to patulin and is also biosynthesized from a 6-MSA precursor.[19] Thus, terreic acid was also selected for investigation.

A. niger is a producer of numerous PKs including asperrubrol (**7**).[20] It has previously been hypothesized that cinnamic acid is a precursor to asperrubrol.[21] Cinnamic acid is a known precursor of benzoic acid in *Phanerochaete chrysosporium*,[22] which means that benzoic acid might also be used to investigate the biosynthesis of cinnamic acid. Because both cinnamic acid and benzoic acid were commercially available as SIL compounds, feeding experiments were performed using both.

The PK YWA1 (**8**)[23] is a key precursor to several different compounds in a variety of different fungal species; in *A. nidulans* YWA1 (produced by WA, encoded by *wA*) is the precursor to the green melanin responsible for pigmentation of conidia.[23] In *A. niger*, YWA1 (produced by AlbA, encoded by *albA*) is also the precursor to conidial pigment; however here the YWA1 is converted into 1,8-dihydroxynaphthalene (1,8-DHN) by chain shortening, after which the 1,8-DHN is polymerized into black melanin. YWA1 is also the precursor to the naphtho-*γ*-pyrones, of which the predominant compounds are the aurasperones.[24,25]

In *Fusarium graminearum*, YWA1 is the first stable intermediate formed during biosynthesis of the red pigment aurofusarin (**12**).[26−28] In *F. graminearum*, YWA1 is biosynthesized by PKS12,[29] an orthologue of the WA PKS in *A. nidulans*,[24] resulting in the formation of a nonreduced heptaketide. Folding of the heptaketide can result in the formation of either YWA1 or isocoumarins.[27] After release from the PKS, YWA1 is converted into nor-rubrofusarin (**9**), rubrofusarin (**10**), 9-hydroxyrubrofusarin, and finally the dimers fuscofusarin (**11**) and aurofusarin.[29]

Antibiotic Y (**13**) (avenacein Y) was first isolated from *F. avenaceum* in 1986, and although its biosynthetic pathway is unknown,[30] it displays several structural features in common with YWA1 and rubrofusarin. This suggest that it may also be formed via the nonreducing polyketide biosynthetic pathway.[5] The carbon backbone of antibiotic Y includes a lactone, which is atypical for nonreduced polyketides, and in this study, we hypothesize that it is formed either by the fusion of a tri- and tetraketide or by a previously undescribed carbon backbone cleavage of YWA1 followed by recondensation into a lactone.

In this study, we have used LC-MS to investigate the biosynthetic pathways of different filamentous fungi using SIL precursors. Both well-known metabolites such as patulin and terreic acid and metabolites biosynthesized from undescribed pathways (antibiotic Y and asperrubrol) were investigated to explore advantages and limitations of the approach.

■ RESULTS AND DISCUSSION

$^{13}C_8$-6-MSA Was Not Incorporated into Patulin or Terreic Acid. Feeding experiments were performed using several organisms that were known to produce patulin (*P.*

Table 1. Results from the Labeling Experiments, Where the Highest Determined Degree of Incorporation Is Listed

target compound	producer organism	precursor	time of precursor addition (d)	degree of incorporation (%, average of duplicates)
patulin (**2**)	*P. griseofulvum, P. paneum, P. carneum, A. clavatus, B. nivea*	6-MSA (**1**)	3	ND[a]
(2*Z*,4*E*)-2-methyl, 4-hexadienoic acid (**4**)				45[b]
terreic acid (**3**)	*A. hortai, A. floccosus*	6-MSA (**1**)	3	ND[a]
(2*Z*,4*E*)-2-methyl, 4-hexadienoic acid (**4**)			6	ND[a]
			3	76[c]
			6	58[c]
asperrubrol (**7**)	*A. niger*	Cinnamic acid (**6**)	3	ND[a]
			6	ND[a]
		Benzoic acid (**5**)	3	1.3[d]
			6	ND[a]
aurofusarin (**12**)	*F. avanaceum, F. graminearum*	YWA1 (**8**)	3	1.2[f]
			7	0.3[g]
			10	0.4[g]
antibiotic Y (**13**)			3	ND[a]
			7	0.7[e]
			10	0.4[e]
rubrofusarin (**10**)			3	0.4[g]
			7	10[g]
			10	17[g]
putative intermediate to antibiotic Y (**14**)			3	ND[a]
			7	2.2[e]
			10	2.2[e]

[a]No incorporation detected. [b]*A. clavatus*. [c]*A. floccosus*. [d]*F. avanaceum* cultivated on DFM. [e]*F. avanaceum* cultivated on Bell's medium. [f]*F. graminearum* cultivated on DFM. [g]*F. graminearum* cultivated on Bell's medium.

griseofulvum, *P. paneum*, *P. carneum*, *A. clavatus*, *B. nivea*) or to produce terreic acid (*A. hortai* and *A. floccosus*).

No changes in morphologies or chemical profiles (acquired base peak chromatograms, BPC) were observed for any of the fungi fed with SIL precursors. Chemical analysis showed no signs of incorporated $^{13}C_8$-6-MSA into either patulin or terreic acid. The analysis was conducted by examining extracted ion chromatograms (EIC, ±0.02 Da) corresponding to both the labeled and unlabeled forms of the compounds (Table 2) and comparing these to reference standards of the compounds.

This was a surprise because 6-MSA is a known precursor to both compounds.[15,19] Since chemical analysis showed that the $^{13}C_8$-6-MSA was removed from the medium, we hypothesize that this result could be due to the fungi degrading the 6-MSA as a source of nutrient. Another explanation could be that the enzymatic activities involved in biosynthesis are linked in a manner that does not allow entry of an advanced precursor. A recent paper by Guo et al.[19] showed that (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid is a shunt product in the terreic acid pathway, and we subsequently detected a peak corresponding to the correct accurate mass of this compound in an extract from *A. floccosus*. Investigation of the mass spectrum also revealed the presence of an ion corresponding to one incorporating $^{13}C_7$ (Supporting Information, Figure S1). We define the degree of labeling as

$$\frac{\text{Signal}_{\text{labeled form}}}{\text{Signal}_{\text{labeled form}} + \text{Signal}_{\text{unlabeled form}}}$$

For (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid, the degree of labeling was thus 76% in *A. floccosus* fed after 3 days (Table 1). Interestingly (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid was also found in the extracts from the patulin producers (Table 1), in both labeled and unlabeled form, showing that it is also a shunt product in the patulin biosynthesis. This strongly indicates that it is a result of a detoxification reaction in the cytoplasm and that patulin and terreic acid are produced in defined compartments. This would make sense, since patulin is an antifungal compound. The need for a detoxification process also seems to be important because (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid was detected in amounts corresponding to 10−20% of the produced patulin as determined using UV. To test for compartmentalization, the peptide sequence of the proteins involved in the terreic acid pathway[19] were analyzed in order to predict any membrane bound proteins, using a range of different prediction tools,[31] including TargetP 1.1,[32] PSORT II,[33] and MultiLoc2.[34] However, no conclusive results were returned on whether the proteins are membrane bound.

Benzoic Acid Is a Precursor to Asperrubrol in *A. niger*. Asperrubrol biosynthesis in *A. niger* was investigated by addition of the two proposed precursors, cinnamic acid and benzoic acid. After feeding with $^2H_7$-cinnamic acid, no changes in morphologies or the BPCs were observed (data not shown).

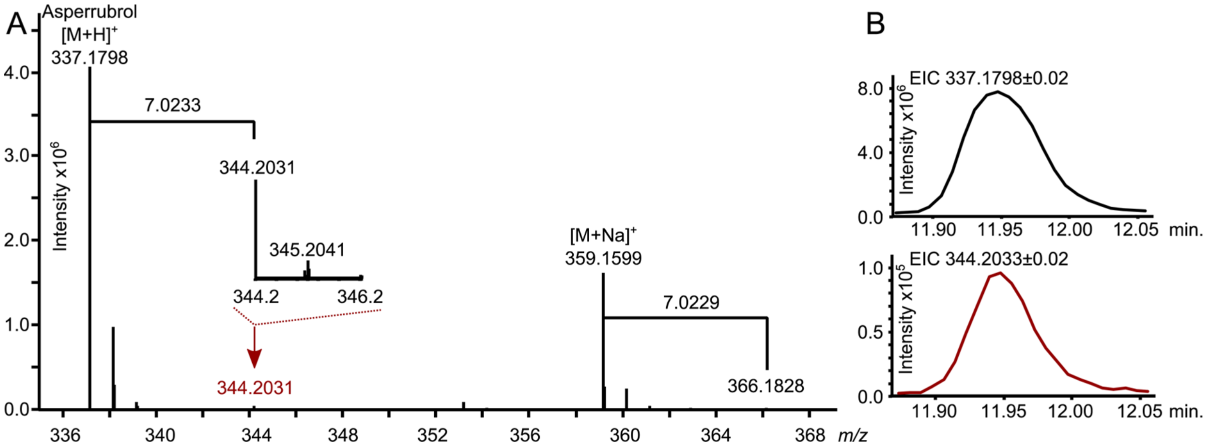


Figure 1. (A) Mass spectrum extracted at RT 12.0 min contained the $[M + H]^+$ (m/z 337.1798, mass deviation m/z 0.06 ppm) and $[M + Na]^+$ (m/z 359.1599). Mass shift of 7.0233 Da (m/z 344.2031, mass deviation 0.60 ppm) suggests incorporation of $^{13}C_7$ (red arrow). (B) EICs corresponding to asperrubrol (**7**, top) and asperrubrol with $^{13}C_7$ incorporated (bottom).

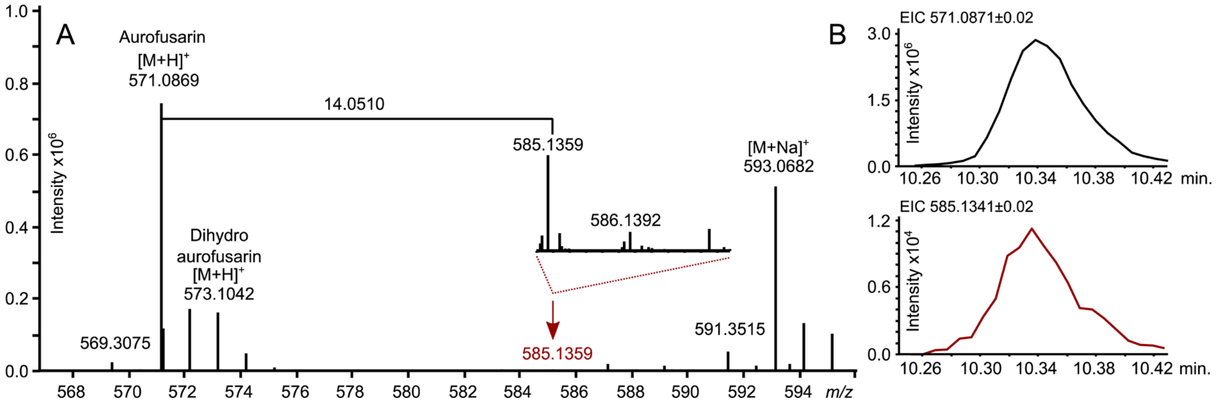


Figure 2. (A) Mass spectrum extracted at RT 10.3 min showing $[M + H]^+$ (m/z 571.0869, mass deviation −0.35 ppm) and $[M + Na]^+$ (m/z 593.0682) pseudomolecular ions. A mass shift of 14.0510 Da (m/z 585.1359 mass deviation, 3.1 ppm) suggests incorporation of $^{13}C_{14}$ (red arrow). (B) EICs corresponding to aurofusarin (**12**, top) and aurofusarin with $^{13}C_{14}$ incorporated (bottom).

Mass spectra of asperrubrol from samples fed with $^2H_7$-cinnamic acid exhibited no changes compared with the control samples. If cinnamic acid was converted into benzoic acid or another advanced precursor prior to incorporation into asperrubrol, extracted ion chromatograms corresponding to asperrubrol labeled with five, six, or seven ^2H atoms should be detectable; our experiments showed this was not the case.

Cultures of samples fed with $^{13}C_7$-benzoic acid also did not exhibit any changes in morphologies nor any peaks appearing or disappearing in the BPCs (Supporting Information, Figure S2), but investigation of the peak corresponding to asperrubrol revealed an ion with m/z 344.2031, corresponding to a difference of m/z 7.0233 compared with the $[M + H]^+$ ion of asperrubrol (Figure 1A). The ion corresponding to the $[M + Na]^+$ pseudomolecular ion of asperrubrol, as well as its labeled form, was also detected. This corresponded to incorporation of $^{13}C_7$ into the asperrubrol molecule. EICs of asperrubrol and its labeled form (Figure 1B) exhibited similar peak shapes and retention time (RT) and had a degree of incorporation of around 1.3% (Table 1).

These results suggest that asperrubrol is indeed biosynthesized from benzoic acid, which may in turn be synthesized from cinnamic acid in a different compartment. These results support the structure of asperrubrol reported by Rabache et al.[20]

Labeling in *Fusarium* spp. The compound YWA1 is known to be a biosynthetic precursor of several compounds including nor-rubrofusarin, rubrofusarin, fuscofusarin, and aurofusarin in fusaria. To investigate the biosynthesis of these, ^{13}C-labeled YWA1 (**8**) was used in labeling studies with two wild-type *Fusarium* strains, as well as two PKS12 deletion strains, deficient in the production of YWA1, grown under conditions that induce production of the compounds of interest. The two wild-type *Fusaria* did not exhibit any changes in morphologies or BPCs as a result of adding labeled substrate (data not shown). The mass spectrum extracted at the RT of the peak corresponding to aurofusarin showed ions corresponding to both unlabeled aurofusarin (**12**) and aurofusarin labeled with $^{13}C_{14}$ (Figure 2A).

EICs corresponding to labeled and unlabeled aurofusarin (Figure 2B) exhibited similar peak shapes and RTs with an incorporation degree of 0.4% (Table 1). No ions corresponding to aurofusarin with incorporation of two labeled YWA1 units were detected. This result was not surprising due to the low frequency of incorporation, that is, the frequency of incorporation of two units into aurofusarin would be $(0.4\%)^2 \approx 0.0016\%$, which is below the limit of detection.

Based on the previously established biosynthetic pathway of aurofusarin,[26,28,29] intermediates of the biosynthesis were investigated to determine if labeling of these could be detected.

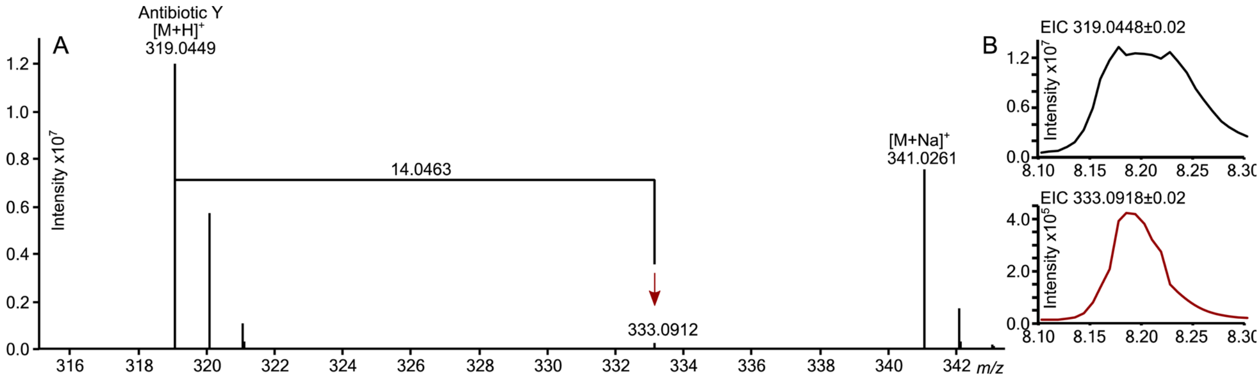


Figure 3. (A) Mass spectrum extracted at RT 8.2 min with $[M + H]^+$ (*m/z* 319.0449, mass deviation 0.18 ppm) and $[M + Na]^+$ (*m/z* 341.0261) pseudomolecular ions corresponding to antibiotic Y. Mass shift of $^{13}C_{14}$ suggest incorporation of labeled YWA1 (red arrow). (B) EICs corresponding to antibiotic Y (**13**; *m/z* 333.0912, mass deviation −1.8 ppm; top) and antibiotic Y with $^{13}C_{14}$ (bottom).

Only one precursor to aurofusarin, rubrofusarin (See Supporting Information, Figure S3), was detected in its labeled form and exhibited an incorporation degree of 20% (Table 1).

The two PKS12 deletion strains, *F. graminearum* Δ*PKS12* P1b and *F. graminearum* PH-1 HUEA (Δ*PKS12*), were also investigated by feeding with $^{13}C_{14}$-YWA1. These should not be able to produce YWA1 or aurofusarin. The PH-1 HUEA strain is thus pale white, while the wild-type *F. graminearum* is deep red. For one of these strains, PH-1 HUEA, addition of YWA1 resulted in visual changes: addition of $^{13}C_{14}$-YWA1 on day three resulted in bright red coloring around the reservoir, and addition after 7 days resulted in brownish coloring (Supporting Information, Figure S4). Addition of $^{13}C_{14}$-YWA1 after 10 days did not result in any color change. The colors of the control samples were unchanged throughout all 14 days. BPCs from the analysis did not reveal any changes in the chemical profiles (Supporting Information, Figure S5). Chemical analysis showed that the samples fed on days three and seven contained a compound with the same RT as aurofusarin. The mass spectrum (Supporting Information, Figure S6) contained an ion (*m/z* 599.1807) corresponding to aurofusarin with two YWA1 units ($^{13}C_{28}$) incorporated. Because this strain is not able to biosynthesize YWA1 on its own, all aurofusarin produced must be a product of the added $^{13}C_{14}$-YWA1, thus allowing detection of aurofusarin with two YWA1 units incorporated. This demonstrated that the fungus is indeed able to take up YWA1 from the medium and that YWA1, as expected, is a precursor to aurofusarin.

To test the hypothesis that antibiotic Y in *F. avanaceum* was also formed from YWA1, a wild-type *F. avenaceum* was fed with $^{13}C_{14}$-YWA1 under conditions that were known to induce production of antibiotic Y. As expected, feeding did not affect the metabolite profile (Supporting Information, Figure S7). However, closer investigation of the mass spectrum from the peak corresponding to antibiotic Y (Figure 3) revealed an ion (*m/z* 333.0912) corresponding to antibiotic Y with $^{13}C_{14}$ incorporated.

EICs corresponding to unlabeled antibiotic Y and antibiotic Y with $^{13}C_{14}$ incorporated (Figure 3B) exhibited similar RT, confirming that the labeled YWA1 precursor is incorporated into antibiotic Y. The unlabeled form was present in high enough amounts to saturate the detector, which accounts for the differences observed for the peak shapes. To calculate the degree of incorporation, the intensity of the $[^{13}C_1M + H]^+$ ion, which was not saturated, was then used to estimate the nonsaturated intensity of $[M + H]^+$, calculated using the theoretical ratio between these two. This showed that the degree of incorporation of YWA1 into antibiotic Y was 0.4% (Table 1). These results confirmed the hypothesis that YWA1 is a precursor to antibiotic Y and that its biosynthesis must depend on a yet undescribed structural rearrangement. To further investigate the biosynthesis of antibiotic Y, several putative intermediates were proposed and their chemical formulas formed the basis for a targeted analysis. One of these putative intermediates to antibiotic Y exhibited a mass spectrum indicative of YWA1 incorporation (Supporting Information, Figure S8), with an incorporation degree of 2.3%.

Comparison of the aurofusarin gene clusters in the genome-sequenced aurofusarin-producing fusaria revealed that the three antibiotic Y producing *F. avenaceum* strains contained an additional gene (*aurE*, FAVG1_08663) located centrally in the gene cluster.[35] AurE is predicted to encode a soluble epoxide hydrolase (EC: 3.3.2.3) based on its enzymatic domains. It is possible that the product of this unique gene is responsible for cleavage of YWA1 (**8**), and molecular genetics studies have been initiated to test this hypothesis.

Degrees of Incorporation. Overall the feeding experiments showed that the degrees of incorporation of the labeled precursors obtained by direct addition to wild-type strains varied significantly from 0.3% to 76%, with two further cases of incorporation into a presumed detoxification product. As expected, strains deficient in production of the precursor showed 100% incorporation. The degree of incorporation seemed to correlate inversely with the quantity of end product biosynthesized, with the signal of (2*Z*,4*E*)-2-methyl-2,4-hexadienedioic acid being very low in the patulin producers that have a 100-fold higher production of the compound than the terreic acid producing strains. In other published labeling studies, the degrees of incorporation of precursor have also varied. In a study of the mycotoxin terretonin by McIntyre et al., incorporation of several different differentially labeled precursors was investigated.[36] They found incorporation degrees of 0.3−2.5% depending on the precursor and cultivation conditions used. A study by Yoshizaws et al. investigated the incorporation of acetate in the biosynthesis of dehydrocurvalarin and found that these were incorporated at approximately 2%.[37] Finally, Yue et al. reported a 6% incorporation of ethyl (2*R*,3*R*)-2-methyl-3-hydroxy pentanoate into tylactone for an investigation of macrolide biosynthesis.[38]

The results revealed several important parameters for successful labeling of a compound through the use of an advanced labeled precursor. The organism must be able to take up the labeled precursor and, if necessary, transport it to a specific biosynthetic compartment in the cell. Second, the labeled compound must be included in the biosynthesis of a compound to act as a precursor. Finally, the precursor must be recognized by the tailoring enzymes as a substrate, and it is dependent on tailoring enzymes that are not physically coupled to the PKS synthesis, for example, as a protein complex. One hypothesis could be that synthesis of the PKs takes place in a so-called metabolon, where the SIL precursor cannot be inserted, as described for the tricarboxylic acid cycle.[39]

Examination of the data showed that the highest degree of incorporation of the labeled precursors was obtained at different time points, which is not surprising because biosynthesis also occurs at different time points during growth. For antibiotic Y, the highest degree of incorporation was obtained by addition after 7 days, but for aurofusarin, the highest incorporation was obtained with addition on day three. Presumably, the best strategy is to add the labeled compound at the onset of biosynthesis for the compound(s) to be studied. Another complication is that produced compounds may be recycled as part of the primary metabolism, as described for the nonribosomal peptide roquefortine C.[40]

Due to the low incorporation degrees observed for wild-type strains, a targeted analysis approach was required for determination of the incorporation levels. This could be combined with more systematic feeding studies, where fungi of interest could be cultivated using a whole panel of SIL precursors to investigate the biosynthesis of more complex compounds, since it is well suited for confirming hypotheses concerning biosynthetic pathways.

EXPERIMENTAL SECTION

General Experimental Procedures. All LC-MS analysis was performed using ultrahigh-performance liquid chromatography (UPHLC) UV/vis diode array detector (DAD) high-resolution MS (HRMS). The equipment used was an Agilent 6550 iFunnel Q-TOF LC/MS system (Torrance, CA) with an electrospray ionization (ESI) source operating in positive polarity, connected to an Agilent 1290 infinity UHPLC. The column used was an Agilent Poroshell 120 phenyl hexyl 2.7 μm, 250 mm × 2.1 mm column.

Chemicals. Solvents were LC-MS grade, and all other chemicals were analytical grade. All were from Sigma-Aldrich (Steinheim, Germany) unless otherwise stated. Water was purified using a Milli-Q system (Millipore, Bedford, MA). Electrospray ionization time-of flight (ESI-TOF) tune mix was purchased from Agilent.

$^{13}C_8$-Labeled 6-MSA (Table 2), 98.7%, had been produced by fermentation of a genetically modified *A. nidulans* by cultivation on labeled media, as described by Holm et al.[6] $^{13}C_7$-Benzoic acid, 99% labeled, and $^{2}H_7$-cinnamic acid, 98%, were purchased from Sigma-Aldrich (Steinheim, Germany).

Table 2. SIL Compounds Used in the Study

compound	elemental composition[a]	monoisotopic mass [Da]	mass difference[b] [Da]
6-MSA	$^{13}C_8H_8O_3$	152.0473	8.0268 (7.0235)[c]
cinnamic acid	$C_9{}^2H_7HO_2$	148.0524	7.0439
benzoic acid	$^{13}C_7H_6O_2$	122.0368	7.0235 (6.0201)[c]
YWA1	$^{13}C_{14}H_{12}O_6$	276.0634	14.0450

[a]Elemental composition denotes the formula of the compound and indicates the presence of labeled atoms. [b]Mass difference denotes the mass difference between the SIL compound and the natural predominant isotype. [c]Mass difference of compound following potential decarboxylation.

Construction of YWA1 Producing Strain. Protoplasting and gene targeting procedures were performed as described previously for *A. nidulans.*[41,42] The *wA* ORF (AN8209) was amplified with primers wA-fw (5′-GAGCGAUATGGAGGACCCATACCGTGT-3′) and wA-rv (5′-TCTGCGAUTATTAGAACCAGAGGATTATTATTGTT-3′) and inserted into the expression vector pDH57 via USER cloning, as described by Holm et al.[6] The gene targeting substrate for insertion of the YWA1 synthase gene was excised from pDH57-wA by *Not*I digestion and transformed into IBT 29539, as previously described.[6] Transformants with *wA* integrated into IS1 were verified by diagnostic PCR as described by Hansen and co-workers.[43]

Production and Purification of $^{13}C_{14}$-Labeled YWA1. The constructed YWA1 producing strain was propagated on solid MM medium prepared as described by Cove[44] and supplemented with 4 mM arginine. Spores were harvested after 14 days incubation at 30 °C with 10 mL of saline (0.9% NaCl in water) with 0.01% Tween 80 and filtered through Miracloth (Merck Millipore, Billerica, MA, USA). The spores were washed twice with saline prior to application. The batch fermentation was initiated by inoculation of 5 × 10^9 spores/L. A 1 L bioreactor (Sartorius, Goettingen, Germany) with a working volume of 0.8 L equipped with two Rushton six-blade disc turbines was used. The pH electrode (Mettler, Greifensee, Switzerland) was calibrated according to manufacturer standard procedures. For batch cultivation, the following media composition was applied: 20 g/L D-glucose-$^{13}C_6$ (99 atom % ^{13}C, Sigma-Aldrich) or D-glucose, 7.5 g/L $(NH_4)_2SO_4$, 1.5 g/L KH_2PO_4, 1.0 g/L $MgSO_4{\cdot}7H_2O$, 1.0 g/L NaCl, 0.1 g/L $CaCl_2$, 0.1 mL of Antifoam 204 (Sigma-Aldrich), 1 mL/L trace element solution (0.4 g/L $CuSO_4{\cdot}5\ H_2O$, 0.04 g/L $Na_2B_2O_7{\cdot}10H_2O$, 0.8 g/L $FeSO_4{\cdot}7H_2O$, 0.8 g/L $MnSO_4{\cdot}H_2O$, 0.8 g/L $Na_2MoO_4{\cdot}2H_2O$, 8.0 g/L $ZnSO_4{\cdot}7H_2O$.

The bioreactor was sparged with sterile atmospheric air, and off-gas concentrations of oxygen and carbon dioxide were measured with a Prima Pro Process mass spectrometer (Thermo-Fischer Scientific, Waltham, MA, USA). Temperature was maintained at 30 °C, and pH was controlled by addition of 2 M NaOH and H_2SO_4. Start conditions were as follows: pH 3.0, stir rate 100 rpm, and air flow 0.1 volume of air per volume of liquid per minute (vvm). These conditions were changed linearly in 720 min to pH 5.0, stir rate 800 rpm, and air flow 1 vvm. The cultivation was ended at glucose depletion, as measured by glucose test strips (Macherey-Nagel, Düren, Germany), and the culture had entered stationary phase as monitored by off-gas CO_2 concentration. The entire volume of the reactor was harvested, and the biomass was removed by filtration through a Whatman No. 1 qualitative paper filter followed by centrifugation at 8000*g* for 20 min to remove fine sediments. The YWA was then recovered from the supernatant by repetitive liquid−liquid extraction using ethyl acetate with 0.5% formic acid. The organic extract was completely dried in vacuo resulting in a crude extract that was redissolved in 20 mL of ethyl acetate and dry loaded onto 3 g of Sepra ZT C_{18} (Phenomenex, Torrence, CA, USA) resin prior to packing into a 25 g SNAP column (Biotage, Uppsala, Sweden) with 22 g of pure resin in the base. The crude extract was fractionated on an Isolera flask purification system (Biotage) using a water−acetonitrile gradient starting at 15:85 going to 100% acetonitrile in 23 min at a flow rate of 25 mL min^{-1} and kept at that level for 4 min. Fractions were collected using UV detection at 210 and 254 nm, resulting in a total of 20 fractions, of which two were pooled and analyzed. The total yield of 0.6 g of $^{13}C_{14}$−YWA1 was estimated to be 90% pure by UHPLC-UV/vis-TOFMS analysis and have a labeling degree of 98.2% based on the $^{13}C_{13}{}^{12}C/^{13}C_{14}$ ratio.

UHPLC-DAD-Quadrupole Time-of-Flight (qTOF) MS. Analysis was performed using UPHLC-DAD-HRMS. The equipment used was an Agilent 6550 iFunnel Q-TOF LC/MS system (Agilent Technologies, Torrence, CA, USA), connected to an Agilent 1290 infinity UHPLC. The column used was an Agilent Poroshell 120 phenyl hexyl 2.7 μm, 250 mm × 2.1 mm, and the column was maintained at 60 °C. The UV was used to measure at 280 nm. A linear water−acetonitrile (LC-MS-grade) gradient was used (both solvents were buffered with 20 mM formic acid) starting from 10% (v/v) acetonitrile and increased to 100% in 15 min, maintaining this rate for 2.5 min before returning to starting conditions in 0.1 min and staying there for 2.4 min before the following run. A flow rate of 0.35 mL/min was used. MS was performed in both ESI$^+$ and ESI$^-$ in the mass range *m/z* 30−1700. Additional parameters and settings are published in Kildgaard et al.[45]

Cultivation of Fungi. Attempted labeling of patulin and terreic acid was carried out using the following fungi: *Penicillium griseofulvum* (IBT 18169), *P. paneum* (IBT 24722), *P. carneum* (IBT 26356), *Byssochlamys nivea* (CBS 546.75), *Aspergillus clavatus* (IBT 27903), *A. hortai* (IBT 26384 = NRRL 274, formerly identified as *A. terreus*), and *A. floccosus* (IBT 22556 = WB 4872 = NRRL 4872, formerly identified *A. terreus* var. *floccosus*). The IBT strains are available from the IBT culture collection at authors' address, NRRL strains from National Center for Agricultural Utilization Research (Peoria, IL, USA), and the CBS strain from Centraalbureau voor Schimmelcultures (Utrecht, Netherlands).

With a 5 mm plug drill, a reservoir was cut in the middle of a solid YES 9 cm media plate (Figure 4), prepared as described Frisvad and Samson.[46] Into this reservoir was added 65 μL of spore suspension, and the fungi were incubated for 7 days at 30 °C in darkness. On day three, 100 μg of ^{13}C-labeled 6-MSA dissolved in 100 μL of EtOH−H_2O (1:4) was added to the reservoir. Control samples without addition and with addition of 100 μL of EtOH−H_2O (1:4) were also prepared. On day seven, five plugs were excised from across the fungus using a 5 mm plug drill, and the plugs were extracted using acidic ethyl acetate−dichloromethane−methanol (3:2:1 vol/vol/vol) as described by Smedsgaard,[47] followed by analysis using LC-MS. All experiments were performed in duplicate.

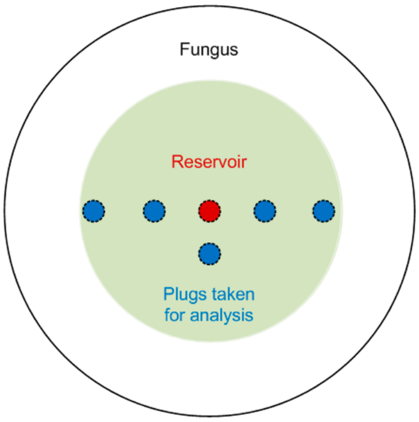


Figure 4. Diagram depicting the experimental setup. A reservoir (red) was cut in the middle of the media in the 9 cm Petri dish, and the fungus was then inoculated therein. At a specific time point, the labeled compound was added to the reservoir. At the end of the experiment plugs (blue) were removed from the fungal colony (green) and extracted as described in the text.

A. niger experiments, for the labeling of asperrubrol, were carried out following the described procedure, with addition of 100 μg of $^{13}C_7$-labeled benzoic acid or $^2H_7$-cinnamic acid dissolved in 100 μL of Milli-Q water on day 3 or 6, respectively. Separate control samples without labeled compounds were also fed to the strains. $^2H_7$-Cinnamic acid was only fed to *A. niger* KB1001. All experiments were prepared in duplicate. Sampling and extraction was performed as described above.

For the *Fusarium* labeling experiments four strains were used: *F. avanaceum* (IBT 41708), *F. graminearum* PH-1 (NRRL 31084) , *F. graminearum* ΔPKS12 P1b,[48] and *F. graminearum* PH-1 HUEA.[49] Fungi were inoculated on both Bells medium[50] and defined *Fusarium* medium (DFM)[51] and cultivated for 14 days at 30 °C in darkness to produce spores for the feeding experiment.

For the feeding experiments, solid Bells and DFM plates were prepared using a plug 5 mm drill to make a reservoir in the middle of the plate. Into this plate was added 65 μL of spore suspension, and the fungi were then cultivated for 14 days at 30 °C in darkness. After 3, 7, and 10 days, respectively, 100 μg of labeled YWA1, dissolved in 55 μL of ACN, was added to the reservoirs in the plates. Separate controls without labeled compounds and controls with 100 μL of ACN were also prepared. All experiments were prepared in duplicate. Sampling and extraction was performed as described above.

ASSOCIATED CONTENT

Supporting Information

Photographs of *F. graminearum* HUEA strain. BPCs from analysis of *A. niger, F. avanaceum*, and *F. graminearum* HUEA. Mass spectra of rubrofusarin, putative intermediate to antibiotic Y, and (2Z,4E)-2-methyl-2,4-hexadienedioic acid indicting labeling. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/np500979d.

AUTHOR INFORMATION

Corresponding Author

*Tel: +45 45 25 26 02. E-Mail: kfn@bio.dtu.dk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The study was supported by Grant 09-064967 from the Danish Council for Independent Research, Technology, and Production Sciences. We are grateful to Agilent technologies for the Thought Leader Donation of the Agilent UHPLC-qTOF system. We also thank Kenneth S. Bruno from Pacific Northwest Laboratory, WA, USA, for donating the *A. niger* ATCC 1015 strain, which is a derivative of ATCC 1015.

REFERENCES

(1) Adrio, J. L.; Demain, A. L. *Int. Microbiol.* **2003**, *6*, 191−199.
(2) Marroquín-Cardona, a G.; Johnson, N. M.; Phillips, T. D.; Hayes, a W. *Food Chem. Toxicol.* **2014**, *69*, 220−230.
(3) Bok, J. W.; Hoffmeister, D.; Maggio-Hall, L. A.; Murillo, R.; Glasner, J. D.; Keller, N. P. *Chem. Biol.* **2006**, *13*, 31−37.
(4) Villa, F. a; Gerwick, L. *Immunopharmacol. Immunotoxicol.* **2010**, *32*, 228−237.
(5) Hertweck, C. *Angew. Chem., Int. Ed.* **2009**, *48*, 4688−4716.
(6) Holm, D. K.; Petersen, L. M.; Klitgaard, A.; Knudsen, P. B.; Jarczynska, Z. D.; Nielsen, K. F.; Gotfredsen, C. H.; Larsen, T. O.; Mortensen, U. H. *Chem. Biol.* **2014**, *21*, 519−529.
(7) Kodicek, E. *Biochem. J.* **1955**, *60*, 25.
(8) Simpson, T. J. *Chem. Soc. Rev.* **1987**, *16*, 123.
(9) Townsend, C.; Christensen, S. *Tetrahedron* **1983**, *39*, 3575−3582.
(10) Steyn, P. S.; Vleggaar, R.; Simpson, T. J. *J. Chem. Soc., Chem. Commun.* **1984**, *3*, 765−767.
(11) Petersen, L. M.; Holm, D. K.; Knudsen, P. B.; Nielsen, K. F.; Gotfredsen, C. H.; Mortensen, U. H.; Larsen, T. O. *J. Antibiot. (Tokyo)* **2014**, 1−5.
(12) Christensen, B.; Nielsen, J. *Biotechnol. Prog.* **2002**, *18*, 163−166.
(13) Bode, H. B.; Reimer, D.; Fuchs, S. W.; Kirchner, F.; Dauth, C.; Kegler, C.; Lorenzen, W.; Brachmann, A. O.; Grün, P. *Chemistry* **2012**, *18*, 2342−2348.
(14) Bennett, B. D.; Yuan, J.; Kimball, E. H.; Rabinowitz, J. D. *Nat. Protoc.* **2008**, *3*, 1299−1311.
(15) Tanenbaum, S. W.; Bassett, E. W. *J. Biol. Chem.* **1959**, *234*, 1861−1866.
(16) Read, G.; Vining, L. *Chem. Commun.* **1968**, 935−937.
(17) Samson, R. A.; Peterson, S. W.; Frisvad, J. C.; Varga, J. *Stud. Mycol.* **2011**, *69*, 39−55.
(18) Boruta, T.; Bizukojc, M. *J. Biotechnol.* **2014**, *175*, 53−62.
(19) Guo, C.-J.; Sun, W.-W.; Bruno, K. S.; Wang, C. C. C. *Org. Lett.* **2014**, *16*, 5250−5253.
(20) Rabache, M.; Neumann, J.; Lavollay, J. *Phytochemistry* **1974**, *13*, 637−642.

(21) Holm, D. K. Development and implementation of novel genetic tools for investigation of fungal secondary metabolism, Ph.D. Thesis, Technical University of Denmark, 2013; p. 269.
(22) Jensen, K.; Evans, K.; Kirk, T. K.; Hammel, K. E. *Appl. Environ. Microbiol.* **1994**, *60*, 709−714.
(23) Watanabe, A.; Fujii, I.; Sankawa, U.; Mayorga, M. E.; Timberlake, W. E.; Ebizuka, Y. *Tetrahedron Lett.* **1999**, *40*, 91−94.
(24) Chiang, Y.-M.; Meyer, K. M.; Praseuth, M.; Baker, S. E.; Bruno, K. S.; Wang, C. C. C. *Fungal Genet. Biol.* **2011**, *48*, 430−437.
(25) Jørgensen, T. R.; Park, J.; Arentshorst, M.; van Welzen, A. M.; Lamers, G.; Vankuyk, P. a; Damveld, R. a; van den Hondel, C. a M.; Nielsen, K. F.; Frisvad, J. C.; Ram, A. F. J. *Fungal Genet. Biol.* **2011**, *48*, 544−553.
(26) Frandsen, R. J. N.; Nielsen, N. J.; Maolanon, N.; Sørensen, J. C.; Olsson, S.; Nielsen, J.; Giese, H. *Mol. Microbiol.* **2006**, *61*, 1069−1080.
(27) Sørensen, J. L.; Nielsen, K. F.; Sondergaard, T. E. *Fungal Genet. Biol.* **2012**, *49*, 613−618.
(28) Frandsen, R. J. N.; Schütt, C.; Lund, B. W.; Staerk, D.; Nielsen, J.; Olsson, S.; Giese, H. *J. Biol. Chem.* **2011**, *286*, 10419−10428.
(29) Rugbjerg, P.; Naesby, M.; Mortensen, U. H.; Frandsen, R. J. *Microb. Cell Fact.* **2013**, *12*, No. 31.
(30) Goliński, P.; Wnuk, S.; Chełkowski, J.; Visconti, A.; Schollenberger, M. *Appl. Environ. Microbiol.* **1986**, *51*, 743−745.
(31) Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. *Nat. Methods* **2011**, *8*, 785−786.
(32) Emanuelsson, O.; Nielsen, H.; Brunak, S.; von Heijne, G. *J. Mol. Biol.* **2000**, *300*, 1005−1016.
(33) Nakai, K.; Horton, P. *Trends Biochem. Sci.* **1999**, *24*, 34−35.
(34) Blum, T.; Briesemeister, S.; Kohlbacher, O. *BMC Bioinformatics* **2009**, *10*, No. 274.
(35) Lysøe, E.; Harris, L. J.; Walkowiak, S.; Subramaniam, R.; Divon, H. H.; Riiser, E. S.; Llorens, C.; Gabaldón, T.; Kistler, H. C.; Jonkers, W.; Kolseth, A.-K.; Nielsen, K. F.; Thrane, U.; Frandsen, R. J. N. *PLoS One* **2014**, *9*, No. e112703.
(36) McIntyre, C.; Scott, F.; Simpson, T.; Trimble, L.; Vederas, J. *Tetrahedron* **1989**, *45*, 2307−2321.
(37) Yoshizawa, Y.; Li, Z.; Reese, P. B.; Vederas, J. C. *J. Am. Chem. Soc.* **1990**, *112*, 3212−3213.
(38) Yue, S.; Duncan, J. S.; Yamamoto, Y.; Hutchinson, C. R. *J. Am. Chem. Soc.* **1987**, *109*, 1253−1255.
(39) Meyer, F. M.; Gerwig, J.; Hammer, E.; Herzberg, C.; Commichau, F. M.; Völker, U.; Stülke, J. *Metab. Eng.* **2011**, *13*, 18−27.
(40) Overy, D. P.; Nielsen, K. F.; Smedsgaard, J. *J. Chem. Ecol.* **2005**, *31*, 2373−2390.
(41) Johnstone, I. L.; Hughes, S. G.; Clutterbuck, A. J. *EMBO J.* **1985**, *4*, 1307−1311.
(42) Nielsen, M. L.; Albertsen, L.; Lettier, G.; Nielsen, J. B.; Mortensen, U. H. *Fungal Genet. Biol.* **2006**, *43*, 54−64.
(43) Hansen, B. G.; Salomonsen, B.; Nielsen, M. T.; Nielsen, J. B.; Hansen, N. B.; Nielsen, K. F.; Regueira, T. B.; Nielsen, J.; Patil, K. R.; Mortensen, U. H. *Appl. Environ. Microbiol.* **2011**, *77*, 3044−3051.
(44) Cove, D. J. *Biochim. Biophys. Acta, Enzymol. Biol. Oxid.* **1966**, *113*, 51−56.
(45) Kildgaard, S.; Mansson, M.; Dosen, I.; Klitgaard, A.; Frisvad, J. C.; Larsen, T. O.; Nielsen, K. F. *Mar. Drugs* **2014**, *12*, 3681−3705.
(46) Samson, R. A.; Houbraken, J.; Thrane, U.; Frisvad, J. C.; Andersen, B. *Food and Indoor Fungi*; Crous, P. W., Samson, R. A., Eds.; CBS-KNAW Fungal Biodiversity Centre: Utrecht, 2010.
(47) Smedsgaard, J. *J. Chromatogr. A* **1997**, *760*, 264−270.
(48) Malz, S.; Grell, M. N.; Thrane, C.; Maier, F. J.; Rosager, P.; Felk, A.; Albertsen, K. S.; Salomon, S.; Bohn, L.; Schäfer, W.; Giese, H. *Fungal Genet. Biol.* **2005**, *42*, 420−433.
(49) Sørensen, J. L.; Hansen, F. T.; Sondergaard, T. E.; Staerk, D.; Lee, T. V.; Wimmer, R.; Klitgaard, L. G.; Purup, S.; Giese, H.; Frandsen, R. J. N. *Environ. Microbiol.* **2012**, *14*, 1159−1170.
(50) Bell, A. a; Wheeler, M. H.; Liu, J.; Stipanovic, R. D.; Puckhaber, L. S.; Orta, H. *Pest Manage. Sci.* **2003**, *59*, 736−747.
(51) Yoder, W.; Christianson, L. *Fungal Genet. Biol.* **1998**, *80*, 68−80.

Combining UHPLC-high resolution MS and feeding of stable isotope labeled polyketide intermediates for linking precursors to end products

Andreas Klitgaard, Rasmus J. N. Frandsen, Dorte M. K. Holm, Peter B. Knudsen, Jens C. Frisvad, Kristian F. Nielsen*

Department of Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

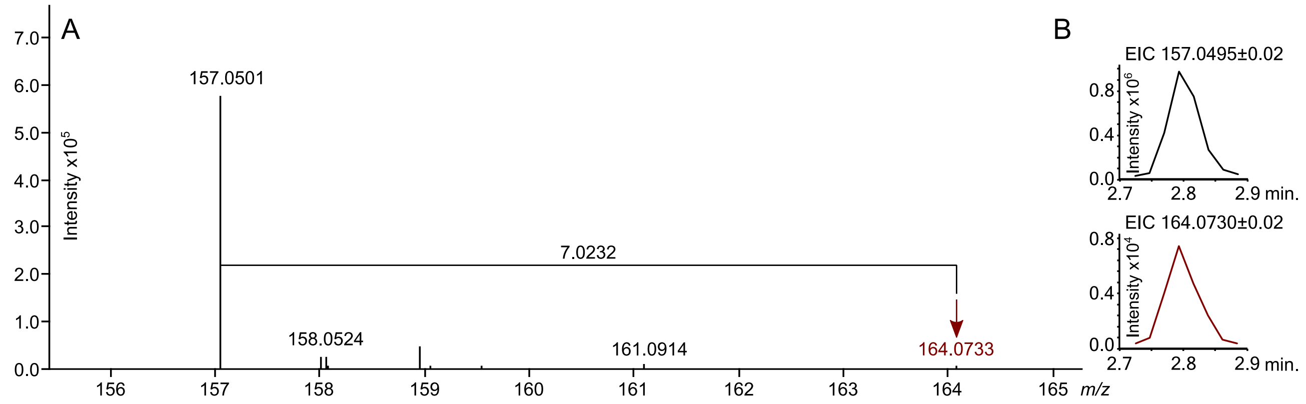


Figure S1 – A) Mass spectrum obtained from (2Z,4E)-2-methyl-2,4-hexadienedioic (**4**) at RT 2.8 min contained the $[M+H]^+$ (*m/z* 273.0761) pseudomolecular ion, as well as a an ion that displayed to a shift in mass indicative of incorporation of $^{13}C^7$-atoms. B) EICs corresponding to (2Z,4E)-2-methyl-2,4-hexadienedioic and (2Z,4E)-2-methyl-2,4-hexadienedioic with $^{13}C_7$-atoms incorporated are shown, and demonstrated the same peak shape and elution time.

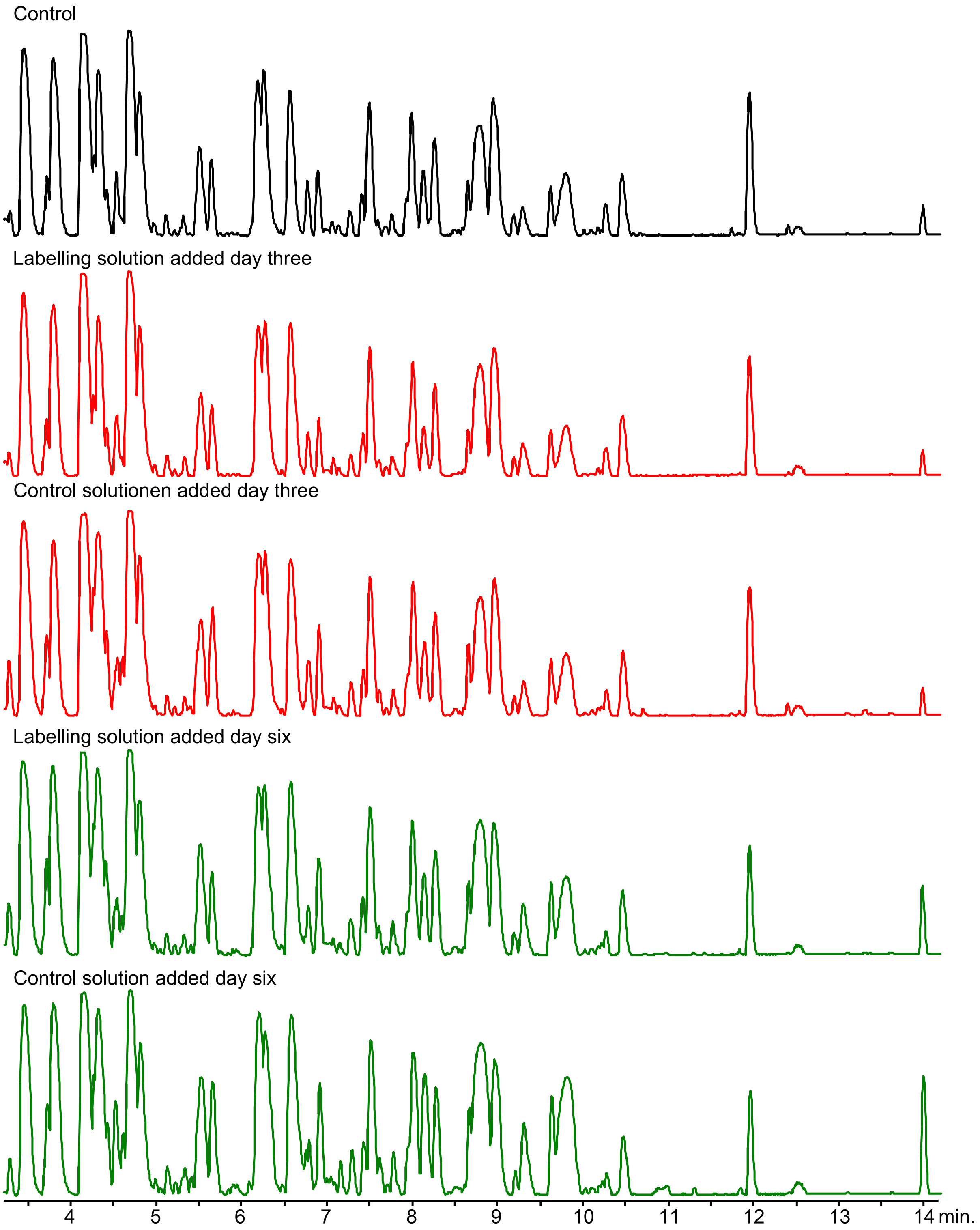


Figure S2 – BPCs from extracts of *A. niger* showed that no changes in the metabolite profiles were detected when the labeling solutions were added. The fungi were cultivated on YES for 7 days at 30 °C in darkness. The chromatograms have been scaled.

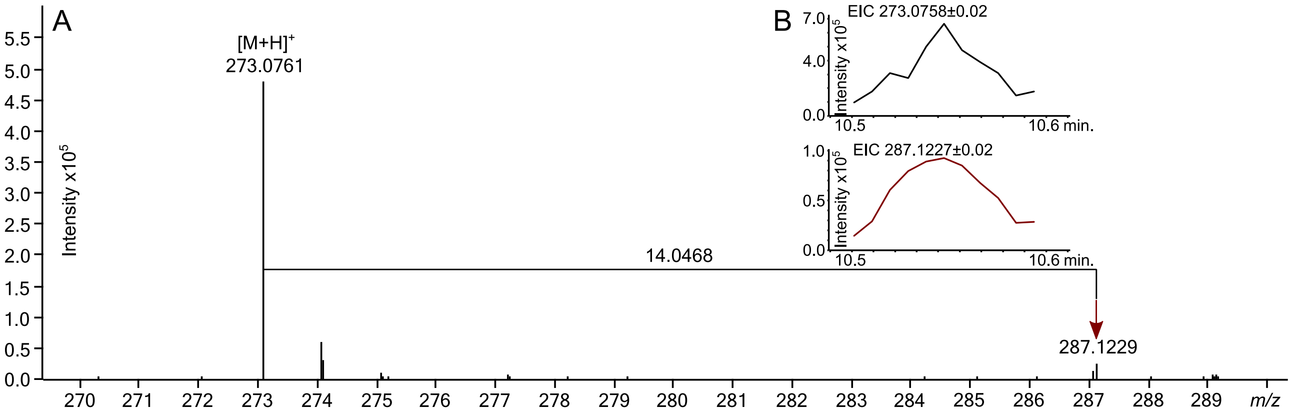


Figure S3 – A) Mass spectrum obtained from rubrofusarin at RT 10.5 min contained the $[M+H]^+$ (*m/z* 273.0761) pseudomolecular ion, as well as a an ion that displayed to a shift in mass indicative of incorporation of 14 ^{13}C-atoms. B) EICs corresponding to rubrofusarin (**10**) and rubrofusarin with 14 ^{13}C-atoms incorporated are shown, and demonstrated the same peak shape and elution time.

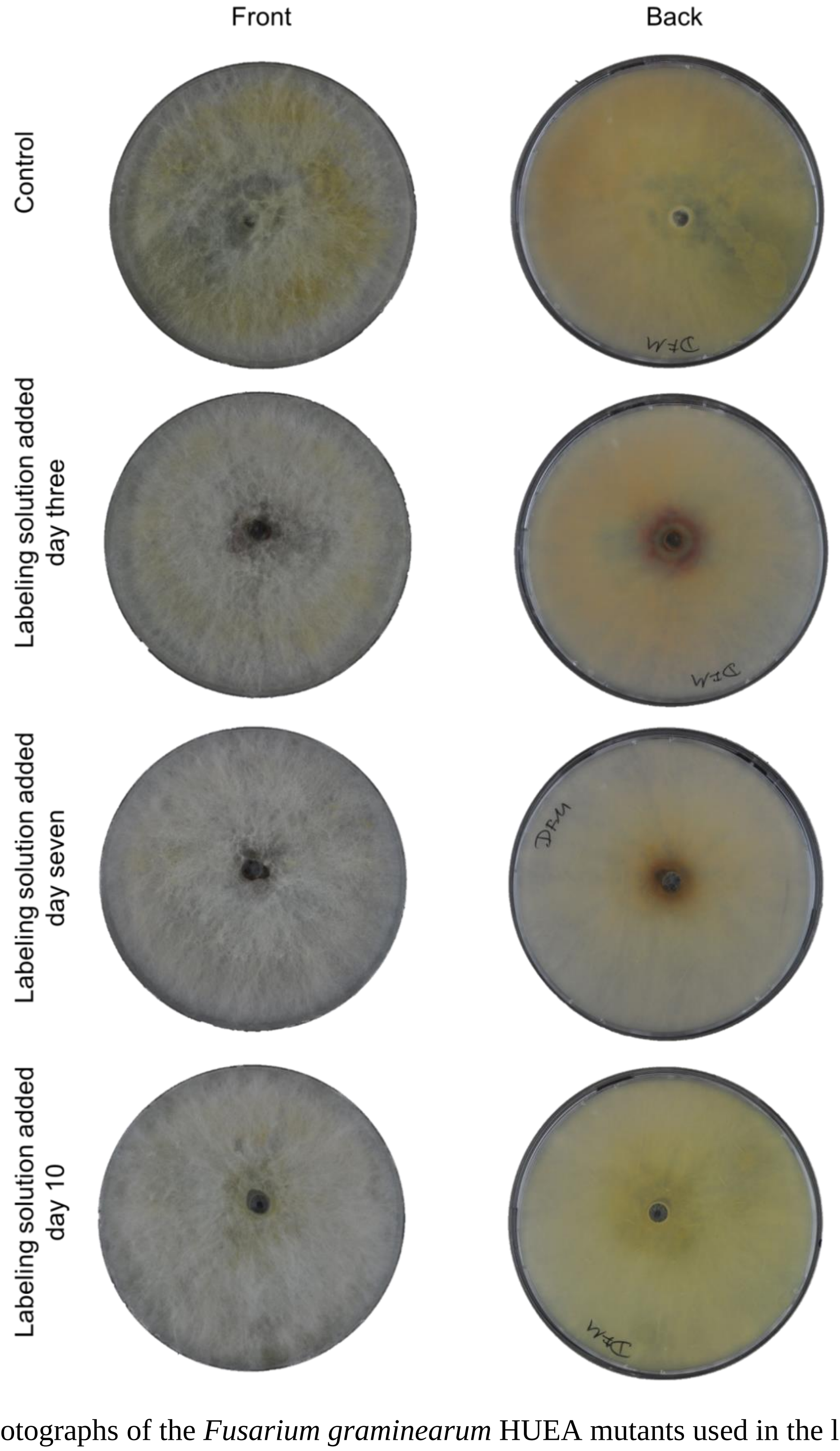

Front
Back
Control
Labeling solution added day three
Labeling solution added day seven
Labeling solution added day 10
DFM
DFM
DFM
DFM

Figure S4 – Photographs of the *Fusarium graminearum* HUEA mutants used in the labeling experiment cultivated on DFM medium at 30 °C for 14 days. Labeling solution was added after three, seven, or 10 days. The photographs show that addition of the labeling solution after three days resulted in a clear red color around the well where the solution was added. Addition after seven days resulted in a brownish coloring around the well, whilst addition after 10 days yielded no change.

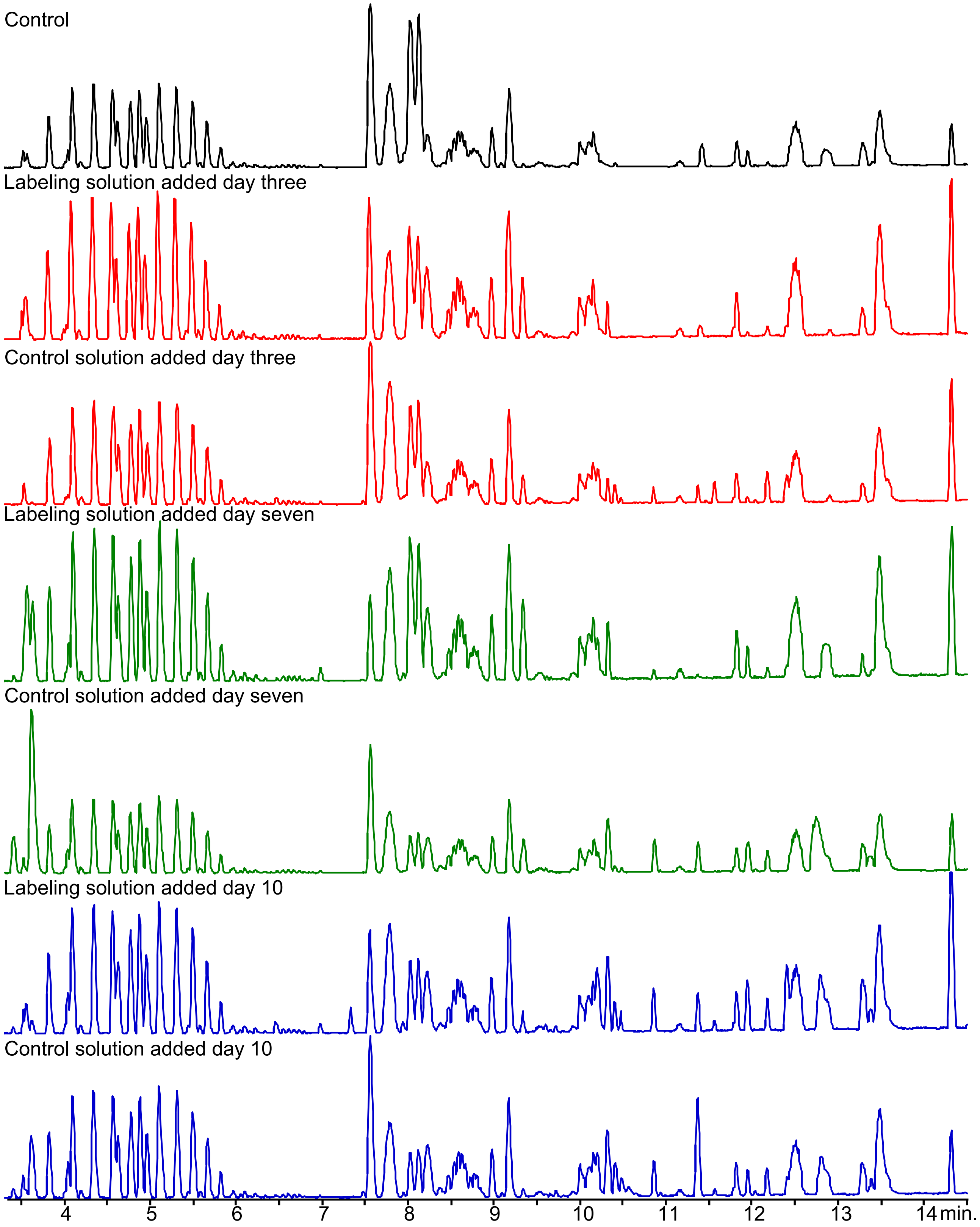


Figure S5 - BPCs from extracts of *F. gramineraum HUEA* showed that no changes in the metabolite profiles were detected when the labeling solutions were added. The fungi were cultivated on DFM for 14 days at 30 °C in darkness. The chromatograms have been scaled.

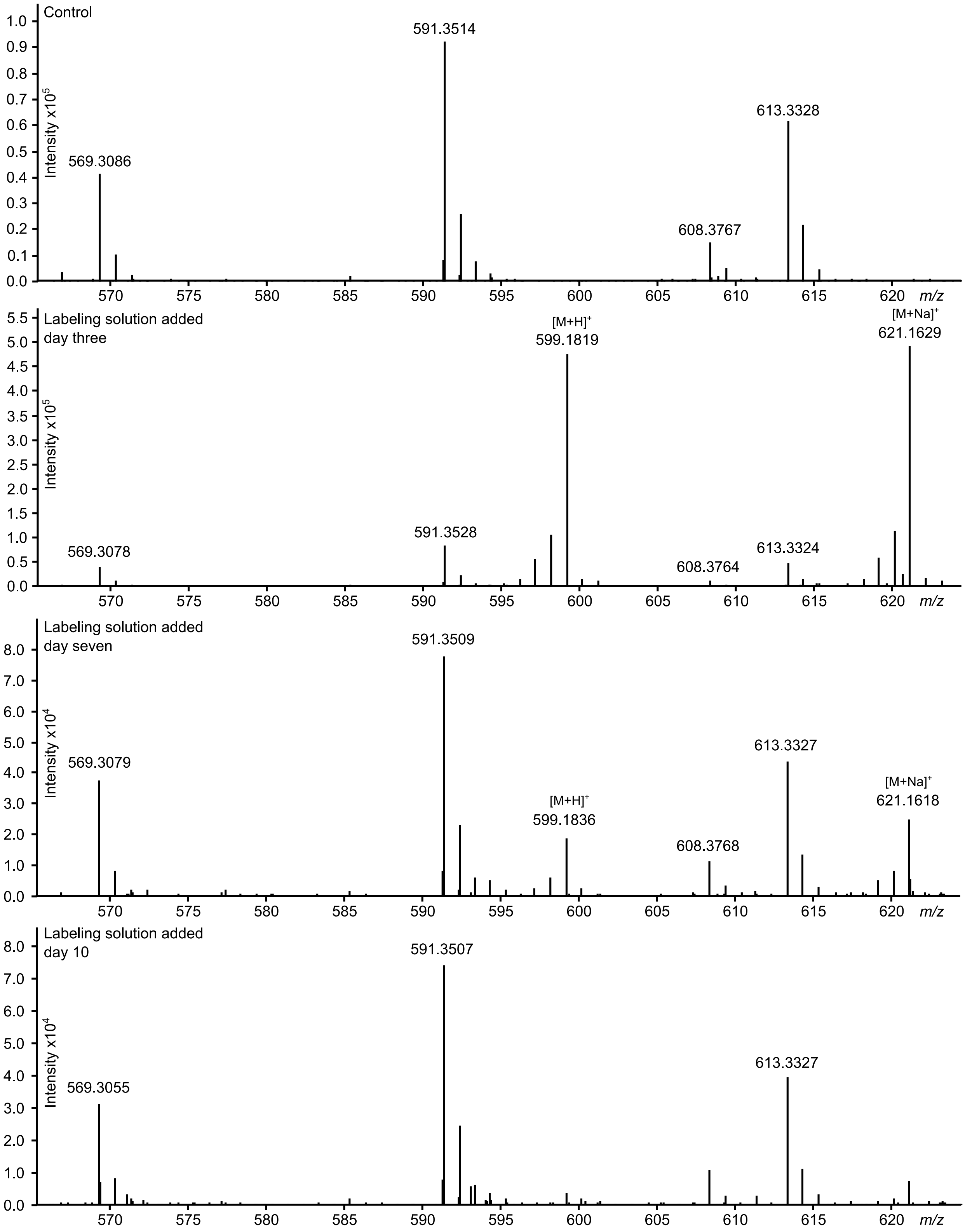


Figure S6 – Mass spectrum extracted at RT 10.3 min contained the $[M+H]^+$ (*m/z* 599.1819) and $[M+Na]^+$ (*m/z* 621.1629) pseudomolecular ions that corresponded aurofusarin with incorporation of two $^{13}C_{14}$-labeled YWA1 units, while showing now traces of the unlabeled form. The ions (*m/z* 569.3079), (*m/z* 591.3509), and (*m/z* 613.3327) were believed to be lipids unrelated to the investigated compounds.

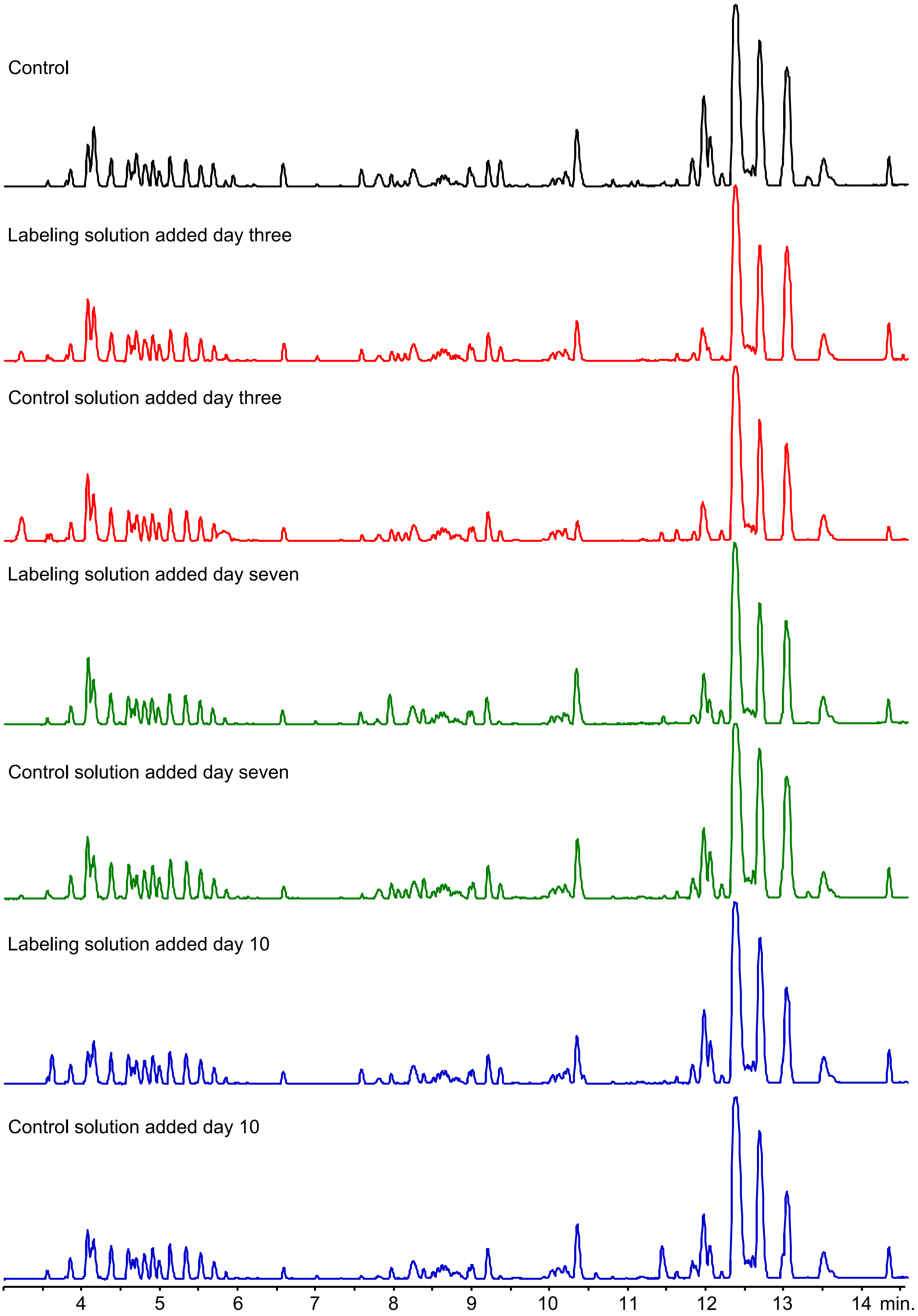


Figure S7 – BPCs from extracts of *F. avanaceum* showed that no changes in the metabolite profiles were detected when the labeling solutions were added. The fungi were cultivated on Bells medium for 14 days at 30 °C in darkness. The chromatograms have been scaled.

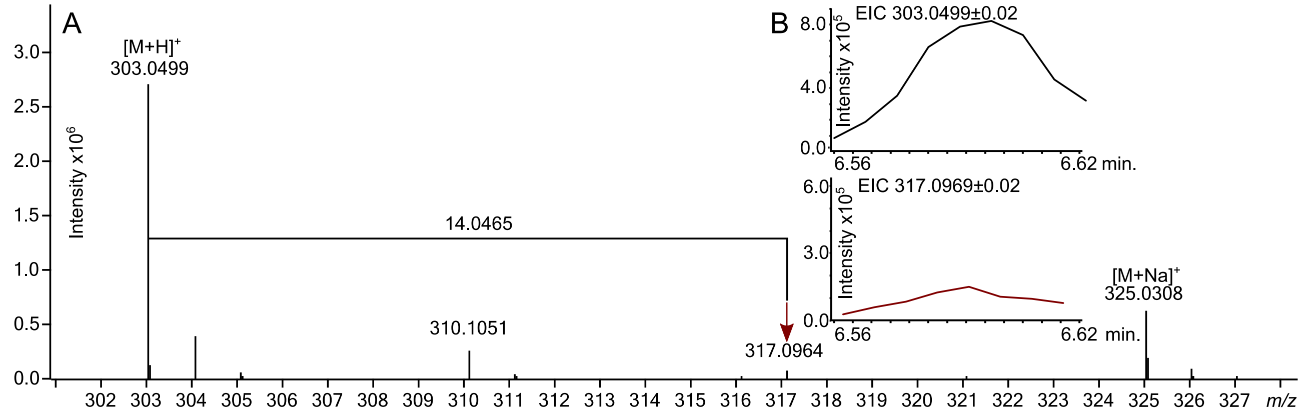


Figure S8 – A) Mass spectrum obtained from the putative intermediate to antibiotic Y (**14**) at RT 6.4 min contained the $[M+H]^+$ (*m/z* 291.0500) pseudomolecular ion, as well as a an ion that displayed to a shift in mass indicative of incorporation of 14 ^{13}C-atoms. B) EICs corresponding to the naturally occurring putative intermediate to antibiotic Y and the putative intermediate with 14 ^{13}C-atoms incorporated are shown, and demonstrated the same peak shape and elution time.

6.5 Paper 5 – Accurate prediction of secondary metabolite gene clusters in filamentous fungi

Andersen, M. R., Nielsen, J. B., **Klitgaard, A.**, Petersen, L. M., Zachariasen, M., Hansen, T. J., Blicher, L. H., Gotfredsen, C. H., Larsen, T. O., Nielsen, K. F., & Mortensen, U. H.

Paper accepted in Proceedings of the National Academy of Sciences of the United States of America (2014)

Accurate prediction of secondary metabolite gene clusters in filamentous fungi

Mikael R. Andersen[a,1], Jakob B. Nielsen[a], Andreas Klitgaard[a], Lene M. Petersen[a], Mia Zachariasen[a], Tilde J. Hansen[a], Lene H. Blicher[b], Charlotte H. Gotfredsen[c], Thomas O. Larsen[a], Kristian F. Nielsen[a], and Uffe H. Mortensen[a]

[a]Center for Microbial Biotechnology, Department of Systems Biology, [b]DTU Multi-Assay Core, Department of Systems Biology, and [c]Department of Chemistry, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved November 19, 2012 (received for review August 20, 2012)

Biosynthetic pathways of secondary metabolites from fungi are currently subject to an intense effort to elucidate the genetic basis for these compounds due to their large potential within pharmaceutics and synthetic biochemistry. The preferred method is methodical gene deletions to identify supporting enzymes for key synthases one cluster at a time. In this study, we design and apply a DNA expression array for *Aspergillus nidulans* in combination with legacy data to form a comprehensive gene expression compendium. We apply a guilt-by-association–based analysis to predict the extent of the biosynthetic clusters for the 58 synthases active in our set of experimental conditions. A comparison with legacy data shows the method to be accurate in 13 of 16 known clusters and nearly accurate for the remaining 3 clusters. Furthermore, we apply a data clustering approach, which identifies cross-chemistry between physically separate gene clusters (superclusters), and validate this both with legacy data and experimentally by prediction and verification of a supercluster consisting of the synthase AN1242 and the prenyltransferase AN11080, as well as identification of the product compound nidulanin A. We have used *A. nidulans* for our method development and validation due to the wealth of available biochemical data, but the method can be applied to any fungus with a sequenced and assembled genome, thus supporting further secondary metabolite pathway elucidation in the fungal kingdom.

aspergilli | natural products | secondary metabolism | polyketide synthases

No other group of biochemical compounds holds as much promise for drug development as the secondary (nongrowth associated) metabolites (SMs). A review from 2012 (1) found that for small-molecule pharmaceuticals, 68% of the anticancer agents and 52% of the antiinfective agents are natural products, or derived from natural products. The fact that SMs are often synthesized as polymer backbones that are subsequently diversified greatly via the actions of tailoring enzymes sets the stage for combinatorial biochemistry (2), because their biosynthesis is modular.

Major groups of SMs include polyketides (PKs) consisting of -CH_2-(C = O)- units, ribosomal and nonribosomomal peptides (NRPs), and terpenoids made from C_5 isoprene units. These polymer backbones are, with the exception of ribosomal peptides, made by synthases or synthetases and are modified by a plethora of tailoring enzymes, including (de)hydratases, oxygenases, hydrolases, methylases, and others.

In fungi, these biosynthetic genes of secondary metabolism are organized in discrete clusters around the synthase genes. Although quite accurate algorithms are available for identification of possible SM biosynthetic genes, particularly PK synthases (PKSs), NRP synthetases (NRPSs), and dimethylallyl tryptophan synthases (DMATSs) (3, 4), the assignment and prediction of the members of the individual clusters solely from the genome sequence have not been accurate. Relevant protein domains can be predicted for some of the genes (e.g., cytochrome P450 genes) (5); however, genes in identified clusters often have unknown functions, which makes predicting their inclusion impossible. Furthermore, SM gene clusters often colocalize on the chromosomes (6), which makes separation of clusters solely based on gene function predictions difficult.

The efficient elucidation of the biosynthetic genes for each SM cluster has thus so far been based on laborious single gene deletion of each of the putative members and chemical profiling of the SMs of the deletion strains. This effort has been especially noticeable in the model fungus *Aspergillus nidulans*, which is presently the fungal species with the largest number (n = 25) of characterized SM synthases/synthetases, due to a massive effort by several groups (7–30). In recent studies, this fungus has also been shown to have cross-chemistry between gene clusters on separate chromosomes (8, 30). Although these reactions are highly interesting for combinatorial chemistry, the identification of gene clusters involved in cross-chemistry is cumbersome because it involves combinatorial deletion of SM synthetic genes, thus greatly increasing the potential number of candidates.

In this study, we propose a general "omics"-based method for the accurate determination of fungal SM gene cluster members. The method is based on an annotated genome sequence and a catalog of gene expression, a set of information that is readily available for many fungal species and can easily be generated for more. To develop, benchmark, and validate this algorithm, we have used *A. nidulans* as a model organism, which is especially well-suited for this purpose due to the above-stated wealth of information. The algorithm is proven to be very powerful in identifying gene cluster members. We furthermore report an extension of the algorithm, which is proven to be successful in identifying cross-chemistry between gene clusters.

Results

Analysis of SMs *A. nidulans* on Complex Solid Medium Identifies 42 Compounds. Initially, we evaluated the production of SMs on four different solid media [oatmeal agar (OTA), yeast extract sucrose (YES), Czapek yeast autolysate (CYA), and CYA with 50 g/L NaCl sucrose (CYAS); *Materials and Methods*] at 4, 8, and 10 d. The object of this was to identify a selection of media that (*i*) gave as many produced SMs as possible, (*ii*) showed one or more SMs unique to each medium, and (*iii*) had SMs that were only produced on two of the selected media.

Author contributions: M.R.A., J.B.N., L.M.P., C.H.G., T.O.L., K.F.N., and U.H.M. designed research; M.R.A., J.B.N., A.K., L.M.P., M.Z., T.J.H., L.H.B., C.H.G., and K.F.N. performed research; M.R.A. and K.F.N. contributed new reagents/analytic tools; M.R.A., J.B.N., A.K., L.M.P., T.J.H., C.H.G., T.O.L., K.F.N., and U.H.M. analyzed data; and M.R.A., J.B.N., A.K., L.M.P., C.H.G., K.F.N., and U.H.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The gene expression data, gene expression microarray data description, and legacy gene expression data reported in this paper are available from the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE39993, GPL15899, GSE12859 and GSE7295).

[1]To whom correspondence should be addressed. E-mail: mr@bio.dtu.dk.

See Author Summary on page 24 (volume 110, number 1).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1205532110/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1205532110

These characteristics should allow us to have as many active gene clusters as possible, as well as ensuring unique production profiles for as many SM gene clusters as possible.

From this initial analysis, we selected the YES, CYA, and CYAS media for transcriptional profiling. On these media, we were able to separate and detect 59 unique SMs, of which we could name 42 by comparison with our extensive in-house library of microbial metabolites (31) and the AntiBase 2010 natural products database. The production profile of the compounds satisfied the three criteria listed above (Fig. 1, Fig. S1, and Dataset S1).

Generation of a Diverse Gene Expression Compendium for *A. nidulans*. Samples were taken for transcriptional profiling from plates cultivated in parallel to those of the SM profiling above. RNA was purified, prepared for labeling, and hybridized to custom-designed Agilent Technologies arrays based on version 5 of the *A. nidulans* annotation (32).

The produced data were combined with previously published microarray data from *A. nidulans* bioreactor cultivations (33, 34) to form a microarray compendium spanning a diverse set of conditions, comprising 44 samples in total. The set includes four strains of *A. nidulans*. Four different growth media are included: three complex media (see above) and one minimal medium. Medium variations include five different defined carbon sources (ethanol, glycerol, xylose, glucose, and sucrose), as well as yeast extract. The combined compendium of expression data is available in Dataset S2.

Correlation-Based Identification of Gene Clusters. To identify gene clusters efficiently around SM synthases, we developed a gene clustering score (CS) based on the Pearson product-moment correlation coefficient. Our CS gives a numerical value for correlation of the expression profile of a given gene with the expression profiles of the three immediate neighbor genes on either side. Only positive correlation is considered. Values for the CS are available in Dataset S2.

Statistical simulation of the distribution of CS on the given dataset showed that CS values ≥2.13 corresponded to a false-positive rate of 0.05 (Fig. S2). Therefore, CS ≥ 2.13 was used as a guideline for identifying the extent of gene clusters.

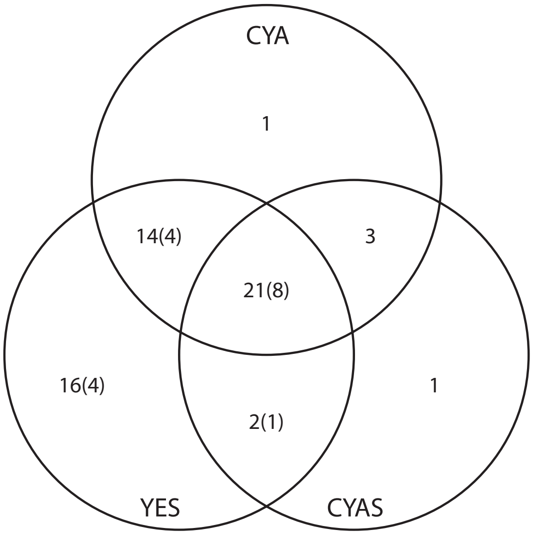


Fig. 1. Venn diagram of SMs found on three different solid media. The number of different metabolites is sorted according to which media the metabolites have been identified on. The number of metabolites unable to be confidently identified are noted in parentheses. Details can be found in Dataset S1, and the chemical structures are illustrated in Fig. S1.

Prediction of the Extent of 51 Gene Clusters. Evaluation of the size of the clusters around SM genes was performed using a precomputed list of 66 putative PKSs, NRPSs, and DMATSs from the secondary metabolite unique regions finder (SMURF) algorithm (3) based on the *A. nidulans* FGSC A4 gene set (35). In addition to these 66 genes, we added one prenyltransferase gene found in the primary literature (30) and three diterpene synthase (DTS) genes predicted by Bromann et al. (25), resulting in 70 putative biosynthetic genes. All 25 experimentally verified PKSs, NRPSs, DTSs, and prenyltransferases were found to be included in this list (Tables 1–3).

For each of the 70 biosynthetic genes, we examined the genes nearby for high CS values and inspected the expression profiles of the genes manually for additional validation and refinement. Apart from 12 genes that were silent under the conditions tested (Table S1), this allowed prediction of the sizes of gene clusters around 58 biosynthetic genes organized in 51 clusters and counting of a total of 254 genes included in the clusters (an example is shown in Fig. 2). The fact that we can map expression for 58 of the 70 biosynthetic genes (a large proportion of the gene clusters) is surprising, considering that many, or even the majority, of the gene clusters are reported to be silent under standard laboratory conditions (13, 14, 20, 36–38). An example of a cluster previously described as silent but identified here is the *inpAB* cluster (39). However, those cultivation experiments were conducted on liquid minimal medium and not on solid complex media, where we find that the expression from most of these genes is most pronounced. We therefore see the large number of active clusters as a confirmation of adequate diversity of the cultivation conditions in our microarray compendium.

Next, we investigated how our cluster predictions matched those published in the literature. This comparison demonstrated that our algorithm generally predicts gene clusters with excellent accuracy. Specifically, we accurately predict the extent of 11 of the 16 known gene clusters (Tables 1–3). In two of the remaining 5 gene clusters, the difference is due to artifacts. For the gene sterigmatocystin cluster (Fig. 2), the difference of 24 genes relative to 25 genes is caused by differences in the current gene annotation compared with the original paper from 1996 (17). Changes in gene calling are also the reason for discrepancy in the terrequinone cluster, where our legacy microarray data only contain data for 3 of the 5 genes, thus impairing the prediction. For the three remaining cases, the 2 gene clusters involved in meroterpenoid (austinol and dehydroaustinol) biosynthesis and the aspyridone cluster, the divergence seems to be biological. For the austinol/dehydroaustinol double-cluster system, we predict 3 extra genes in one cluster (around AN8383) and 2 extra genes in the other cluster (around AN9259) in addition to genes identified by Lo et al. (30). We individually deleted the 3 extra genes (AN8375, AN8376, and AN8380) in the AN8383 cluster; however, apart from differences in the austinol/dehydroaustinol ratio, we could only confirm the results of Lo et al. (30) of these genes not being essential for austinol/dehydroaustinol biosynthesis (Fig. S3). Because the size of most of the clusters was accurately predicted by our algorithm, we speculate that some or all of the extra genes are involved in biosynthesis of derivatives of austinol/dehydroaustinol. In agreement with this scenario, it is not uncommon that newly detected compounds are linked to known PKS pathways. For example, shamixanthones and arugusins were recently discovered to be products derived from the monodictyphenone cluster (8, 11), and this cluster has been redefined several times (9, 10). For the remaining case, the *apdG* gene of the aspyridone cluster (20), misprediction of the cluster members is due to a complete divergence between the transcription profiles of *apdG* and the remainder of the gene cluster. In general, we conclude that the use of

Table 1. Prediction of PKS gene clusters

GeneID	Gene	Compound (if known)	Cluster size		Medium	Ref(s).
			Predicted	Known		
AN0150	*mdpG*	Monodictyphenone/emodin	12	12	Solid	(7–10)
AN7903		Violaceol I and II	12	?	Solid	(11)
AN6448	*pkbA*		8	?	Solid	(24)
AN7084			8		Solid	
AN8209	*wA*	Green conidial pigment	6	?	Solid	
AN7909	*orsA*	Orsellinic acid/F9975/violaceols	5	5	Solid	(11–14)
AN1784			4		Solid	
AN9005			4		Solid	
AN6000	*aptA*	Asperthecin	3	3	Solid	(15)
AN6431			3		Solid	
AN11191			2		Solid	
AN7489			1		Solid	
AN3273			1		Solid	
AN2547	*easB*	Emericellamide	4	4	Both	(16)
AN3230	*pkfA*	Orsellinaldehydes	6	?	Both	(24)
AN7071	*pkgA*	Alternariol/isocoumarins	7	?	Both	(24)
AN7825	*stcA*	Sterigmatocystin	24*	25	Both	(17–19)
AN7815	*stcJ*	Sterigmatocystin	24*	25	Both	(17–19)
AN8383	*ausA*	Austinol	7	4	Both	(11, 24, 30)
AN2032	*pkhA*	Unknown	10	?	Liquid	(24)
AN2035	*pkhB*	Unknown	10	?	Liquid	(24)
AN8412	*adpA*	Aspyridone	7†	8	Liquid	(20)
AN6791			1		Liquid	
AN8910			1		Liquid	

This table contains predicted PKSs as well as PKS-like genes (AN7489 and AN7815) and a PKS/hybrid gene (AN8412). The medium column describes under which type of medium (liquid, solid, or both) the cluster is expressed. For gene clusters with identified functions and gene members, the number of identified cluster members is given as well as references to the original papers. Further details on the cluster members and the expression profiles of the individual clusters may be found in Dataset S2 and Fig. S4. Chemical structures of all compounds may be found in Fig. S1.

*Difference seemingly due to the current gene calling diverging from the original paper from 1996 (17).

†Algorithm was not able to predict the inclusion of *apdG*, the outmost gene hypothesized to be a part of the cluster (20). The expression profile of *apdG* diverges from the rest of the cluster.

CS values in combination with inspection of the expression profiles is a very effective tool to predict the extent of gene clusters, because the borders of 13 of 16 clusters were accurately predicted (when predictions were adjusted to compensate for the two artifacts discussed above) and there was near-accurate prediction of all 16 clusters.

Diverse Gene Expression Compendium Is Important for Accurate Prediction. To evaluate the compendium size needed for accurate predictions, we used principal component analysis (PCA) on our matrix of expression values (Dataset S2). Greater than 95% of the variation within the set can be described in the first three principal components. This suggests that a theoretical lower limit for this type of analysis would be three arrays if one could select conditions with a near-perfect difference in expression levels, ideally high, medium, and low expression for all genes, and with a maximum difference between all clusters and their surrounding genes. This would be nearly impossible to achieve for all clusters. However, if one is only interested in a single or a few gene clusters of interest, and has the appropriate prior knowledge, it should be possible to select three to five conditions and achieve accurate predictions. Very informative studies have been performed with two conditions, but the boundaries of the cluster can be difficult to determine (e.g., ref. 25).

To test how much it was possible to reduce our dataset, we used an unsupervised PCA-based analysis for incremental reduction of the dataset. In this, we found (unsurprisingly) that our biological replicate samples contain the smallest amount of unique information. Ten of 44 samples can be removed with only an approximately 10% loss in the data variation, and 25 of 44 samples (all replicates) can be removed with less than a 35% loss in data variation. The time sample series on a solid medium presented in this study were not reduced from the set until all biological replicates were reduced. We conclude that in selection of samples for cluster elucidation, one should sample as diversely as possible. Biological replicates are not cost-effective unless already available from prior studies.

Clustering of Synthase Expression Profiles Identifies Superclusters. Recent work has identified two cases of cross-chemistry between clusters located on separate chromosomes. The production of austinol and derived compounds (the meroterpenoid pathway) has been shown to be dependent on two separate clusters (11, 30), and the biosynthesis of prenyl xanthones is dependent on three separate clusters (8). We were interested in seeing whether this is a general phenomenon and whether such cross-chromosomal “superclusters” could be detected using our expression data.

A full gene-to-gene comparison of expression profiles between all predicted NRPSs, PKSs, DTSs, and prenyl transferases found in the array data was conducted, and the genes were clustered (Fig. 3). This clustering is not based directly on the expression profiles, because expression index variation from silent conditions distorts clustering. Instead, we clustered on the basis of a Spearman-based score of similarity to the expression profiles of the other synthases, which effectively eliminates noise.

The method is efficient for clustering the synthases and transferases according to shared products. Seven of eight sets of genes

Table 2. Prediction of NRPS gene clusters

			Cluster size			
GeneID	Gene	Compound (if known)	Predicted	Known	Medium	Source
AN9226			18		Solid	
AN6444			8		Solid	
AN4827			7		Solid	
AN8105			8		Solid	
AN8513	*tdiA*	Terrequinone A	3*	5	Solid	(21, 22)
AN1242	*nlsA*	Nidulanin A	3		Solid	This study
AN6961			2		Solid	
AN0016			1		Solid	
AN10486			1		Solid	
AN7884			14		Both	
AN3495	*inpA*	Unknown	7	7	Both	(25, 39)
AN3496	*inpB*	Unknown	7	7	Both	(25, 39)
AN2545	*easA*	Emericellamide	4	4	Both	(16)
AN2621	*acvA/pcbAB*	Penicillin G	3	3	Both	(25, 27, 28)
AN3396	*mica*	Microperfuranone	3	3[†]	Both	(29)
AN2924			2		Both	
AN10576	*ivoA*	*N*-acetyl-6-hydroxytryptophan	2	2	Both	(23, 26)
AN0607	*sidC*	Siderophores	1	1	Both	(55)
AN10297			1		Both	
AN5318			1		Both	
AN1680			1		Liquid	
AN2064			1		Liquid	
AN9129			1		Liquid	
AN9291			1		Liquid	

This table contains predicted NRPSs as well as NRPS-like genes (AN3396, AN5318, and AN9291). The medium column describes under which type of medium (liquid, solid, or both) the cluster is expressed. For gene clusters with identified functions and gene members, the number of identified cluster members is given as well as references to the original papers. Further details on the cluster members and the expression profiles of the individual clusters may be found in Dataset S2, and Fig. S4. Chemical structures of all compounds may be found in Fig. S1.

*Extent of the gene cluster is predicted correctly. The difference is due to the absence of two of the genes on the legacy microarray data, which removes them from the prediction.

[†] Yeh et al. (29), who examined this cluster, found increased transcription of the two extra genes we predict, but they found them to be nonessential for microperfuranone production.

predicted to be in the same biosynthetic clusters by the method above are found to cluster together in this representation. The exception is AN2032 and AN2035, which do not cocluster due to very low signals from the AN2032 probes on the microarray. Furthermore, the clustering is accurate in terms of cross-chemistry. In examining the two examples of cross-chemistry between gene clusters, it is found that these are predicted correctly. The meroterpenoid pathway includes the PKS AN8383 and the DMATS AN9259, which are illustrated to colocate in Fig. 3. The other example is the prenylxanthone biosynthetic pathway, which includes

Table 3. Prediction of gene clusters around prenyltransferases and diterpene synthases

				Cluster size			
GeneID	Type	Gene	Compound (if known)	Predicted	Known	Medium	Source
AN11194	DMATS			18		Solid	
AN11202	DMATS			18		Solid	
AN9259	DMATS			12	10	Both	(30)
AN8514	DMATS	*tdiB*	Terrequinone A	3*	5	Solid	(21, 22)
AN11080	DMATS	*nptA*	Nidulanin A	1		Both	This study
AN10289	DMATS			1		Solid	
AN6784	DMATS	*xptA*	Variecoxanthone A	1	1	Solid	(8–10)
AN1594	DTS		Ent-pimara-8(14),15-diene	9	9	Solid	(25)
AN3252	DTS			7		Solid	
AN9314	DTS			2		Solid	

This table contains predicted DMATSs, functionally prenyltransferases, and three DTSs predicted by Bromann et al. (25). The medium column describes on which type of medium (liquid, solid, or both) the cluster is expressed. For gene clusters with identified functions and gene members, the number of identified cluster members is given as well as references to the original papers. Further details on the cluster members and the expression profiles of the individual clusters may be found in Dataset S2, and Fig. S4. Chemical structures of all compounds may be found in Fig. S1.

*Extent of the gene cluster is predicted correctly. The difference is due to the absence of two of the genes on the legacy microarray data, which removes them from the prediction.

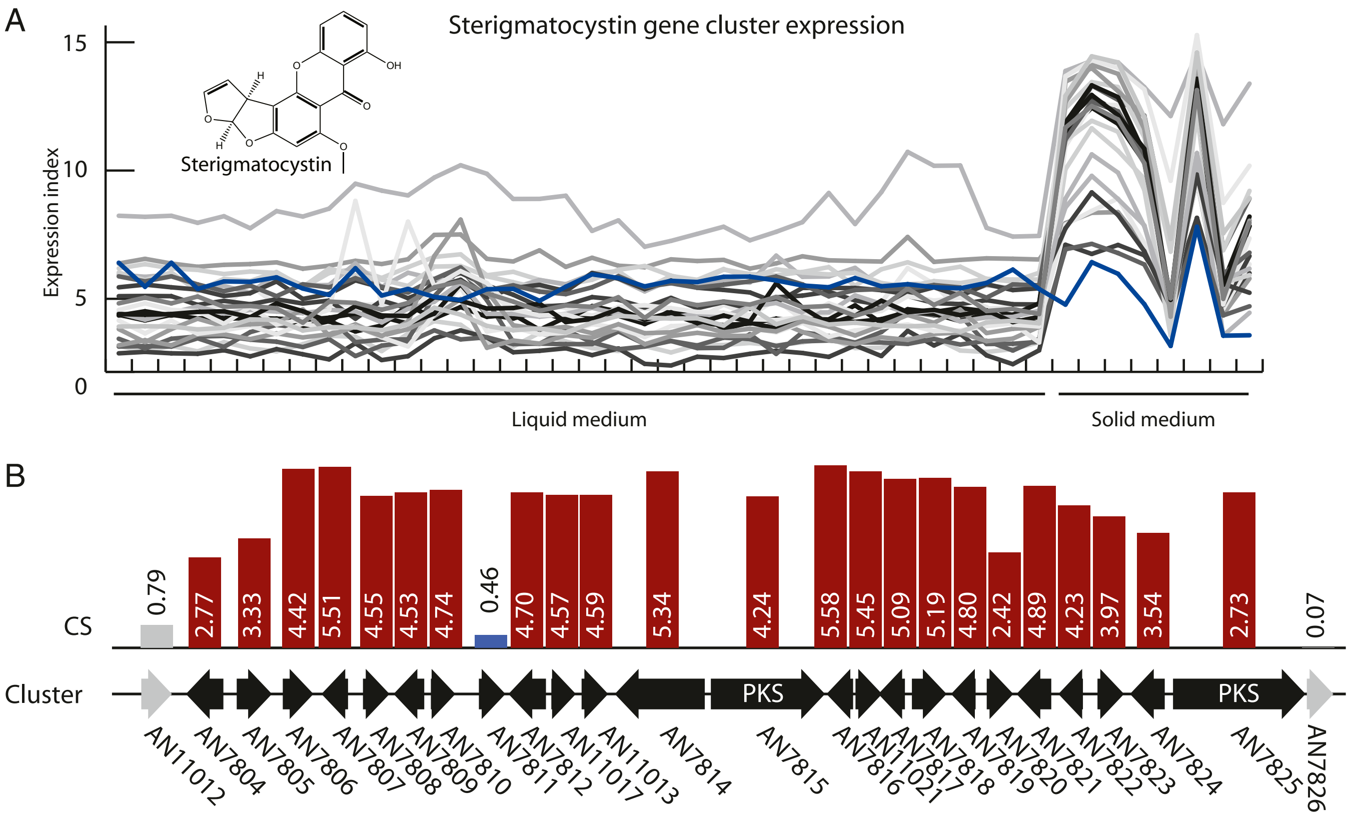


Fig. 2. Identification of the sterigmatocystin biosynthetic cluster. (*A*) Gene expression profiles across 44 experiments for the 24 genes (marked in black in *B*) predicted to be in the sterigmatocystin biosynthetic cluster (liquid and solid cultures are marked for reference). The expression profile of AN7811(*stcO*) is marked in blue. (*B*) Illustration of the values of the gene CS for the 24 genes and the two immediate neighbors. Genes included in the predicted cluster are marked in black. AN7811(*stcO*) did not have a CS above the used cutoff of 2.13 denoting clustering but was added due to the similarity of the expression profile, as shown in blue. The predicted extent of the cluster corresponds with the cluster as originally described by Brown et al. (17), when correcting for the fact that the gene models have changed since then. Full data for all predicted clusters may be found in Dataset S2.

the PKS AN0150 and the DMATS AN6784. These two genes are also found close to each other in Fig. 3.

We further use the maximum separation distance of two genes in the same biosynthetic cluster in the heat map of Fig. 3 as a cutoff distance for cross-chemistry. This allowed the genes to be sorted into seven larger superclusters. Details on the expression profiles of the individual clusters in each supercluster can be found in Fig. S4. Although we cannot directly separate tight coregulation from cross-chemistry with this method, the presence of these superclusters consisting of individual clusters with similar expression profiles suggests a larger extent of cross-chemistry in *A. nidulans* than what has been reported to date. To test the predictive power of this clustering further, we performed a gene deletion study within supercluster 5, which contains clusters located on six of the eight chromosomes.

Identification of the Chemical Structure of Nidulanin A Confirms Prediction of Cross-Chemistry Between NRPS AN1242 (NlsA) and Prenyltransferase AN11080 (NptA). To test the hypothesis of superclusters and whether the analysis above could be used to elucidate cross-chemistry, we constructed a deletion mutant of the NRPS AN1242 and evaluated the SMs found in the mutant relative to a reference strain. Four related compounds (compounds 1–4) were found to be absent in the ΔAN1242 strain (Fig. S5). MS isotope patterns as well as tandem MS (MS/MS) analysis showed compound 1 to have the molecular formula $C_{34}H_{45}N_5O_5$, with compounds 2 and 3 likely being oxygenated forms with one and two extra oxygen molecules, respectively. Compounds 1–3 all seem to be prenylated, as shown by spontaneous loss of a prenyl-like fragment, C_5H_8, in a small fraction of the ions during MS analysis. Compound 4 has a molecular formula of (1)-C5H8, suggesting it to be the unprenylated precursor of compound 1.

We thus isolated and elucidated the structure of compound 1, henceforth called nidulanin A, based on NMR spectroscopy. The stereochemistry of compound 1 was examined using Marfey's method (40) and was supported by bioinformatic analysis of the protein domains of AN1242 (*SI Text*). Altogether, nidulanin A is proposed to be a tetracyclopeptide with the sequence -L-Phe-L-Kyn-L-Val-D-Val- and an isoprene unit *N*-linked to the amino group of L-kynurenine (Fig. 4).

Because no prenyltransferase genes are found near AN1242, cross-chemistry catalyzed by an *N*-prenylating DMATS is a likely assumption. Examination of supercluster 5 in Fig. 3, where the NRPS AN1242 is found, shows AN11080 to be the DMATS with the expression profile most similar to AN1242. Gene deletion of AN11080 and subsequent ultra-high-performance liquid chromatography (UHPLC) high-resolution MS (HRMS) analysis of the ΔAN11080 strain show that the deprenylated compound 4, but none of the three prenylated forms, is present, thus confirming that nidulanin A and the two oxygenated forms (compounds 3 and 4) are synthesized by cross-chemistry between AN1242 (now NlsA) on chromosome VIII and AN11080 (now NptA) on chromosome V (Fig. S5).

Furthermore, we note that the masses corresponding to compound 3 (nidulanin A + O) and compound 4 (nidulanin A + O_2) are not found in the reference strain or in the ΔAN11080 strain. This suggests that compounds 3 and 4 are oxidized after the prenylation step.

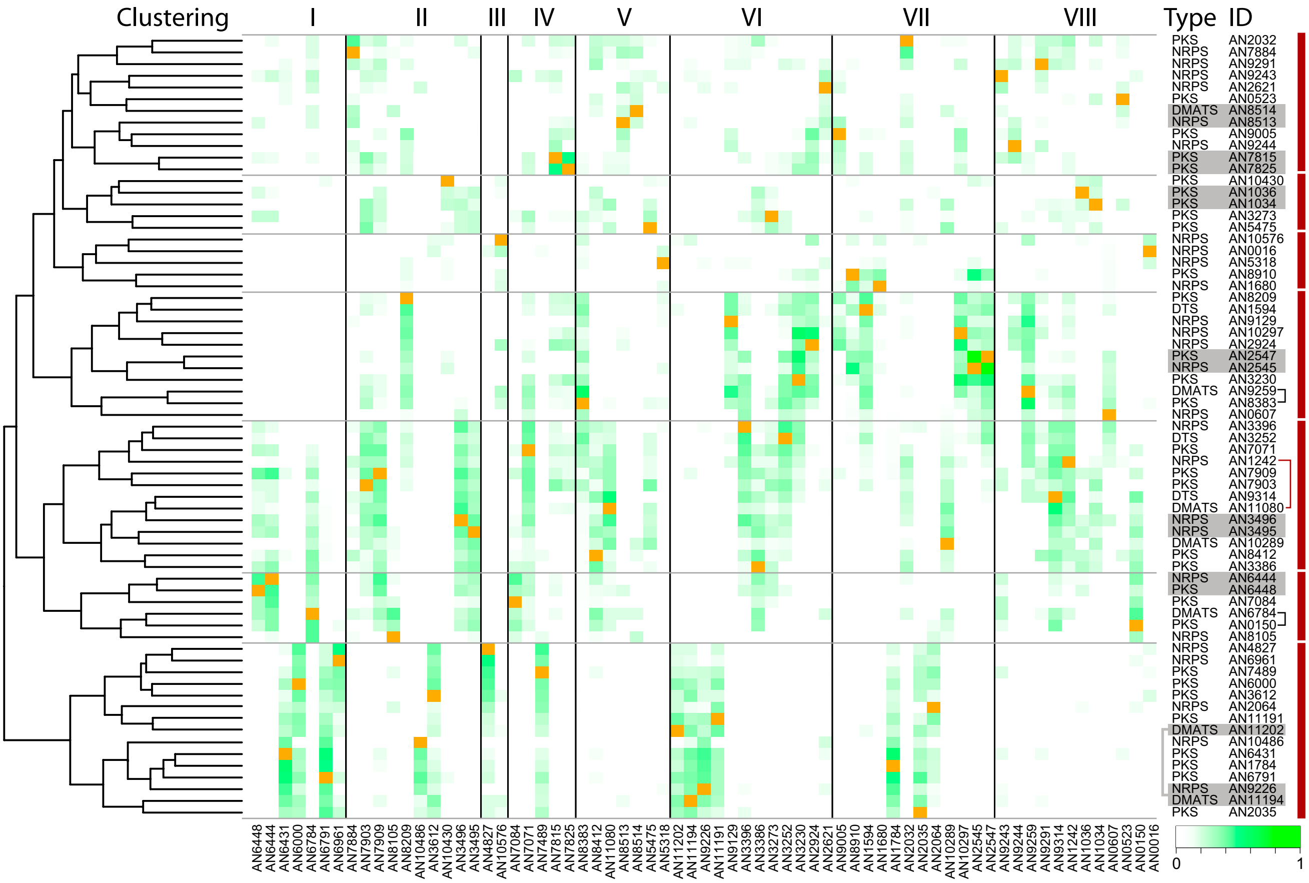
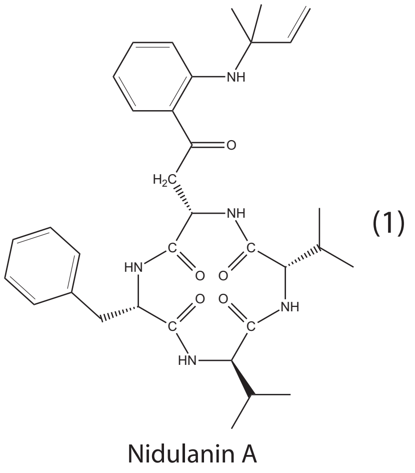


Fig. 3. Cross-chromosomal clustering. Matrix diagram of the correlation between 67 predicted and known biosynthetic genes. Each square in the matrix shows the compounded squared Spearman correlation coefficient for comparison of the expression profile of the genes color-coded from 0 (white) to 1 (green). Genes are sorted horizontally according to their location on the chromosomes (marked in orange) and vertically according to their scores (*Left*, marked with a dendrogram). (*Right*) Genes located in the same clusters are highlighted with a gray box, which is connected with a gray bracket in one case. Genes with known cross-chemistry are marked with a black bracket. An example of cross-chemistry found in this study is marked with a red bracket. Seven putative superclusters are marked. Further details of the clusters may be found in Fig. S4.

Discussion

In this study, we present a method for fungal SM cluster estimation based on similarity of expression profiles for neighboring genes. For the given organism *A. nidulans*, comparison with legacy data has verified the method to be highly accurate and effective for a large proportion of the gene clusters.



(1)

Fig. 4. Proposed absolute structure of nidulanin A. Details on the structural elucidation are available in *SI Text*.

It is clear from our results that the composition of the gene expression compendium has a significant effect on cluster predictions. We show here that it is important with a diverse set of samples, including both liquid and agar cultures as well as minimal medium and complex medium. This is in accordance with previous observations (11, 13, 14, 20, 36) stating that at a given set of conditions, only a fraction of the clusters are active. A reduction analysis of our own data has further shown that the inclusion of biological replicates in the dataset does not improve the analysis as much as inclusion of more unique samples. A diverse set of conditions should remedy regulation at the transcriptional level as well as chromatin-level regulation, which has been shown to have significant effects in fungi (13, 41). Another factor of importance is the quality of genome annotation. Erroneous gene calls inside clusters decrease the value of the CS for genes within a distance of three genes. Furthermore, problems with gene calls can affect expression profiling if a nontranscribed region is included in the gene cluster. However, neither of these seems to be a problem in the data presented here. Including the expression profiles of seven genes in the calculation of the CS also increases the robustness of the method toward erroneous gene calls.

The stated robustness of the CS has the disadvantage that the CS alone performs poorly for clusters with four or fewer genes, because the maximum value of CS for *n* genes is $n - 1$. However, in the cases of small clusters, the clustering can still be predicted from the transcription profiles, as shown in this study.

In some cases, we also see that cluster calling based on expression profiles outperforms the combination of gene KO and metabolomics. If a given detected metabolite is not the end product of the biosynthetic pathway, gene deletions will only identify a part of an SM cluster as being relevant for that metabolite, thus missing genes. An example of this is seen in the emodin/monodictyphenone cluster (PKS AN0150), where a subset of the genes is only required for some of the metabolites, resulting in a two-step elucidation of the gene cluster (7, 8). The CS method correctly calls the full cluster.

One aspect of the method is the ability to identify gene clusters simply from identifying groups of genes with high CS values, and not using a seeding set of synthases as was done in this case. This allows the unbiased identification of gene clusters throughout the entire genome. Although we see a surprising amount of these clusters (Dataset S2) not limited to the predicted SM synthases, we have not evaluated these in this study, because data for appropriate benchmarking is not available. However, we believe that there is great potential for biological discoveries to be made here, both in terms of promoter and chromatin-based transcriptional regulation.

The final extension of the algorithm is its ability to identify biosynthetic superclusters scattered across different chromosomes. Although this is a recently reported phenomenon (8), we believe that this is a common phenomenon, at least in *A. nidulans* and possibly in fungi in general. It is important to note that our method does not allow one to discriminate between tight coregulation and cross-chemistry between two distant clusters. It is therefore most efficient in cases in which it is evident that a given gene cluster does not hold all enzymatic activities required to synthesize the associated compound. In those cases, the use of a diverse transcription catalog, such as the one applied here, is a powerful strategy for identifying cross-chemistry, as shown for the NRPS AN1242 and the assisting prenyltransferase AN11080 in the synthesis of nidulanin A and derived compounds.

In summary, this study provides (*i*) an updated gene expression DNA array for *A. nidulans*, (*ii*) a wealth of information advancing the cluster elucidation in the model fungus *A. nidulans*, (*iii*) a powerful tool for prediction of SM cluster gene members in fungi, (*iv*) a proven methodology for prediction of SM gene cluster cross-chemistry, and (*v*) a proposed structure for the compound nidulanin A.

Materials and Methods

Strains. *A. nidulans* FGSC A4 was used for all transcriptomic experiments in this study. Furthermore, legacy data using the FGSC A4, *A. nidulans* AR16msaGP74 (expressing the *msaS* gene from *Penicillium griseofulvum*) (34), *A. nidulans* AR1phk6msaGP74 (expressing the *msaS* gene from *P. griseofulvum* and overexpressing the *A. nidulans xpkA*) (34), and *A. nidulans* AR1phkGP74 (overexpressing the *A. nidulans xpkA*) (33), were applied.

The *A. nidulans* FGSC A4 stock culture was maintained on CYA agar at 4 °C. *A. nidulans* strain IBT 29539 (*veA1*, *argB2*, *pyrG89*, and *nkuA*Δ) was used for all gene deletions. Gene deletion strains (see below) are available from the IBT fungal collection as *A. nidulans* IBT 32029, (AN1242Δ::*AfpyrG*, *veA1*, *argB2*, *pyrG89*, and *nkuA*Δ) and *A. nidulans* IBT 32030, (AN11080Δ::*AfpyrG*, *veA1*, *argB2*, *pyrG89*, and *nkuA*Δ). For chemical analyses, *A. nidulans* IBT 28738 (*veA1*, *argB2*, *pyrG89*, and *nkuA-trS*::*AfpyrG*) was used as reference strain.

Metabolite Profiling Analysis. *A. nidulans* strains were inoculated on CYA agar, OTA, YES agar, and CYAS agar (42). All strains were three-point inoculated on these media and incubated at 32 °C in darkness for 4, 8, or 10 d, after which three to five plugs (6-mm diameter) along the diameter of the fungal colony were cut out and extracted (43).

Samples were subsequently analyzed by UHPLC-UV/vis diode array detector (DAD)-HRMS on a maXis G3 quadrupole time-of-flight mass spectrometer (Bruker Daltonics) equipped with an electrospray injection (ESI) source. The mass spectrometer was connected to an Ultimate 3000 UHPLC system (Dionex). Separation of 1-μL samples was performed at 40 °C on a 100-mm × 2.1-mm inner diameter (ID), 2.6-μm Kinetex C_{18} column (Phenomenex) using a linear water-acetonitrile gradient (both buffered with 20 mM formic acid) at a flow rate of 0.4 mL/min starting from 10% (vol/vol) acetonitrile and increased to 100% acetonitrile in 10 min, keeping this for 3 min. HRMS was performed in ESI^+ with a data acquisition range of 10 scans per second at *m/z* 100–1,000. The mass spectrometer was calibrated using sodium formate automatically infused before each analytical run, providing a mass accuracy better than 1.5 ppm. Compounds were detected as their $[M + H]^+$ ion ± 0.002 Da, often with their $[M + NH_4]^+$ and/or $[M + Na]^+$ ion used as a qualifier ion with the same narrow mass range. SMs with a peak areas >10,000 counts (random noise peaks of approximately 300 counts) were integrated and identified by comparison with approximately 900 authentic standards available from previous studies (31, 44) and dereplicated against the approximately 18,000 fungal metabolites listed in AntiBase 2010 by ultraviolet-visible (UV/Vis) spectra, retention time, adduct pattern, and high-resolution data (<1.5 ppm mass accuracy and isotope fit better than 40 using SigmaFit; Bruker Daltonics) (31, 45).

Array Design. Initial probe design was done using OligoWiz 2.0 software (46) from the coding sequences of predicted genes from the genome sequence of *A. nidulans* FGSC A4 (35), using version 5 of the *A. nidulans* gene annotation, downloaded from the *Aspergillus* Genome Database (32).

For each gene, a maximum of three nonoverlapping, perfect-match 60-mer probes was calculated using the OligoWiz standard scoring of cross-hybridization, melting temperature, folding, position preference, and low complexity. A position preference for the probes was included in the computations. Pruning of the probe sequences was done by removing duplicate probe sequences.

Also included on the chip were 1,407 standard controls designed by Agilent Technologies. Details of the array are available from the National Center for Biotechnology Information Gene Expression Omnibus (accession no. GPL15899).

Microarray Gene Expression Profiling. ***Mycelium harvest and RNA purification.*** Whole colonies from three-stab agar plates were sampled for transcriptional analysis by scraping the mycelium off the agar with a scalpel and transferring the agar directly into a 50-mL Falcon tube containing approximately 15 mL of liquid nitrogen. Care was taken to transfer a minimum of agar to the Falcon tube. The liquid nitrogen was allowed to evaporate before capping the lid and recooling the tube in liquid nitrogen before storing the tube at −80 °C until use for RNA purification.

For RNA purification, 40–50 mg of frozen mycelium was placed in a 2-mL microcentrifuge tube precooled in liquid nitrogen containing three steel balls (two balls with a diameter of 2 mm and one ball with a diameter of 5 mm). The tubes were then shaken in a Retsch Mixer Mill at 5 °C for 10 min until the mycelium was ground to a powder. Total RNA was isolated from the powder using the Qiagen RNeasy Mini Kit according to the protocol for isolation of total RNA from plant and fungi, including the optional use of the QiaShredder column. Quality of the purified RNA was verified using a NanoDrop ND-1000 spectrophotometer and an Agilent 2100 Bioanalyzer (Agilent Technologies).

Microarray hybridization. A total of 150 ng in 1.5-μL total RNA was labeled according to the One Color Labeling for Expression Analysis, Quick Amp Low Input (QALI) manual, version 6.5, from Agilent Technologies. Yield and specific activity were determined on the ND-1000 spectrophotometer and verified on a Qubit 2.0 fluorometer (Invitrogen). A total of 1.65 μg of labeled cRNA was fragmented at 60 °C on a heating block, and the cRNA was prepared for hybridization according to the QALI protocol. A 100-μL sample was loaded on a 4 × 44 Agilent Gasket Slide situated in a hybridization chamber (both from Agilent Technologies). The 4 × 44 array was placed on top of the Gasket Slide. The array was hybridized at 65 °C for 17 h in an Agilent Technologies hybridization oven. The array was washed following the QALI protocol and scanned in a G2505C Agilent Technologies Micro Array Scanner.

Analysis of transcriptome data. The raw array signal was processed by first removing the background noise using the normexp method, and signals between arrays were made comparable using the quantiles normalization method as implemented in the Limma package (47). Multiple probe signals per gene were summarized into a gene-level expression index using Tukey's medianpolish, as performed in the last step of the robust multiarray average (RMA) processing method (48). The data are available from the Gene Expression Omnibus database (accession no. GSE39993).

The generated data from the Agilent Technologies arrays were combined with legacy Affymetrix data (accession nos. GSE12859 and GSE7295) using the qspline normalization method (46) to combine the two normalized sets of data to one microarray catalog with expression indices in comparable ranges.

Calculation of the Gene CS. The CS is calculated for each individual gene along the chromosomes according to the following equation:

$$CS_{\pm 3} = \sum_{i=-3}^{3}\left(\frac{s_{0,i} + \left\|s_{0,i}\right\|}{2}\right)^2 + \sum_{i=1}^{3}\left(\frac{s_{0,i} + \left\|s_{0,i}\right\|}{2}\right)^2, \qquad [1]$$

where $s_{0,i}$ is the Spearman coefficient for the expression indices of the gene in question and the gene located *i* genes away in a positive or negative direction relative to the chromosomal coordinate of the gene. The absolute term is added to set inverse correlations to 0. The CS assigned to a specific gene is the average of the CS for the liquid cultures and the CS for the solid cultures to adjust for background expression levels. Genes located less than four genes away from the ends of the supercontigs are assigned a CS of 0. All calculations were performed in the R software suite v. 2.14.0 (49), using the Bioconductor package (50, 51) for handling of array data. An adaptable R script for calculation of the CS is available on request.

Generation of Random Values for Evaluation of CS Significance. To estimate significance levels of the CS, a random set of scores was generated by selecting six genes at random as simulated neighbors for each of the 10,411 genes in the dataset. Examining this random distribution showed 95% of the population to have a CS <2.13 (Fig. S2). This value was used to have a false discovery rate of 0.05. All calculations were performed in R (49).

Identification of Gene Clusters. Gene clusters were defined around each NRPS, PKS, and DMATS by examination of the transcription profile of all surrounding genes with a CS ≥2.13 as well as three flanking genes in either direction. All genes with similar expression profiles were included in the cluster.

PCA-Based Analysis of Dataset Variation. PCA analysis was performed on the data of Dataset S2 using the prcomp-function of R (49). For stepwise reduction of the dataset, all principal components were calculated in each iteration and a sample was eliminated based on the one that had the largest contribution to the last principal component (i.e., with the smallest amount of unique information).

Generation of *A. nidulans* Gene Deletion Mutants. The genetic transformation experiments were performed with *A. nidulans* strain IBT 29539 [*veA1*, *argB2*, *pyrG89*, and *nkuA*Δ as described by Nielsen et al. (52)]. Fusion PCR-based bipartite gene targeting of substrates using the AF*pyrG* marker for selection and deletion of AN1242 was performed as described by Nielsen et al. (52), with the exception that all PCR assays were performed with the PfuX7 DNA polymerase (53). The deletion construct for AN11080 was assembled by uracil-specific excision reagent (USER) cloning. Specifically, sequences upstream and downstream of the gene to be deleted were amplified by PCR using primers containing a uracil residue (Table S2). The two PCR fragments were simultaneously inserted into the *Pac*I/Nt.*Bbv*CI USER cassette of pU20002A by USER cloning (54, 55). As a result, AF*pyrG* is now flanked by the two PCR fragments to complete the gene targeting substrate. The gene targeting substrate was released from the resulting vector pU20002A-AN11080 by digestion with *Swa*I. All restriction enzymes are from New England Biolabs. Primer sequences for deletion of the targeted genes and verification of strains are listed in Table S2. In addition, internal AF*pyrG* primers were used in combination with the check primers listed in Table S2 for confirmation of correct integration of DNA substrates (52). Transformants and AF*pyrG* pop-out recombinant strains were rigorously tested for correct insertions as well as for the presence of heterokaryons by touchdown spore-PCR analysis on conidia with an initial denaturation at 98 °C for 20 min.

MS/MS-Based Characterization of Compounds 1–4. Analysis was performed as stated above for the UHPLC-DAD-HRMS but in MS/MS mode, where analysis of the target mass and 6 *m/z* units up (to maintain isotopic pattern) was performed both via a targeted MS/MS list for the target compounds of interest and by the data-dependent MS/MS mode with an exclusion list, such that the same compound was selected several times. MS/MS fragmentation energy was varied from 18 to 55 eV.

Isolation and Structural Elucidation of Nidulanin A. Two hundred plates of minimal medium were inoculated with *A. nidulans*, from which SMs were extracted and nidulanin A was isolated in pure form. One-dimensional and 2D NMR spectra were recorded on a Bruker Daltonics Avance 800-MHz spectrometer with a 5-mm TCI Cryoprobe at the Danish Instrument Centre for NMR Spectroscopy of Biological Macromolecules at Carlsberg Laboratory. Stereoisometry of the amino acids was elucidated using Marfey's method (40). Details are provided in *SI Text*, Table S3, and Figs. S6–S8.

NRPS protein domains were predicted to identify adenylation domains and epimerase domains (56). Adenylation-domain specificities were predicted using NRPSpredictor (57). Details are provided in *SI Text*.

ACKNOWLEDGMENTS. We thank Peter Dmitrov, who treated the raw microarray data. We acknowledge Laurent Gautier for good scientific discussion of experimental design for microarray experiments, Marie-Louise Klejnstrup for assistance in retrieving MS data, and Dorte Koefoed Holm and Francesca Ambri for analysis of the austinol gene deletion mutants. We also thank the Danish Instrument Center for NMR Spectroscopy of Biological Macromolecules for NMR time. This work was supported by the Danish Research Agency for Technology and Production Grants 09-064967 and FI 2136-08-0023.

1. Newman DJ, Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 75(3):311–335.
2. Liu T, Chiang YM, Somoza AD, Oakley BR, Wang CC (2011) Engineering of an "unnatural" natural product by swapping polyketide synthase domains in Aspergillus nidulans. *J Am Chem Soc* 133(34):13314–13316.
3. Khaldi N, et al. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* 47(9):736–741.
4. Medema MH, et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web server issue):W339–W346.
5. Kelly DE, Krasevec N, Mullins J, Nelson DR (2009) The CYPome (Cytochrome P450 complement) of Aspergillus nidulans. *Fungal Genet Biol* 46(Suppl 1):S53–S61.
6. Palmer JM, Keller NP (2010) Secondary metabolism in fungi: Does chromosomal location matter? *Curr Opin Microbiol* 13(4):431–436.
7. Chiang YM, et al. (2010) Characterization of the Aspergillus nidulans monodictyphenone gene cluster. *Appl Environ Microbiol* 76(7):2067–2074.
8. Sanchez JF, et al. (2011) Genome-based deletion analysis reveals the prenyl xanthone biosynthesis pathway in Aspergillus nidulans. *J Am Chem Soc* 133(11):4010–4017.
9. Simpson TJ (2012) Genetic and biosynthetic studies of the fungal prenylated xanthone shamixanthone and related metabolites in Aspergillus spp. revisited. *ChemBioChem* 13(11):1680–1688.
10. Schätzle MA, Husain SM, Ferlaino S, Müller M (2012) Tautomers of anthrahydroquinones: Enzymatic reduction and implications for chrysophanol, monodictyphenone, and related xanthone biosyntheses. *J Am Chem Soc* 134(36):14742–14745.
11. Nielsen ML, et al. (2011) A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in Aspergillus nidulans. *FEMS Microbiol Lett* 321(2):157–166.
12. Sanchez JF, et al. (2010) Molecular genetic analysis of the orsellinic acid/F9775 gene cluster of Aspergillus nidulans. *Mol Biosyst* 6(3):587–593.
13. Bok JW, et al. (2009) Chromatin-level regulation of biosynthetic gene clusters. *Nat Chem Biol* 5(7):462–464.
14. Schroeckh V, et al. (2009) Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans. *Proc Natl Acad Sci USA* 106(34): 14558–14563.
15. Szewczyk E, et al. (2008) Identification and characterization of the asperthecin gene cluster of Aspergillus nidulans. *Appl Environ Microbiol* 74(24):7607–7612.
16. Chiang YM, et al. (2008) Molecular genetic mining of the Aspergillus secondary metabolome: Discovery of the emericellamide biosynthetic pathway. *Chem Biol* 15(6): 527–532.
17. Brown DW, et al. (1996) Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in Aspergillus nidulans. *Proc Natl Acad Sci USA* 93(4):1418–1422.
18. Kelkar HS, Keller NP, Adams TH (1996) Aspergillus nidulans stcP encodes an O-methyltransferase that is required for sterigmatocystin biosynthesis. *Appl Environ Microbiol* 62(11):4296–4298.
19. Keller NP, Watanabe CM, Kelkar HS, Adams TH, Townsend CA (2000) Requirement of monooxygenase-mediated steps for sterigmatocystin biosynthesis by Aspergillus nidulans. *Appl Environ Microbiol* 66(1):359–362.
20. Bergmann S, et al. (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. *Nat Chem Biol* 3(4):213–217.
21. Bouhired S, Weber M, Kempf-Sontag A, Keller NP, Hoffmeister D (2007) Accurate prediction of the Aspergillus nidulans terrequinone gene cluster boundaries using the transcriptional regulator LaeA. *Fungal Genet Biol* 44(11):1134–1145.
22. Schneider P, Weber M, Hoffmeister D (2008) The Aspergillus nidulans enzyme TdiB catalyzes prenyltransfer to the precursor of bioactive asterriquinones. *Fungal Genet Biol* 45(3):302–309.
23. Clutterbuck AJ (1969) A mutational analysis of conidial development in Aspergillus nidulans. *Genetics* 63(2):317–327.
24. Ahuja M, et al. (2012) Illuminating the diversity of aromatic polyketide synthases in Aspergillus nidulans. *J Am Chem Soc* 134(19):8212–8221.
25. Bromann K, et al. (2012) Identification and characterization of a novel diterpene gene cluster in Aspergillus nidulans. *PLoS ONE* 7(4):e35450.

26. Birse CE, Clutterbuck AJ (1990) N-acetyl-6-hydroxytryptophan oxidase, a developmentally controlled phenol oxidase from Aspergillus nidulans. *J Gen Microbiol* 136(9):1725–1730.
27. MacCabe AP, et al. (1991) Delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase from Aspergillus nidulans. Molecular characterization of the acvA gene encoding the first enzyme of the penicillin biosynthetic pathway. *J Biol Chem* 266(19):12646–12654.
28. Martin JF (1992) Clusters of genes for the biosynthesis of antibiotics: regulatory genes and overproduction of pharmaceuticals. *J Ind Microbiol* 9(2):73–90.
29. Yeh HH, et al. (2012) Molecular genetic analysis reveals that a nonribosomal peptide synthetase-like (NRPS-like) gene in Aspergillus nidulans is responsible for microperfuranone biosynthesis. *Appl Microbiol Biotechnol* 96(3):739–748.
30. Lo H-C, et al. (2012) Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in Aspergillus nidulans. *J Am Chem Soc* 134(10):4709–4720.
31. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO (2011) Dereplication of microbial natural products by LC-DAD-TOFMS. *J Nat Prod* 74(11):2338–2348.
32. Arnaud MB, et al. (2010) The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. *Nucleic Acids Res* 38(Database issue):D420–D427.
33. Panagiotou G, et al. (2008) Systems analysis unfolds the relationship between the phosphoketolase pathway and growth in Aspergillus nidulans. *PLoS ONE* 3(12):e3847.
34. Panagiotou G, et al. (2009) Studies of the production of fungal polyketides in Aspergillus nidulans by using systems biology tools. *Appl Environ Microbiol* 75(7):2212–2220.
35. Galagan JE, et al. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 438(7071):1105–1115.
36. Brakhage AA, et al. (2008) Activation of fungal silent gene clusters: A new avenue to drug discovery. *Prog Drug Res* 66(1):3–12.
37. Bok JW, et al. (2006) Genomic mining for Aspergillus natural products. *Chem Biol* 13 (1):31–37.
38. Cullen D (2007) The genome of an industrial workhorse. *Nat Biotechnol* 25(2):189–190.
39. Bergmann S, et al. (2010) Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. *Appl Environ Microbiol* 76(24):8143–8149.
40. Marfey P (1984) Determination of D- amino acids. II. Use of a bifunctional reagent, 1,5-di-fluoro-2,4-dinitrobenzene. *Carlsberg Res Commun* 49(6):591–596.
41. Nützmann HW, et al. (2011) Bacteria-induced natural product formation in the fungus Aspergillus nidulans requires Saga/Ada-mediated histone acetylation. *Proc Natl Acad Sci USA* 108(34):14282–14287.
42. Frisvad JC, Samson R (2004) Polyphasic taxonomy of Penicillium subgenus Penicillium. A guide to identification of the food and air-borne terverticillate Penicillia and their mycotoxins. *Stud Mycol* 49:1–173.
43. Smedsgaard J (1997) Micro-scale extraction procedure for standardized screening of fungal metabolite production in cultures. *J Chromatogr A* 760(2):264–270.
44. Nielsen KF, Smedsgaard J (2003) Fungal metabolite screening: Database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry methodology. *J Chromatogr A* 1002(1-2):111–136.
45. Månsson M, et al. (2010) Explorative solid-phase extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification. *J Nat Prod* 73(6):1126–1132.
46. Workman C, et al. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 3(9):research0048.
47. Smyth GK (2005) *Limma: Linear models for microarray data* (Springer, New York), pp 397–420.
48. Irizarry RA, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31(4):e15.
49. R Development Core Team (2007) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), Available at www.R-project.org.
50. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80.
51. Nielsen ML, Albertsen L, Lettier G, Nielsen JB, Mortensen UH (2006) Efficient PCR-based gene targeting with a recyclable marker for Aspergillus nidulans. *Fungal Genet Biol* 43(1):54–64.
52. Nielsen JB, Nielsen ML, Mortensen UH (2008) Transient disruption of non-homologous end-joining facilitates targeted genome manipulations in the filamentous fungus Aspergillus nidulans. *Fungal Genet Biol* 45(3):165–170.
53. Nørholm MH (2010) A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol* 10:21.
54. Hansen BG, et al. (2011) Versatile enzyme expression and characterization system for Aspergillus nidulans, with the Penicillium brevicompactum polyketide synthase gene from the mycophenolic acid gene cluster as a test case. *Appl Environ Microbiol* 77(9):3044–3051.
55. Eisendle M, Oberegger H, Zadra I, Haas H (2003) The siderophore system is essential for viability of Aspergillus nidulans: Functional analysis of two genes encoding l-ornithine N 5-monooxygenase (sidA) and a non-ribosomal peptide synthetase (sidC). *Mol Microbiol* 49(2):359–375.
56. Bachmann BO, Ravel J (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181–217.
57. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res* 33(18):5799–5808.

Supporting Information

Andersen et al. 10.1073/pnas.1205532110

SI Materials and Methods

Fungal Growth, Extraction, and Isolation of Nidulanin A. *Aspergillus nidulans* (IBT 22600) was inoculated as three-point stabs on 200 plates of MM and incubated in the dark at 30 °C for 7 d. The fungi were harvested and extracted twice overnight with EtOAc. The extract was filtered and concentrated in vacuo. The combined extract was dissolved in 100 mL of MeOH and H_2O (9:1), and 100 mL of heptane was added after the phases were separated. Eighty milliliters of H_2O was added to the MeOH/H_2O phase, and metabolites were then extracted with 5 × 100 mL of dichloromethane (DCM). The phases were then concentrated separately in vacuo. The DCM phase (0.2021 g) was absorbed onto diol column material and dried before packing into a 10-g SNAP column [coefficient of variation (CV) = 15 mL; Biotage] with diol material. The extract was then fractionated on an Isolera flash purification system (Biotage) using seven steps of heptane-DCM-EtOAc-MeOH. A flow rate of 20 mL·min^{-1} was used, and fractions were automatically collected with 2 × 2 CVs for each step. Solvents used were of HPLC grade, and H_2O was milliQ-water (purified and deionized using a Millipore system through a 0.22-μm membrane filter). Two of the Isolera fractions were subjected to further purification on separate runs on semipreparative HPLC (Waters 600 Controller with a 996-photodiode array detector). This was achieved using a Luna II C_{18} column (250 mm × 10 mm, 5 μm; Phenomenex). A linear water-MeCN gradient was used starting with 15% MeCN and increasing to 100% over 20 min using a flow rate of 4 mL·min^{-1}. MeCN was of HPLC grade, and H_2O was milliQ-water (purified and deionized using the Millipore system through a 0.22-μm membrane filter); both were added to 50 ppm of TFA. The fractions obtained from the separate runs were pooled, and a final purification using the same method yielded 1.5 mg of nidulanin A.

Marfey's Method. Stereoisometry of the amino acids was elucidated using Marfey's method (1). One hundred micrograms of the peptide was hydrolyzed with 200 μL of 6 M HCl at 110 °C for 20 h. To the hydrolysis product (or 2.5 μmol of standard D- and L-amino acids) was added 50 μL of water, 20 μL of 1 M $NaHCO_3$ solution, and 100 μL of 1% 1-fluoro-2-4-dinitrophenyl-5-L-alanine amide (FDAA) in acetone, followed by reaction at 40 °C for 1 h. The reaction mixture was removed from the heat and neutralized with 10 μL of 2 M HCl, and the solution was diluted with 820 μL of MeOH to a total volume of 1 mL. The retention times of the FDAA derivatives were compared with retention times of the standard amino acid derivatives.

Analysis. Analysis was performed using ultra-high-performance liquid chromatography (UPHLC) UV/Vis diode array detector (DAD) high-resolution MS (HRMS) on a maXis G3 orthogonal acceleration (OA) quadrupole–quadrupole time of flight (QQ-TOF) mass spectrometer (Bruker Daltonics) equipped with an electrospray injection (ESI) source and connected to an Ultimate 3000 UHPLC system (Dionex). The column used was a reverse-phase Kinetex 2.6-μm C_{18}, 100 mm × 2.1 mm (Phenomenex), and the column temperature was maintained at 40 °C. A linear water-acetonitrile gradient was used (both solvents were buffered with 20 mM formic acid) starting from 10% (vol/vol) MeCN and increased to 100% in 10 min, maintaining this rate for 3 min before returning to the starting conditions in 0.1 min and staying there for 2.4 min before the following run. A flow rate of 0.4 mL·min^{-1} was used. HRMS was performed in ESI$^+$ with a data acquisition range of 10 scans per second at m/z 100–1,000. The mass spectrometer was calibrated using bruker daltonics high precision calibration (HPC) by means of the use of the internal standard sodium formate, which was automatically infused before each run. UV spectra were collected at wavelengths from 200 to 700 nm. Data processing was performed using DataAnalysis software (Bruker Daltonics). HRMS analysis of nidulanin A was measured to 604.3497 Da corresponding to a molecular formula of $C_{34}H_{45}N_5O_5$ (deviation of –0.6 ppm).

NMR. The 1D and 2D spectra were recorded on a Bruker Daltonics Avance 800-MHz spectrometer equipped with a 5-mm TCI Cryoprobe at the Danish Instrument Centre for NMR Spectroscopy of Biological Macromolecules at Carlsberg Laboratory. Spectra were acquired using standard pulse sequences, and a 1H spectrum, as well as COSY, NOESY, heteronuclear single quantum coherence (HSQC), and heteronuclear multiple bond correlation (HMBC) spectra, were acquired. The deuterated solvent was acetonitrile-d_3, and signals were referenced by solvent signals for acetonitrile-d_3 at δ_H = 1.94 ppm and δ_C = 1.32/118.26 ppm. The NMR data were processed using Topspin 3.1 (Bruker Daltonics). Chemical shifts are reported in parts per million (δ), and scalar couplings are reported in hertz. The sizes of the J coupling constants reported in the tables are the experimentally measured values from the spectra. There are minor variations in the measurements, which may be explained by the uncertainty of J. NMR data for nidulanin A are presented in Table S3, and the structure is shown in Fig. S6.

Protein Domain Predictions. Nonribosomal peptide synthase (NRPS) protein domains were predicted using the analysis tool of Bachmann and Ravel (2) with the standard settings. Only domains with significant P values ($P < 0.05$) were included in the analysis. Adenylation domain specificities were predicted using NRPSpredictor (3).

Structural Elucidation. The 1H NMR spectrum of nidulanin A displayed four resonances at δ_H 8.16, 7.91, 7.64, and 7.51 ppm, which were identified as amide protons indicative of a nonribosomal peptide type of compound. For each resonance, a COSY correlation to a proton further up-field in the α-proton area could be observed. This coupled each of the amide protons to Hα protons at resonances of δ_H 4.82, 3.92, 4.56, and 3.85 ppm, respectively. Investigation of the NOESY connectivities allowed for assembling of the peptide backbone, which revealed a cyclical tetrapeptide as illustrated in Fig. S7.

The two protons at δ_H 7.64 and 4.56 ppm were part of a larger spin system with correlations to a couple of diastereotopic protons at δ_H 3.02 [1H, doublet of doublets (dd), 14.4, 8.0] and 2.82 (1H, dd, 14.3, 7.5) ppm, as well as five aromatic protons at δ_H 7.14 [1H, multiplet (m)], 7.21 (2H, m), and 7.22 (2H, m). HMBC correlations from the diastereotopic pair as well as the aromatic protons revealed a quaternary carbon with a carbon chemical shift of 137.5 ppm. This information, put together, led to the amino acid phenylalanine. The protons at δ_H 7.91 and 3.92 ppm, as well as the protons at δ_H 7.51 and 3.85 ppm, had very similar spin systems. In both spin systems, a single proton appeared (δ_H 1.93 and 1.96, both multiplets), as well as two methyl groups as doublets (δ_H 0.71/0.78 ppm and 0.84/0.79 ppm). In both cases, the amino acid could be established as valine. Elucidation of the final part of the structure showed that this was not one of the standard proteinogenic amino acids. For this final part, three different spin systems, as well as two isolated methyl groups, were present, which could be linked together by HMBC correlations as well as NOESY connectivities. The first spin system consisted of the amide proton at δ_H 8.16 ppm, the H_α proton at 4.82 ppm, and a diastereotopic pair of protons at δ_H 3.63 (1H, dd, 17.7, 9.7) and 3.09 (1H, dd, 17.6, 4.9) ppm. The second spin system consisted of four aromatic protons at δ_H 7.79 (1H, dd, 8.2, 1.5), 7.28 [1H, doublet of doublets of doublets (ddd), 8.6, 7.0, 1.5], 6.81 (1H, dd, 8.7, 0.7), and 6.57 (1H, ddd, 8.6, 7.0, 1.1) ppm, whereas the third and final spin system contained three protons located in the double-bond area at δ_H 5.95 (1H, dd, 17.6, 10.7), 5.13 (1H, dd, 10.7, 1.0), and 5.15 (1H, dd, 17.6, 1.0) ppm. The latter was shown to be connected to the two methyl groups at δ_H 1.39 [3H, singlet (s)] and 1.38 (3H, s) ppm, and the presence of a quaternary carbon at δ_C 53.7 ppm linked this part as an isoprene unit. The entire residue and key HMBC correlations for the structural elucidation of this part are shown in Fig. S8. The residue contains the amino acid L-kynurenine, which is an intermediate in the tryptophan degradation pathway. In this structure, L-kynurenine has been further modified, because the aforementioned isoprene unit has been incorporated onto the amine located at the aromatic ring.

To establish the stereochemistry of nidulanin A, Marfey's analysis (1) was performed. This technique enables one to determine the absolute configuration of amino acids in peptides (1). The analysis showed the phenylalanine residue present was L-phenylalanine, whereas the analysis for valine showed equal amounts of L- and D-valine.

We used bioinformatics prediction algorithms to identify the stereochemistry of the added amino acids further. Both NRPS protein domain predictions and adenylation domain specificity predictors identify four adenylation domains, corresponding to the four amino acids of the cyclopeptide. By comparison of predictions and the known sequence, the specificity and sequence of the adenylation domains were assigned as predicted to Phe-Kyn-Val-Val. The last two adenylation domains give similar predictions, further supporting both to be specific for valine.

The structure with the proposed absolute chemistry is given in Fig. 4. The absolute configuration of the kynurenine, as well as the order of the L- and D-valine, which is based solely on the bioinformatic studies, has not been verified chemically.

1. Marfey P (1984) Determination of D-amino acids. II. Use of a bifunctional reagent,1,5-difluoro-2,4-dinitrobenzene. *Carlsberg Res Commun* 49:591.
2. Bachmann BO, Ravel J (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181–217.
3. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res* 33(18): 5799–5808.

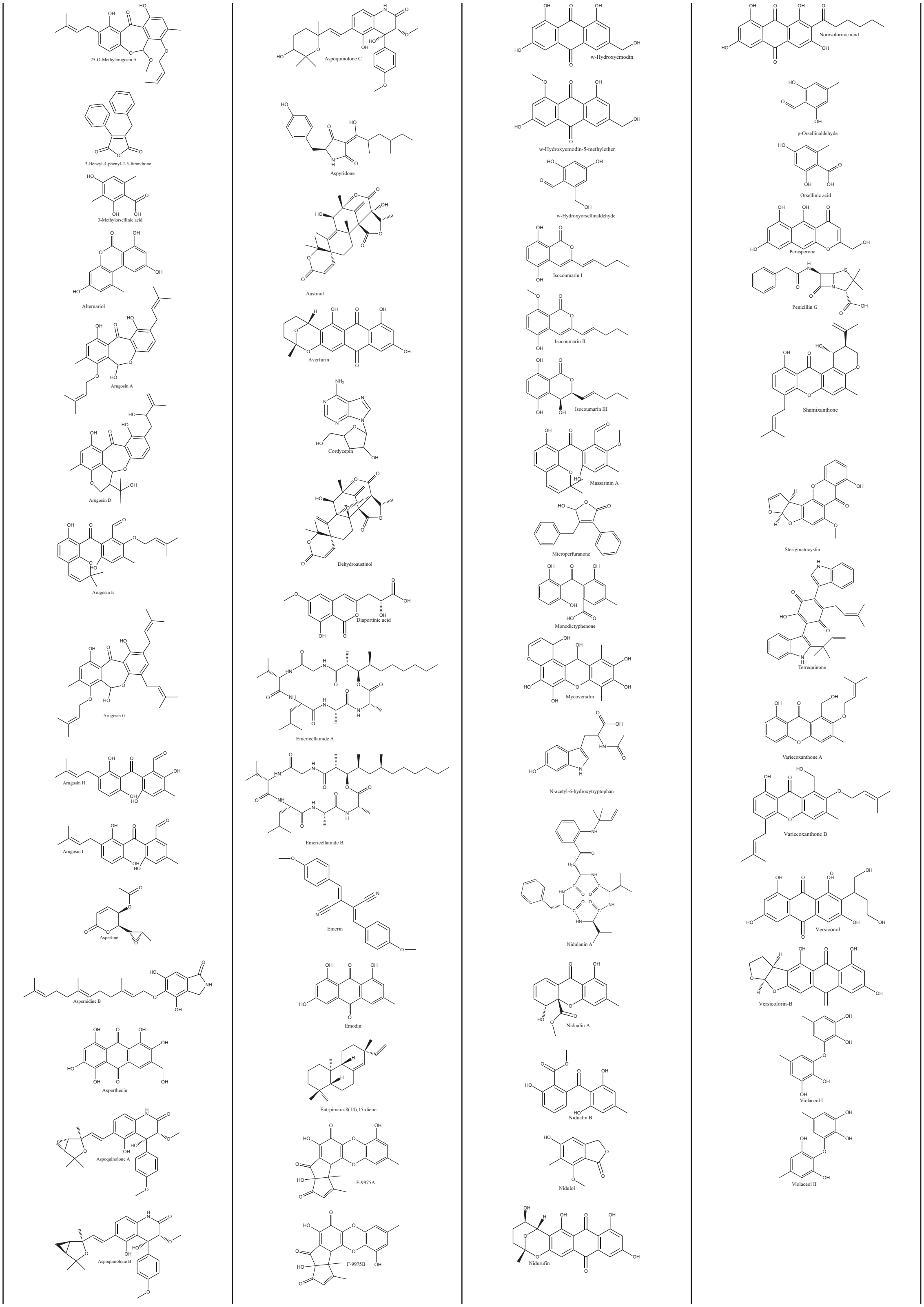


Fig. S1. Chemical structures of secondary metabolites. Structures are shown in alphabetical order in columns from left to right.

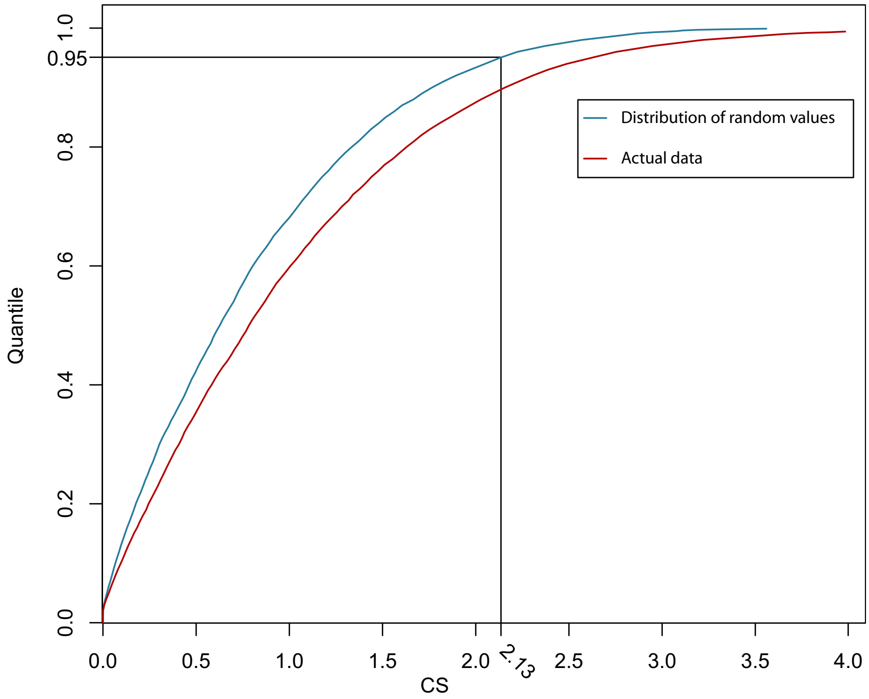


Fig. S2. Quantile plot of clustering scores (CSs). The gray line plots the quantile for a given value of the CS based on a random combination of genes (*Materials and Methods*). Ninety-five percent of the values attained are 2.13 or below (as shown). The red line is a plot of the quantiles of actual values for the genes, as can be found in Dataset S2.

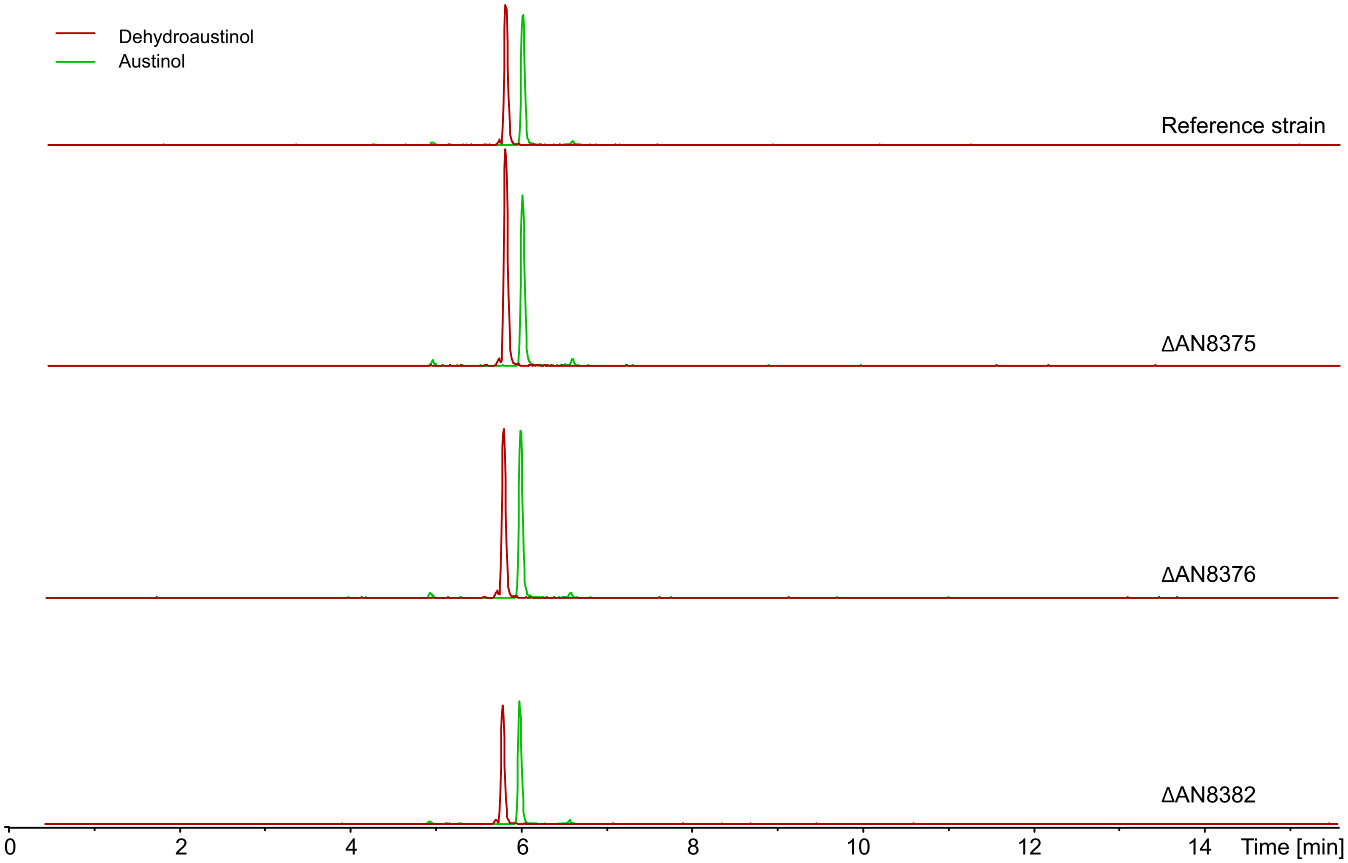


Fig. S3. Extracted ion chromatograms (EICs) for austinol and dehydroaustinol (mass tolerance ± 0.005 Da) from UHPLC-DAD-HRMS of chemical extractions from the reference strain and the ΔAN8375, ΔAN8376 and ΔAN8382 strains. DAD, diode array detector.

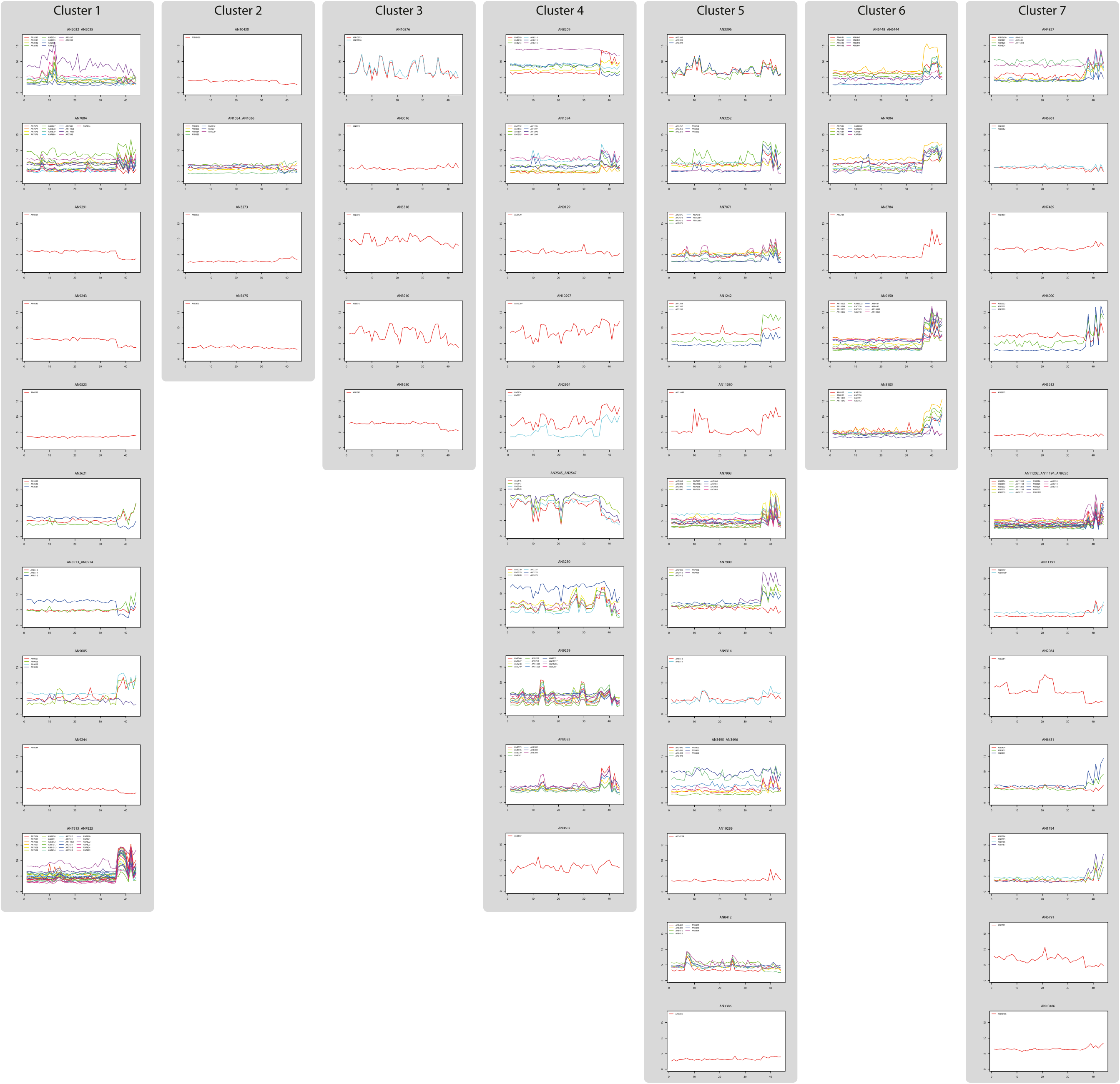


Fig. S4. Overview of the gene expression profiles for all predicted members of the biosynthetic gene clusters (Tables 1–3 and Dataset S2). The *y* axis indicates the gene expression index on a $\log_2$ scale, and the *x* axis represents the 44 experimental conditions included in the microarray compendium. The biosynthetic clusters are sorted into the Superclusters indicated in Fig. 3.

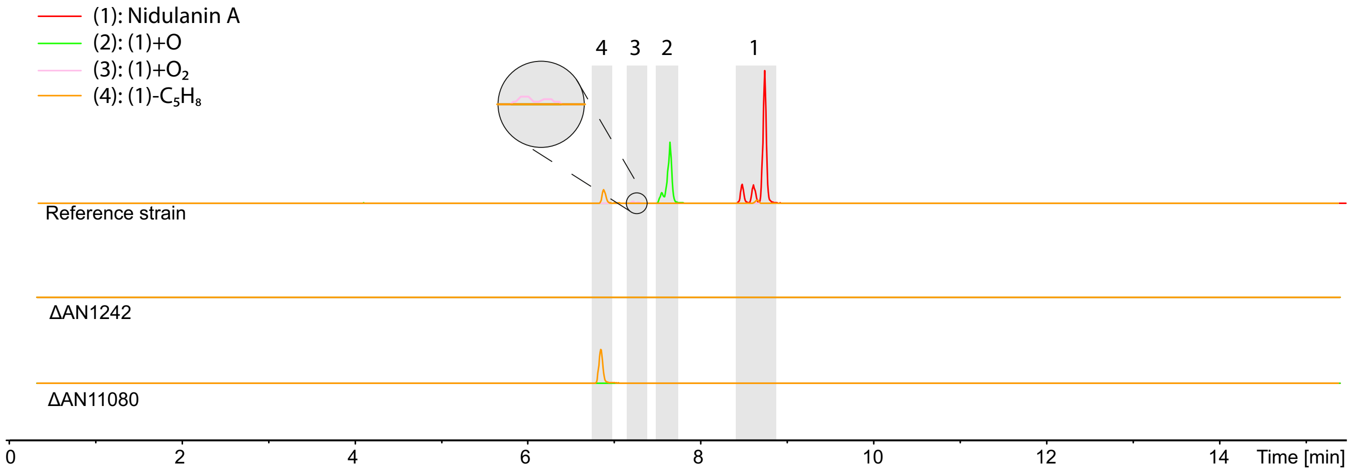


Fig. S5. Extracted ion chromatograms (EICs) for compounds 1–4. Mass tolerance ± 0.005 Da from UHPLC-DAD-HRMS of chemical extractions from the reference strain and ΔAN1242 and ΔAN11080 strains. DAD, diode array detector.

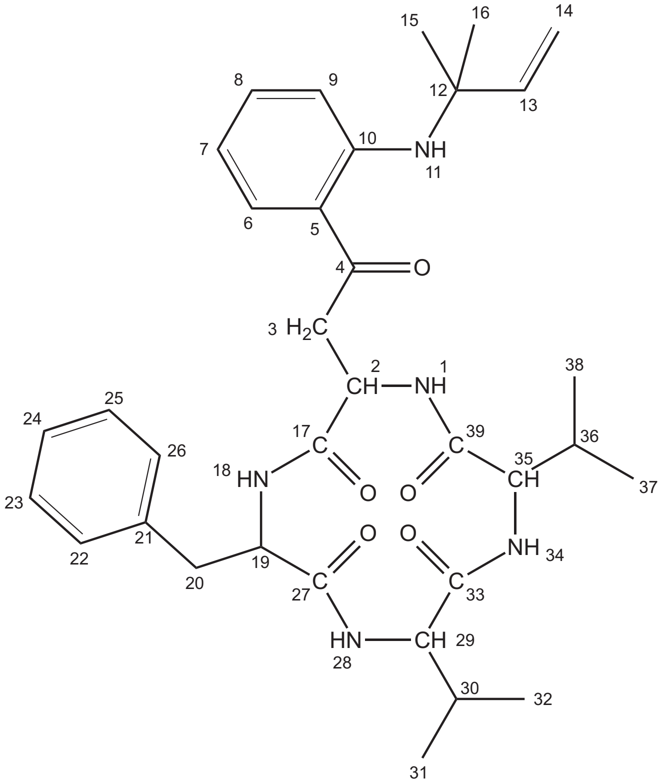


Fig. S6. Structure of nidulanin A, including numbering of individual atoms.

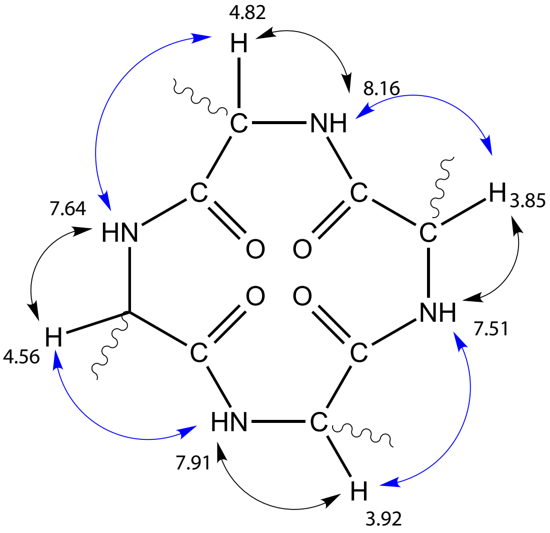


Fig. S7. COSY (black) and NOESY (blue) connectivities lead to the cyclical tetrapeptide.

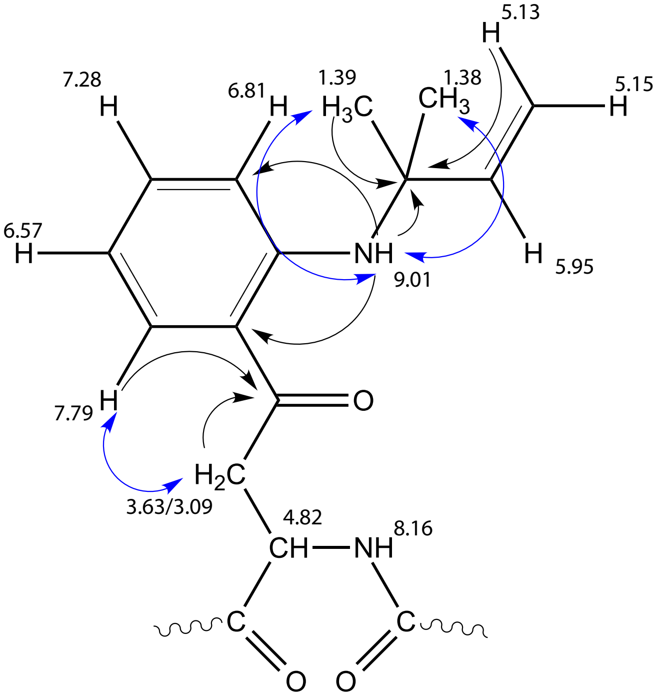


Fig. S8. Key HMBC (black) correlations and NOESY (blue) connectivities link the different spin systems.

Table S1. List of predicted biosynthetic genes where low expression indices with low variation across the 44 conditions were found

GeneID	Type	Gene name	Compound (if known)	Gene no.	Known	Refs.
AN0523	PKS				4	(1)
AN1034	PKS	*afoE*	Asperfuranone	5	7	(2, 3)
AN1036	PKS	*afoG*	Asperfuranone	5	7	(2, 3)
AN10430	PKS					
AN3273	PKS					
AN3386	PKS					
AN3612	PKS					
AN5475	PKS					
AN6961	NRPS					
AN9243	NRPS					
AN9244	NRPS					
AN6810	DTS					

These genes are assumed to be silent in all 44 conditions. DTS, diterpene synthase; NRPS, nonribosomal peptide synthase; PKS, polyketide synthase.

1. Bromann K, et al. (2012) Identification and characterization of a novel diterpene gene cluster in Aspergillus nidulans. *PLoS ONE* 7(4):e35450.
2. Chiang YM, et al. (2008) Molecular genetic mining of the Aspergillus secondary metabolome: Discovery of the emericellamide biosynthetic pathway. *Chem Biol* 15(6):527–532.
3. Bergmann S, et al. (2010) Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. *Appl Environ Microbiol* 76(24):8143–8149.

Table S2. Primer sequences

Primer name	Sequence
AN1242-DL-Up-F	GAGATCGTCGATGGAGTGGCG
AN1242-DL-Up-Rad	gatccccgggaattgccatgCTGCGAGGCACATCATGTTGCC
AN1242-DL-Dw-Fad	aattccagctgaccaccatgGGGTCTGGGTACGCGGGTTTG
AN1242-DL-Dw-R	GATGTGTAGGCGCGACATGGG
AN1242-CHK-Up-F	CCGTCATCATCGTTATAGCC
AN1242-CHK-Dw-R	GCACCCGCTATCACATAC
AN1242-GAPCHK-F	GGCATTATGTGAGCTGTCGTG
AN1242-GAPCHK-R	GATGGAGGGCTTGGTCTTGG
AN1242-INTCHK-R	GATCGAGACGGGTCGTTTAGG
AN11080-DL-Up-FU	GGGTTTAAUGGCAGGTACCAATAATGA
AN11080-DL-Up-RU	GGACTTAAUAGATATACGAGTATGCGG
AN11080-DL-Dw-FU	GGCATTAAUAGTGCCTGATAACTCTGC
AN11080-DL-Dw-RU	GGTCTTAAUGTTGAATCCCTCTGCCTT
AN11080-CHK-Up-F	GGACGGCCCATATTCAGA
AN11080-CHK-Dw-R	AATAAGCTGTAGCGGCGA

Table S3. NMR data for nidulanin A in acetonitrile-d_3

Atom assignment	1H-chemical shift, ppm/J coupling constants, Hz	^{13}C-chemical shift, ppm	HMBC correlations	NOE connectivities
1	8.16 (1H, d, 9.2)	—	—	2, 3, 3′, 35
2	4.82 (1H, ddd, 8.6, 7.0, 1.5)	48.2	—	1, 3, 3′, 18
3	3.63 (1H, dd, 17.7, 9.7)	50.0	2, 4, 17	1, 2, 3′, 6
3′	3.09 (1H, dd, 17.6, 4.9)	50.0	4	1, 2, 3, 6
4	—	198.9	—	—
5	—	116.6	—	—
6	7.79 (1H, dd, 8.2, 1.5)	131.8	4, 8, 10	3, 3′, 7
7	6.57 (1H, ddd, 8.6, 7.0, 1.1)	114.3	5, 9	6
8	7.28 (1H, ddd, 8.6, 7.0, 1.5)	134.0	6, 10	—
9	6.81 (1H, dd, 8.7, 0.7)	114.9	5, 7	15, 16
10	—	148.7	—	—
11	9.01 (1H, s)	—	5, 9, 12	15, 16
12	—	53.7	—	—
13	5.95 (1H, dd, 17.6, 10.7)	145.0	—	14, 14′
14	5.13 (1H, dd, 10.7, 1.0)	113.3	12	13
14′	5.15 (1H, dd, 17.6, 1.0)	113.3	12, 13	13
15	1.39 (3H, s)	27.8	12, 13, 16	9, 11
16	1.38 (3H, s)	27.5	12, 13, 15	9, 11
17	—	172.1	—	—
18	7.64 (1H, d, 8.4)	—	—	2, 19, 20, 20′
19	4.56 (1H, q, 8.0)	53.7	—	18, 28, 20, 20′
20	3.02 (1H, dd, 14.4, 8.0)	*	19, 21, 22/26	18, 19, 20′
20′	2.82 (1H, dd, 14.3, 7.5)	*	19, 21, 22/26, 27	18, 19, 20
21	—	137.5	—	—
22	7.22 (1H, m)	128.0	23/25	—
23	7.21 (1H, m)	127.7	21	—
24	7.14 (1H, m)	126.4	—	—
25	7.21 (1H, m)	127.7	21	—
26	7.22 (1H, m)	128.0	23/25	—
27	—	172.7	—	—
28	7.91 (1H, d, 9.2)	—	—	19, 29, 30, 31
29	3.92 (1H, d, 9.5)	59.1	27, 33	28, 30, 31, 32, 34
30	1.93 (1H, m)	26.3	—	28, 29, 31, 32
31	0.71 (3H, d, 6.6)	18.1	29, 30, 32	28, 29, 30
32	0.78 (3H, d, 6.6)	18.9	29, 30, 31	29, 30
33	—	172.2	172.6	—
34	7.51 (1H, d, 9.6)	—	—	29, 36
35	3.85 (1H, dd, 9.8, 10.7)	59.4	33, 39	1, 36, 37, 38
36	1.96 (1H, m)	26.9	—	34, 35, 37, 38
37	0.84 (3H, d, 6.7)	18.1	35, 36, 38	35, 36
38	0.79 (3H, d, 6.6)	18.9	35, 36, 37	35, 36
39	—	172.1	—	—

1H NMR spectrum and 2D spectra were recorded at with a Bruker Daltonics Avance 800 MHz spectrometer at Carlsberg Laboratory. Signals were referenced to the solvent signals for acetonitrile-d_3 at δ_H = 1.94 ppm and δ_C = 1.32/118.26 ppm. There are minor variations in the measurements which may be explained by the uncertainty of *J*. d, doublet; dd, doublet of doublet; ddd, doublet of doublets of doublets; m, multiplet; q, quartet; s, singlet.

*Cannot be unambiguously assigned.

Dataset S1. Overview of UHPLC-DAD-HRMS analysis of chemical extractions from the reference strain on three solid media after 4, 8, or 10 d (4d, 8d, and 10d, respectively)

Dataset S1

Values given are extracted ion chromatogram peak areas. DAD, diode array detector.

Dataset S2. Gene expression indices from 44 experimental conditions sorted according to chromosomal coordinates

Dataset S2

Locus names and annotation from the *Aspergillus* Genome Database (www.ASPGD.org) are given where available. Clustering scores and cluster members are given.

6.6 Paper 6 – Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization

Klitgaard, A., Nielsen, J. B., Frandsen, R. J. N., Andersen, M. R., & Nielsen, K. F.

Paper accepted in Analytical Chemistry (2015)

analytical chemistry

Article

pubs.acs.org/ac

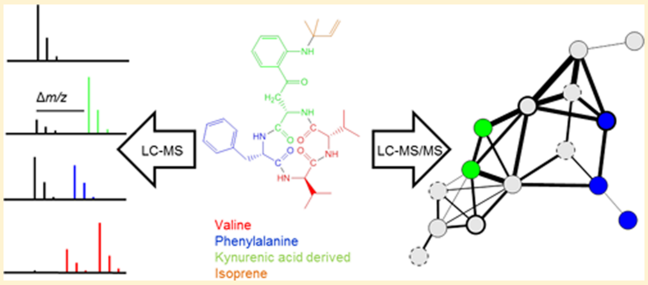
Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization

Andreas Klitgaard, Jakob B. Nielsen, Rasmus J. N. Frandsen, Mikael R. Andersen, and Kristian F. Nielsen*

Department of Systems Biology, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

Supporting Information

ABSTRACT: Filamentous fungi are a rich source of bioactive compounds, ranging from statins over immunosuppressants to antibiotics. The coupling of genes to metabolites is of large commercial interest for production of the bioactives of the future. To this end, we have investigated the use of stable isotope labeled amino acids (SILAAs). SILAAs were added to the cultivation media of the filamentous fungus *Aspergillus nidulans* for the study of the cyclic tetrapeptide nidulanin A. Analysis by UHPLC-TOFMS confirmed that the SILAAs were incorporated into produced nidulanin A, and the change in observed *m/z* could be used to determine whether a compound (known or unknown) incorporated any of the added amino acids. Samples were then analyzed using MS/MS and the data used to perform molecular networking. The molecular network revealed several known and unknown compounds that were also labeled. Assisted by the isotope labeling, it was possible to determine the sequence of several of the compounds, one of which was the known metabolite fungisporin, not previously described in *A. nidulans*. Several novel analogues of nidulanin A and fungisporin were detected and tentatively identified, and it was determined that these metabolites were all produced by the same nonribosomal peptide synthase. The combination of stable isotope labeling and molecular network generation was shown to very effective for the automated detection of structurally related nonribosomal peptides, while the labeling was effective for determination of the peptide sequence, which could be used to provide information on biosynthesis of bioactive compounds.



Filamentous fungi are prolific producers of small bioactive compounds, and the secondary metabolites (SMs) are especially interesting as a source for pharmaceuticals. These include compounds such as the cholesterol-lowering drug lovastatin, the immunosuppressive mycophenolic acid, and the antimicrobial griseofulvin and penicillin.[1] SMs are categorized on the basis of their biosynthetic origin, where the major classes are the polyketides (PKs),[2] nonribosomal peptides (NRPs),[3] and terpenoids,[4] all produced by synthases/synthetases encoded by complex biosynthetic genes clusters. In fungi, NRP synthases (NRPSs) consist of modules responsible for the binding of amino acids (AAs) and stepwise coupling of the peptide. Unfortunately, it is still not possible to accurately predict the AAs encoded by these modules and, hence, the product of the NRPS. This makes it difficult to predict the products of a given synthetase and the involved biosynthetic pathway.

Studies of biosynthetic pathways using radioactive labeled substrates have been performed since the 1950s[5] using sensitive radiation detectors.[6,7] However, advances in GC/MS and LC-MS instrumentation has made it possible to use stable isotope labeled (SIL), without the risks associated with handling radioactive material.[8] One approach is ^{13}C biosynthetic pathway elucidation where a known precursor of a compound of interest is added to the cultivation media of an organism, and the mass spectrum of a given compound is then compared to the predicted ^{13}C labeling pattern.[8] This approach has been used in many experiments, including studies of the aflatoxin pathway,[7] the asticolorin pathway,[9] and recently the yanuthone D pathway.[10] Studies in bacteria have shown that cultivation in the presence of labeled AAs could be used to aid characterization of linear NRPs by tandem MS analysis.[11−13] However, even though interpretation of fragmentation spectra of linear NRPs is a well-established technique, fragmentation patterns of cyclic peptides (often containing nonproteinogenic AAs and/or organic acids) are known to be complex,[14] making the characterization of them by MS/MS difficult at best. Fungi are able to take up AAs from their environment,[15] a property that has been used previously to study incorporation of stable isotope labeled amino acids (SILAAs) into proteins from filamentous fungi using LC-MS.[16,17] SILAAs might therefore be a suitable route for introducing NRP precursors into fungi to probe the NRP pathways.

To investigate the biosynthesis of compounds, the molecular networking method developed by Dorrestein and co-workers[18] can be used to investigate compounds of interest. The method is based on characterizing molecules using MS/MS, after which the fragmentation spectra of the molecules are clustered on the basis of similarity. This can be visualized in a network, in a way

Received: December 10, 2014
Accepted: May 28, 2015
Published: May 28, 2015

DOI: 10.1021/acs.analchem.5b01934

where compounds exhibiting similar fragmentation spectra are grouped together in clusters. Compounds that are biosynthetically related can share structural similarities, which can result in similar fragmentation spectra, leading to the formation of clusters of biosynthetically related compounds. This approach has been used in the investigation of peptides from *Streptomyces roseosporus*[19] and analysis of compounds produced by gut microbiota.[20] This new approach allows for faster examination of biosynthesis compared to traditional labor intensive methods relying on stepwise gene deletion and analysis.

One of the most extensively investigated filamentous fungi is *Aspergillus nidulans* which houses a high number of putative SM pathways.[21] Recent genome-based studies report that *A. nidulans* has 12 NRPSs, 14 NRPS-like proteins, 32 PKSs, 1 PKS-NRPS, and 26 terpene synthase or cyclase encoding genes,[22−24] and much of this metabolic potential is still to be characterized. New products are still being discovered, such as the mixed NRP-terpene nidulanin A (Figure 1A).[21] This is a cyclic tetrapeptide consisting of one L-phenylalanine (Phe) residue, one L-valine (Val) and one D-Val residues, one L-kynurenine residue, and an isoprene unit. Nidulanin A proved difficult to isolate in sufficient quantities for structure elucidation by NMR, and thus, two putative analogs were not isolated and fully characterized.[21]

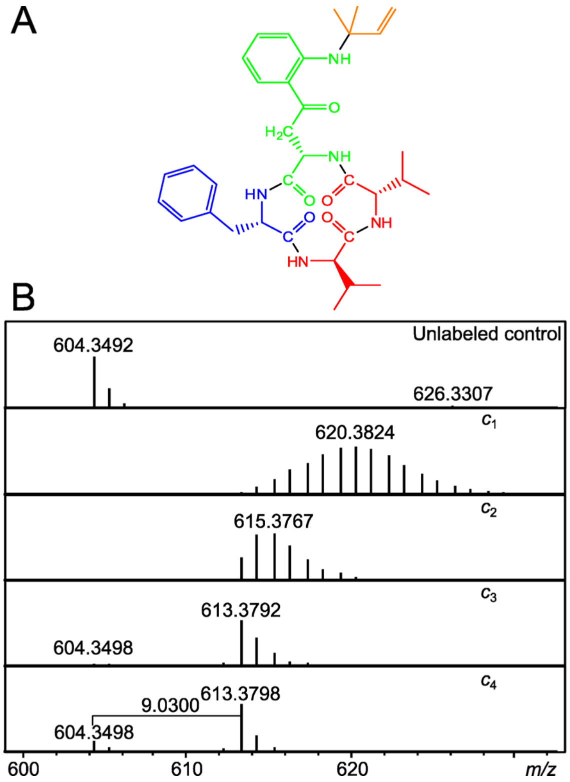


Figure 1. (A) Structure of nidulanin A. The coloring illustrates the different biosynthetic units that make up the metabolite: blue, Phe; green, kynurenine; red, Val; orange, isoprene. (B) Mass spectra of nidulanin A at different concentrations of $^{13}C_9{}^{15}N$-labeled Phe added to *A. nidulans* IBT 4887, cultivated at 25 °C in the darkness for 7 days on MM.

In this study, we propose a new method for characterization of SM biosynthetic pathways. The method combines an experimental protocol as well as recently developed MS/MS networking tools and proves to be very powerful for: (i) highlighting novel compounds produced by the organism, (ii) assisting in characterizing the biosynthetic pathways responsible of their synthesis; (iii) and assisting in probing the structure of NRPs by MS/MS. We illustrate the workflow and demonstrate the effectiveness of the method by applying it to the study of the biosynthesis of the compound nidulanin A and related products produced by *A. nidulans.*

■ EXPERIMENTAL SECTION

Chemicals. Solvents were LC-MS grade, and all other chemicals were analytical grade. All were from Sigma-Aldrich (Steinheim, Germany) unless otherwise stated. Water was purified using a Milli-Q system (Millipore, Bedford, MA). ESI-TOF tune mix was purchased from Agilent Technologies (Torrance, CA, USA).

The labeled AAs were purchased from Cambridge Isotope Laboratories (Andelover, MA, USA) and Sigma-Aldrich. The AAs were labeled to different degrees: L-valine ($^{13}C_5$, 97−99%), L-phenylalanine ($^{13}C_9{}^{15}N$, 98% ^{13}C, 98% ^{15}N), anthranilic acid (ring $^{13}C_6$, 99%), L-tryptophan (D_8, 98%), and L-tyrosin ($^{13}C_9{}^{15}N$, 96%).

LC-MS Analysis. All samples were analyzed as described in previously published work.[25] In summary, samples were analyzed on a Dionex Ultimate 3000 UHPLC system (Thermo Scientific, Dionex, Sunnyvale, California, USA) equipped with a Kinetex C_{18} column (100 × 2.1 mm, 2.6 μm particles) (Phenomenex, Torrance, CA, USA) running an acidic water/ACN gradient. This was coupled to Bruker maXis 3G quadrupole time-of-flight mass spectrometer (Q-TOF-MS) system (Bruker Daltonics, Bremen, Germany) equipped with an ESI source operating in positive polarity.

LC-MS/MS Analysis for Molecular Network Analysis. Samples for the molecular network analysis were analyzed using the same system described in previously published work.[26] Samples were analyzed using an Agilent LC-MS system comprising an Agilent 1290 Agilent 1290 infinity UHPLC (Agilent Technologies, Torrence, CA, USA) equipped with an Agilent Poroshell 120 phenyl-hexyl column (250 mm × 2.1 mm, 2.7 μm particles), running an acidic water/ACN gradient. This was coupled to an Agilent 6550 Q-TOF-MS equipped with an iFunnel ESI source operating in positive polarity.

For the network analysis, automated data-dependent MS/HRMS was performed for ions detected in the full scan at an intensity above 1.500 counts at 10 scans/s in the range of m/z 200−900, with a cycle time of 0.5 s, a quadrupole isolation width of $m/z \pm 0.65$ using a collision energy of 25 eV and a maximum of 3 selected precursors per cycle, and an exclusion time of 0.04 min. Differentiation of molecular ions, adducts, and fragment ions was done by chromatographic deconvolution and identification of the $[M + Na]^+$ ion.[25]

Molecular Network Analysis. Samples for the molecular networking analysis were analyzed using the Agilent LC-MS system. The network was created using data from fungi cultivated without SILAA as well as data from fungi cultivated with one type of SILAA. No data from fungi cultivated with multiple SILAAs were included. Data was converted from the standard .d (Agilent standard data-format) to .mgf (Mascot Generic Format) using the software MSConvert which is part of the ProteoWizard[27] (vers. 3.0.4738) project. The converted data-files were processed using the molecular networking method developed by Dorrestein and co-workers.[18] The following settings were used for generation of the network: Minimum pairs, Cos 0.65; parent mass tolerance, 2.0 Da; ion

DOI: 10.1021/acs.analchem.5b01934

Table 1. SILAAs Used in the Experiment[a]

AA	elemental composition	monoisotopic mass [Da]	mass difference [Da]	start concentration in media (c)			
				c_1[M]	c_2 [M]	c_3 [M]	c_4[M]
Phe	$^{13}C_9H_{11}{}^{15}NO_2$	175.1062	10.0272 (9.0302)[b]	1.7×10^{-2}	5.7×10^{-3}	1.9×10^{-3}	6.4×10^{-4}
Val	$^{13}C_5H_{11}NO_2$	122.0958	5.0168	4.7×10^{-3}	1.6×10^{-4}	5.2×10^{-5}	1.7×10^{-6}
anthranilic acid	$^{13}C_6{}^{12}CH_7NO_2$	143.0678	6.0201	7.4×10^{-3}	2.5×10^{-3}	8.3×10^{-4}	2.8×10^{-4}
Trp	$C_{11}D_8H_4N_2O_2$	212.1401	8.0502	5.6×10^{-3}	1.9×10^{-3}	3.2×10^{-3}	1.1×10^{-3}
Tyr	$^{13}C_9H_{11}{}^{15}NO_3$	191.1011	10.0272 (9.0302)[b]	2.9×10^{-2}	9.7×10^{-3}	3.2×10^{-3}	1.1×10^{-3}

[a] Mass difference denotes the mass difference between the SILAA and the naturally predominant isotope. [b] Mass difference due only to ^{13}C-labeling.

tolerance, 0.5; network topK, 100; minimum matched peaks, 6; minimum cluster size, 2.

The molecular networking workflow is publically available online.[28] The molecular networking data was analyzed and visualized using Cytoscape (vers. 2.8.2).[29]

Preparation of Fungi. Three different wild-type strains of *A. nidulans*, IBT 4887 (A4), 22818, and 25683 were three-point inoculated on solid Czapek yeast autolyzate (CYA)[30] media. Fungal strains are available from the IBT culture collection at the authors' address. The *nlsA*Δ mutant strain (AN1242Δ, IBT 30029)[21] was inoculated on solid CYA media with added arginine supplements (4 mM). All fungi were incubated at 25 °C in darkness for 7 days in standard 9 cm diameter Petri plates. To each plate was then added 2.5 mL of autoclaved Milli-Q water, and the spores were suspended using a Drigalski spatula. The AA sequence for HcpA (CAP93139.1) was obtained from GenBank (NCBI), and sequences for An08g02310 and AN1242.5 were obtained from the AspGD portal. Pairwise alignments for sequence similarity were conducted in the NCBI/BLAST/blastp suite.[31]

Preparation of Labeling Solutions. Several different concentrations (Table 1) of AAs in the media were tested to determine the best for incorporation. Solutions were prepared by dissolving the AAs in Milli-Q water followed by sterile filtering of the solutions.

Inoculation of Fungi. Liquid minimal media (MM) was prepared as in Nielsen et al.[32] but without addition of agar. The fungi were cultivated in 12-well plates with a well size of 2 mL from Nunc (Cat. No. 150200, Roskilde, Denmark). To each well 1.2 mL of MM was added followed by 0.4 mL of AA solution when testing one AA or 0.2 mL of each solution when testing two AAs. Finally, the fungus was inoculated by transfering 5 μL of one of the spore suspensions to the well. The plate was then sealed with an Aeraseal breathable sealing film (Cat. No. A9224-50EA, Excel Scientific, Victorville, Ca, USA) to prevent contamination while allowing for exchange of gases. The fungi were kept stagnant while incubated at 25 °C in the darkness for 7 days.

Extraction of Fungi. After incubation, the mat-like biomass was removed from the wells using a needle and transferred to a 4 mL glass vial. The biomass was extracted using acidic ethyl acetate−dichloromethane−methanol (3:2:1 v/v/v) as described by Smedsgaard.[33]

RESULTS AND DISCUSSION

Exploring Labeling of Nidulanin A. A visual inspection (Supporting Information, Figure S1) of the fungi revealed a slight increase in sporulation at the highest tested concentrations (c_1) of AAs (Table 1). This was most likely because the AAs were used as the carbon and nitrogen source for the organism leading to a richer growth medium. Addition of anthranilic acid completely inhibited growth in the three highest tested concentrations but resulted in no changes at the lowest tested concentration (c_4). Additions of the AAs did not result in the immediate detection of any new compounds; however, it did alter the intensities of some compounds up to 10-fold, although none of these up-regulated compounds showed any signs of AA incorporation in their mass spectra (Supporting Information, Figure S2).

Peaks corresponding to known NRPs produced by *A. nidulans*, such as nidulanin A and the emericellamides, were investigated to determine if incorporation of SILAAs could be detected. However, nidulanin A was initially the only compound for which incorporation of SILAAs could be detected as its production seems to be linked to biomass production. The $^{13}C_9{}^{15}N$-labeled Phe used in the experiment should induce a mass shift of m/z 10.0223 if incorporated directly into a compound. However, during cellular uptake, the nitrogen atom will be exchanged, meaning that incorporation leads to a mass shift of m/z 9.0302. The mass spectra seen in Figure 1B exhibited changes depending on the concentration of labeled Phe in the growth medium. At the highest tested concentration (c_1), there was no trace of the m/z 604.3490 ion of protonated nidulanin A, and instead, the mass spectrum took on a bell shape centered on m/z 620.382. This bell shape occurred because Phe was both used as a substrate for the central carbon cycle and directly incorporated into nidulanin A. This means that the general concentration of ^{13}C in the medium in the fungal cells was increased enough to lead to distorted isotope patterns. At the lowest tested concentration (c_4), the mass spectrum exhibited two distinct signals. One is the protonated ion corresponding to the $[M + H]^+$ ion of nidulanin A, while the other was m/z 9.0300 higher corresponding to the mass difference of a substitution of $^{12}C_9$ to $^{13}C_9$ atoms. A similar type of experiment should be conducted prior to studying the effect of SILAAs on other species of fungi, media, and culture conditions, as it would be expected to vary depending on cellular metabolism.

The kynurenine residue in nidulanin A should be biosynthetically derived from Trp, and it was there to test whether Trp could be added to the fungus and catabolized into kynurenine acid followed by incorporation into nidulanin A. However, addition of labeled Trp did not result in incorporation at any of the tested concentrations. To investigate whether the kynurenine unit was formed prior to incorporation into nidulanin A, $^{13}C_6$-labeled anthranilic acid was tested as it is a precursor to Trp and hence kynurenine. The mass spectrum of nidulanin A at the lowest concentration (c_4) of anthranilic acid (Figure 2) showed the occurrence of a new ion at m/z 610.3701, a shift of m/z 6.0200 compared to unlabeled nidulanin A, corresponding to the incorporation of

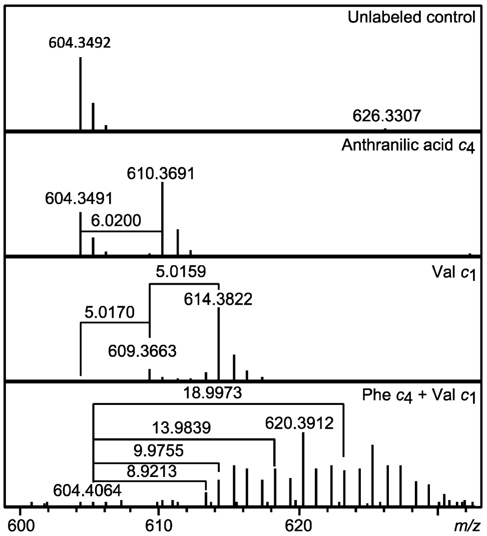


Figure 2. Mass spectra of nidulanin A showing incorporation of tested anthranilic acid and Val. High incorporation was observed in the case of the addition of $^{13}C_6$-labeled anthranilic acid. The addition of $^{13}C_5$-labeled Val formed two distinct ions as incorporation of both the one and two residues was observed. Addition of both $^{13}C_9$-Phe and $^{13}C_5$-labeled Val lead to a complex isotope pattern, containing ions corresponding to incorporation of both Val and Phe.

$^{13}C_6$.This indicated that anthranilic acid was used as a substrate by the NRPS and further biosynthesized into kynurenine.

Mass spectra obtained from analysis of *A. nidulans* cultivated with $^{13}C_5$ labeled Val (Figure 2) showed no trace of unlabeled nidulanin A, but it showed two ions with a *m*/*z* difference of 5.0163, corresponding to nidulanin A with one and two Val residues ($^{13}C_5$) incorporated, respectively. Nidulanin A contains two Val residues, and the results showed a very high degree of labeled Val incorporation.

In the original paper describing nidulanin A,[21] two putative analogues differing in mass corresponding to incorporation of one and two oxygen atoms, respectively, were reported. It was hypothesized that one of these analogs could be a compound where Tyr was incorporated instead of Phe. To test the hypothesis, cultivation experiments were performed using $^{13}C_9{}^{15}N$-labeled Tyr. Mass spectra of the two analogues (See Supporting Information, Figures S3 and S4) showed the incorporation of $^{13}C_9$ atoms indicating incorporation of Tyr, thus confirming the previous hypothesis.

After the initial successful experiments, addition of multiple different AAs to the growth medium at the same time was tested, using the concentrations (c_4) that were found to have the best results. In the experiment, both labeled Phe and Val was added to the growth medium, which was predicted to result in incorporation of three AAs. The mass spectrum obtained from the analysis (Figure 2) depicts a very complex substitution pattern. This was most likely because the nidulanin A could possibly be labeled with both Phe and Val in different amounts leading to five different possible combinations of the labeling (Phe, Val, 2 Val, Phe + Val, and Phe +2 Val).

Spectra (obtained from the Bruker maXis) contained many of the same ions identified in the previous experiment, but the mass accuracy was poor even after calibration. The sample was reanalyzed to investigate whether the poor mass accuracy and isotopic pattern could be caused by insufficient resolution of the MS during recording of data in centroid mode. However,

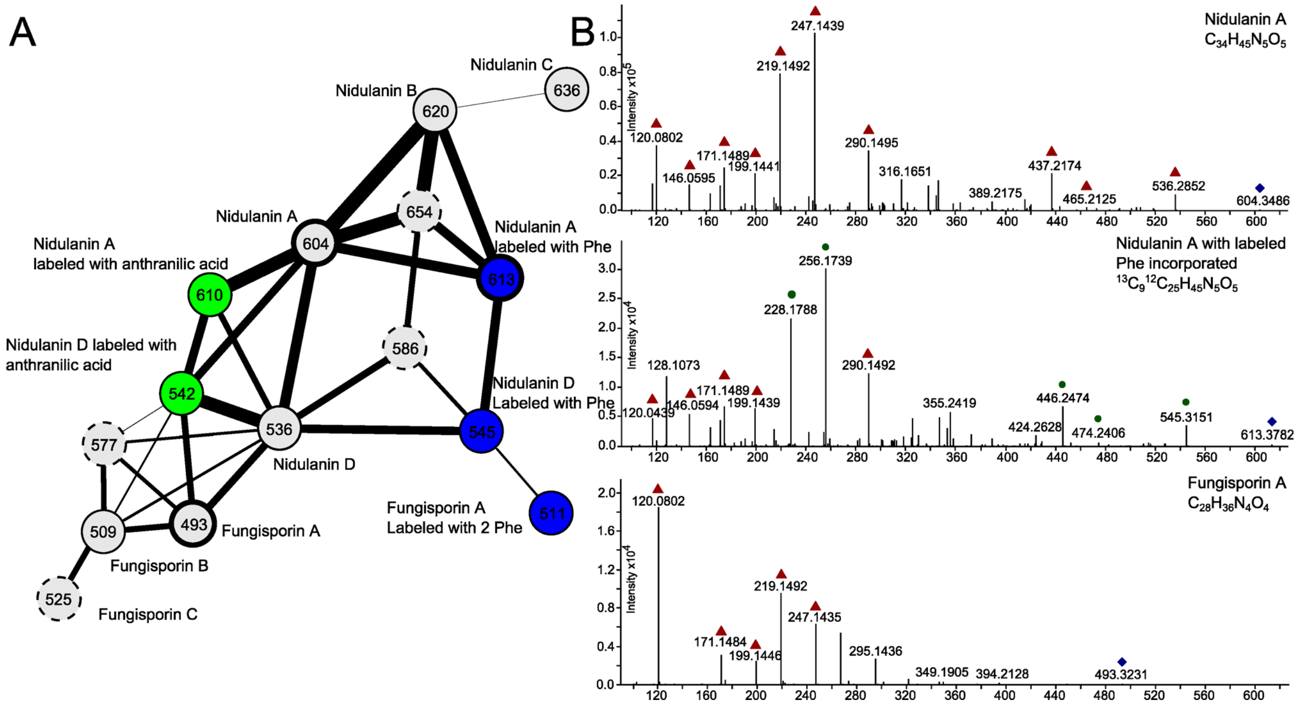


Figure 3. (A) Subcluster containing a node corresponding to nidulanin A and several previously described analogues. The circles represent the consensus MS/MS spectrum for a given parent mass (decimals removed for legibility). The thickness of the black lines connecting the nodes (circles) indicates the similarity of the MS/MS spectra for the connected nodes, as scored by the networking algorithm. Previously undescribed compounds are marked with a dashed outline. (B) MS/MS spectra of three nodes in bold are shown. Blue diamonds denote the product ion of the compounds; red triangles denote fragments formed by the unlabeled nidulanin A, while the green circles denote fragments found in nidulanin A that now contain labeled atoms.

Table 2. Investigated Compounds[a]

						labeling information				
name	RT [min]	molecular formula	AA composition	modification	m/z [M + H]$^+$	Phe	Val	Ant	Tyr	Trp
nidulanin A	8.7	$C_{34}H_{45}N_5O_5$	Phe-Kyn-Val-Val	prenylated	604.3493	1	2	1	–	–
nidulanin B	7.7	$C_{34}H_{45}N_5O_6$	Tyr-Kyn-Val-Val	prenylated	620.3443	1	2	1	1	–
nidulanin C	7.3	$C_{34}H_{45}N_5O_7$	not determined	prenylated	636.3392	1	2	1	1	–
nidulanin D	7.0	$C_{29}H_{37}N_5O_5$	Phe-Kyn-Val-Val		536.2867	1	2	1	–	–
fungisporin A	7.3	$C_{28}H_{36}N_4O_4$	Phe-Phe-Val-Val[b]		493.2809	2	2	–	–	–
fungisporin B	6.3	$C_{28}H_{36}N_4O_5$	Tyr-Phe-Val-Val[b]		509.2758	2	2	–	1	–
fungisporin C	5.4	$C_{28}H_{36}N_4O_6$	Tyr-Tyr-Val-Val[c]		525.2708	1	1	–	1	–
	7.3	$C_{33}H_{44}N_4O_5$	not determined[c]		577.3384	2	2	–	1	–
	7.9	$C_{35}H_{39}N_7O_6$	not determined[c]	prenylated	654.3035	1	–	1	1	–
	6.7	$C_{30}H_{31}N_7O_6$	not determined[c]		586.2401	1	–	1	1	–
fungisporin D	7.2	$C_{30}H_{37}N_5O_4$	Phe-Trp-Val-Val[b]		532.2924	1	2	1	–	–

[a] The column labeling information denotes the number of specific labeled AA residues detected for each compound. (−) no detection of incorporation. AA composition refers to the identity and sequence of the AAs in the compound. For some compounds, the AA composition could not be determined. [b] Compound also described by Ali et al.[35] [c] Previously undescribed compound.

no difference was observed when the samples were reanalyzed in profile mode. As in the experiment with addition of Phe (Figure 1B), the isotope pattern formed a bell shaped pattern centered on m/z 620.3, indicating that the concentrations of the AAs used were too high.

Molecular Network Analysis Revealing New Analogs. Samples were taken from fungi cultivated both with and without labeled AAs. The entire molecular network generated (Supporting Information, Figure S5) contained several distinct smaller separate subnetworks. Utilizing the information from the labeling experiment, the masses of the nodes were investigated to find nodes that differed in m/z according to the predicted shifts obtained from incorporation of the SILAAs. A subnetwork containing a node corresponding to nidulanin A, as well as several nodes corresponding to nidulanin A labeled with AAs, was identified, as seen in Figure 3A. MS/MS spectra that exhibit the same fragment ions or the same neutral losses will be connected in the network. The thickness of the line indicates the similarity of the MS/MS spectra of the compounds, as scored by the networking algorithm. Biosynthetically similar compounds might therefore be grouped together using the generated molecular networks. This subnetwork also contained several nodes that corresponded to the previously reported[21] oxygenated forms of nidulanin A that contained one and two extra oxygen molecules, respectively, as well as an unprenylated form. In addition, several nodes corresponding to unknown compounds were also found, as described in Table 2. The subnetwork is depicted in the Supporting Information, Figure S6, with all decimals for the masses.

SILAA Incorporation Supports Structure Determination. A comparison of the MS/MS spectra from nidulanin A and nidulanin A labeled with Phe (Figure 3B) as well nidulanin A labeled with Val and anthranilic acid (Supporting Information, Figure S7) and the unprenylated form (Supporting Information, Figure S8) allowed for easier assignment of the fragments, as the labels conferred information about the substructure. This information was used to determine the AA sequence of the peptide, although it gives no information on the stereochemistry. Investigation of the fragment m/z 247 showed that it was composed of both a Phe and Val residue. This was supported by results from the feeding studies where addition of labeled $^{13}C_9$-Phe and $^{13}C_5$-Val lead to the formation of fragments of m/z 256 and m/z 252, respectively, corresponding to incorporation of the labeled AAs. Assigned fragments (Supporting Information, Table S1) could also be used to provide structural information on the unknown compound m/z 493, as its MS/MS spectrum displayed several of the same fragments. By using these fragments, it was possible to determine that the unknown compound contained a Phe-Val-Val peptide and that, on the basis of the fragmentation spectrum, the peptide was most likely cyclic. By examining the labeling pattern of the unknown compound (Supporting Information, Figure S9), it was found that Phe was not incorporated when using the lowest concentration (c_4) but only when using higher concentrations (c_1−c_3). The mass spectrum of the compound showed incorporation of two $^{13}C_9$-labeled Phe residues as well as two $^{13}C_5$-labeled Val-residues, while the MS/MS spectrum also exhibited a fragmentation ion corresponding to two linked Phe residues (Supporting Information, Table S2).

Reinvestigation of the subcluster also showed a node corresponding to m/z 511, which fit with the incorporation of two labeled Phe-residues. On the basis of the labeling pattern and fragmentation spectra, the compound was shown to be a cyclic tetrapeptide with the sequence Phe-Phe-Val-Val. This compound has previously been described in the literature as the compound fungisporin and has been isolated from spores from several species of *Penicillium* and *Aspergillus*.[34]

Two nodes corresponding to fungisporin with one and two oxygen atoms incorporated were also detected, analogous to the ones detected for nidulanin A. Labeling experiments again showed (See Supporting Information, Figures S10 and S11) that Tyr was incorporated into the metabolites; see Table 2. The production of fungisporin has recently been linked to a specific NRPS, HcpA, in *P. chrysogenum* by Ali and coworkers.[35] In that study, 10 different cyclic tetrapeptides were found to be produced by the NRPS, including fungisporin and an analog containing a Tyr instead of a Phe residue. The authors also found this pool of 10 cyclic tetrapeptides to be produced by *A. niger*. Pairwise alignments of amino acid sequences of HcpA to the orthologous NRPS of *A. niger* and NlsA in *A. nidulans* indeed showed a relatively high degree of conservation with 55% and 51% identity on the amino acid level, respectively. Moreover, the order of predicted domains is equivalent for the three orthologous proteins except for the lack of the cryptic condensation domain in HcpA. However, we do not observe indications of NlsA being unusual and non-canonical as was reported for HcpA.[21]

Investigation of the two previously reported analogues of nidulanin A showed incorporation of one Tyr residue, accounting for the analog with one extra oxygen. Unfortunately, we were unable to determine the full structure of the analog with a molecular mass corresponding to the incorporation of two extra oxygen atoms. The subnetwork (Figure 3A) contained three additional nodes corresponding to unknown compounds. From the MS/MS spectra of these compounds (See Supporting Information, Figure S12), labeling spectra (See Supporting Information, Figures S13−S16), and fragments (Table 2), it was likely that the compounds with m/z 586 and 654 were prenylated and unprenylated forms of the same compound. MS/MS spectra obtained of the compounds exhibited a formation of fragments corresponding to two linked Val-residues, but the compounds were present in too small quantities to allow for full structure determination. Examination of the results from the Tyr-labeling showed that mass spectra exhibited mass shifts indicative of incorporation of one Tyr-residue, but none containing two Tyr residues was detected. A plausible explanation could be that the degree of incorporation was too low to observe this, and it is speculated that the real structure does indeed contain two Tyr residues.

Nidulanin and Fungisporin Are Products Originating from the Same Biosynthetic Gene. It was investigated if the other cyclic tetrapeptides described by Ali et al.[35] were produced by *A. nidulans*, and the analysis showed that one additional form, a *cyclo*-Phe-Trp-Val-Val peptide, was produced, as confirmed by data from the labeling experiment (See Supporting Information, Figure S16). Analysis of the AN1242 deletion strain, which did not express the NlsA gene, showed that none of the cyclic tetrapeptides from Table 2 were produced, demonstrating that the compounds were most likely products of the NlsA gene. This was also supported by the bioinformatic study, which revealed that the NlsA gene from *A. nidulans* showed a relatively high conservation when compared to the HcpA gene in *A. niger*, which has been shown to encode the NRPS responsible for the production of fungisporin.

Molecular networking has previously been used as a dereplication strategy for natural products.[36] Using this approach, the network can be “seeded” by including data-files obtained from analysis of different standards. However, when working with undescribed natural products, standards are of course not available. This can also be the case for compounds isolated and described by other research groups. In some cases, a biosynthetic analog of a compound is not formed in large enough amounts to record a MS/MS spectrum of sufficient quality. In that case, incorporation of SIL precursors could be used to form labeled compounds that would have similar MS/MS spectra to the unlabeled form, thereby helping the molecular network generation. In the case where the recorded MS/MS spectrum of a compound is not found to be similar to any other in the molecular network, stable isotope labeling could then be used to artificially form a similar compound that would then cluster with the compound of interest. This could potentially be used to expand the usage of molecular networking for compounds that do not form as characteristic fragments as NRPs, for instance PKs.

■ CONCLUSION

In this study, we demonstrated a combined approach for elucidation and characterization of biosynthetic pathways. By combining SIL and molecular networking, it was possible to find new and undescribed metabolites in *A. nidulans*, one of the most investigated filamentous fungi. The effectiveness of the method was illustrated using the secondary metabolite nidulanin A from the filamentous fungus *A. nidulans*. The experiments were conducted in three different wild-type strains and showed that it was possible to simply add SILAAs to the growth medium, leading to incorporation of these AAs into produced metabolites, which could be confirmed by LC-HRMS/MS. By using the molecular networking algorithm, it was possible to find several new analogues of the metabolite, as well as to detect known metabolites that were structurally related. The fact that these compounds have not been reported before also highlights the ability of combined approaches to extract spectral features from compounds that might otherwise be overlooked. This was the case for fungisporin and its two different analogues that had not previously been reported from *A. nidulans*. The MS/MS data obtained could be used to determine the order in which the AAs were coupled in the cyclic peptide nidulanin A and could be used to tentatively determine the structure of new metabolites, thus complimenting other techniques such as NMR. It was determined that nidulanin A, fungisporin, and nine other NRPs were produced by the same NRPS, a coupling that had not previously been realized. The described method has been demonstrated to be useful as an exploratory tool, especially when molecular biology can provide information about what AAs are used in the biosynthesis. Further studies, employing a large number of different AAs for different fungi in an automated system, could be used to probe the NRP production of the organisms. Data from these experiments could be investigated in a targeted manner for a specific case like the probing of nidulanin A or for investigation of the whole NRP production in an organism.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Photographs of fungi cultivated with SILAAs and BPCs and mass spectra for all SILAA additions from the experiments. Full molecular network, fragmentation spectra, and assigned fragments. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b01934.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +45 45 25 26 02. E-mail: kfn@bio.dtu.dk.

Author Contributions

The manuscript was written through contributions of all authors.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The study was supported by Grant 09-064967 from the Danish Council for Independent Research, Technology, and Production Sciences. We are grateful to Agilent Technologies for the Thought Leader Donation of the Agilent UHPLC-QTOF system.

■ REFERENCES

(1) Pearce, C. *Adv. Appl. Microbiol.* **1997**, *44*, 1−80.
(2) Hertweck, C. *Angew. Chem., Int. Ed.* **2009**, *48*, 4688−4716.

(3) Finking, R.; Marahiel, M. A. *Annu. Rev. Microbiol.* **2004**, *58*, 453−488.

(4) Keller, N. P.; Turner, G.; Bennett, J. W. *Nat. Rev. Microbiol.* **2005**, *3*, 937−947.

(5) Hanahan, D.; Al-Wakil, S. *Arch. Biochem. Biophys.* **1952**, *37*, 167−171.

(6) Griffith, G. *Mycologist* **2004**, *18*, 177−183.

(7) Townsend, C.; Christensen, S. *Tetrahedron* **1983**, *39*, 3575−3582.

(8) Tang, J. K.-H.; You, L.; Blankenship, R. E.; Tang, Y. J. *J. R. Soc. Interface* **2012**, *9*, 2767−2780.

(9) Steyn, P. S.; Vleggaar, R.; Simpson, T. J. *J. Chem. Soc. Chem. Commun.* **1984**, *3*, 765−767.

(10) Holm, D. K.; Petersen, L. M.; Klitgaard, A.; Knudsen, P. B.; Jarczynska, Z. D.; Nielsen, K. F.; Gotfredsen, C. H.; Larsen, T. O.; Mortensen, U. H. *Chem. Biol.* **2014**, *21*, 519−529.

(11) Bode, H. B.; Reimer, D.; Fuchs, S. W.; Kirchner, F.; Dauth, C.; Kegler, C.; Lorenzen, W.; Brachmann, A. O.; Grün, P. *Chemistry* **2012**, *18*, 2342−2348.

(12) Proschak, A.; Lubuta, P.; Grün, P.; Löhr, F.; Wilharm, G.; De Berardinis, V.; Bode, H. B. *ChemBioChem* **2013**, *14*, 633−638.

(13) Fuchs, S. W.; Sachs, C. C.; Kegler, C.; Nollmann, F. I.; Karas, M.; Bode, H. B. *Anal. Chem.* **2012**, *84*, 6948−6955.

(14) Liu, W.-T.; Ng, J.; Meluzzi, D.; Bandeira, N.; Gutierrez, M.; Simmons, T. L.; Schultz, A. W.; Linington, R. G.; Moore, B. S.; Gerwick, W. H.; Pevzner, P. A.; Dorrestein, P. C. *Anal. Chem.* **2009**, *81*, 4200−4209.

(15) Helmstaedt, K.; Braus, G. H.; Braus-Stromeyer, S.; Busch, S.; Hofmann, K.; Goldman, G. H.; Draht, O. W. In *The Aspergilli Genomics, Medical Aspects, Biotechnology, and Research Methods*; Goldman, G. H., Osmani, S. A., Eds.; CRC Press: Boca Raton, 2007; pp 143−175.

(16) Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. a; Muddiman, D. C. *Anal. Chem.* **2008**, *80*, 4994−5001.

(17) Georgianna, D. R.; Hawkridge, A. M.; Muddiman, D. C.; Payne, G. A. *J. Proteome Res.* **2008**, *7*, 2973−2979.

(18) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743−E1752.

(19) Liu, W.-T.; Lamsa, A.; Wong, W. R.; Boudreau, P. D.; Kersten, R.; Peng, Y.; Moree, W. J.; Duggan, B. M.; Moore, B. S.; Gerwick, W. H.; Linington, R. G.; Pogliano, K.; Dorrestein, P. C. *J. Antibiot. (Tokyo)* **2014**, *67*, 99−104.

(20) Rath, C. M.; Alexandrov, T.; Higginbottom, S. K.; Song, J.; Milla, M. E.; Fischbach, M. A.; Sonnenburg, J. L.; Dorrestein, P. C. *Anal. Chem.* **2012**, *84*, 9259−9267.

(21) Andersen, M. R.; Nielsen, J. B.; Klitgaard, A.; Petersen, L. M.; Zachariasen, M.; Hansen, T. J.; Blicher, L. H.; Gotfredsen, C. H.; Larsen, T. O.; Nielsen, K. F.; Mortensen, U. H. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E99−E107.

(22) Nielsen, M. L.; Nielsen, J. B.; et al. *FEMS Microbiol. Lett.* **2011**, *321*, 157−166.

(23) Bromann, K.; Toivari, M.; Viljanen, K.; Vuoristo, A.; Ruohonen, L.; Nakari-Setälä, T. *PLoS One* **2012**, *7*, No. e35450.

(24) Ahuja, M.; Chiang, Y.-M.; Chang, S.-L.; Praseuth, M. B.; Entwistle, R.; Sanchez, J. F.; Lo, H.-C.; Yeh, H.-H.; Oakley, B. R.; Wang, C. C. C. *J. Am. Chem. Soc.* **2012**, *134*, 8212−8221.

(25) Klitgaard, A.; Iversen, A.; Andersen, M. R.; Larsen, T. O.; Frisvad, J. C.; Nielsen, K. F. *Anal. Bioanal. Chem.* **2014**, *406*, 1933−1943.

(26) Kildgaard, S.; Mansson, M.; Dosen, I.; Klitgaard, A.; Frisvad, J. C.; Larsen, T. O.; Nielsen, K. F. *Mar. Drugs* **2014**, *12*, 3681−3705.

(27) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. *Nat. Biotechnol.* **2012**, *30*, 918−920.

(28) *GnPS: Global Natural Products Social Molecular Networking*; http://gnps.ucsd.edu (accessed Jul 10, 2014).

(29) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. *Bioinformatics* **2011**, *27*, 431−432.

(30) Samson, R. A.; Houbraken, J.; Thrane, U.; Frisvad, J. C.; Andersen, B. *Food and indoor fungi*; Crous, P. W., Samson, R. A., Eds.; CBS-KNAW Fungal Biodiversity Centre: Utrecht, 2010.

(31) Altschul, S.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−410.

(32) Nielsen, M. L.; Nielsen, J. B.; Rank, C.; Klejnstrup, M. L.; Holm, D. K.; Brogaard, K. H.; Hansen, B. G.; Frisvad, J. C.; Larsen, T. O.; Mortensen, U. H. *FEMS Microbiol. Lett.* **2011**, *321*, 157−166.

(33) Smedsgaard, J. *J. Chromatogr. A* **1997**, *760*, 264−270.

(34) Miyao, K. *J. Agric. Chem. Soc. Japan* **1955**, *19*, 86−91.

(35) Ali, H.; Ries, M. I.; Lankhorst, P. P.; van der Hoeven, R. A.; Schouten, O. L.; Noga, M.; Hankemeier, T.; van Peij, N. N.; Bovenberg, R. A.; Vreeken, R. J.; Driessen, A. J. *PLoS One* **2014**, *9*, No. e98212.

(36) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686−1699.

Combining stable isotope labeling and molecular networking for biosynthetic pathway characterization

Andreas Klitgaard, Jakob B. Nielsen, Rasmus J. N. Frandsen, Mikael R. Andersen, Kristian F. Nielsen*

Department of Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

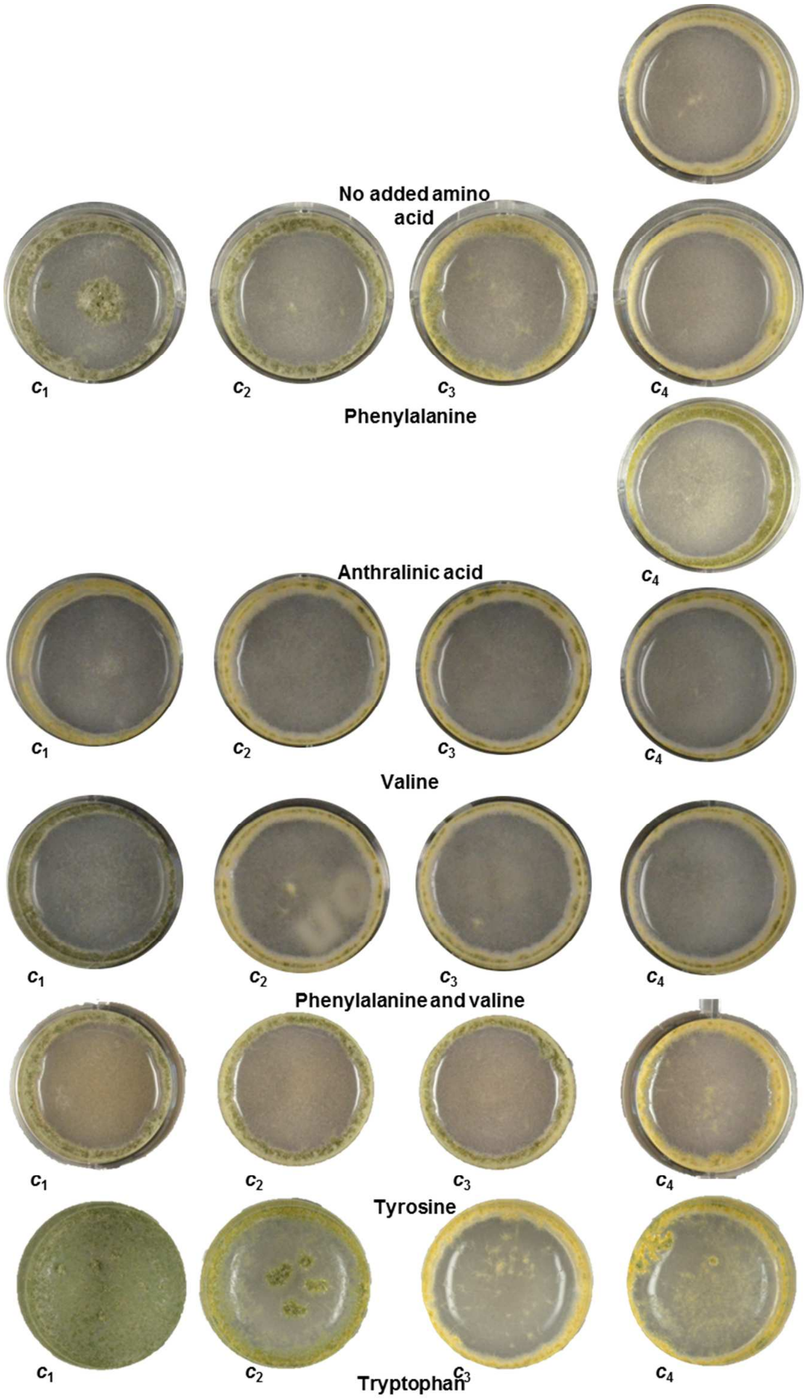


Figure S1. Photographs of *Aspergiullus nidulans* IBT4887 used in the study. The top row is a photo of *A. nidulans* cultivated without the addition of any amino acids. The other rows are photographs of *A. nidulans* cultivated with the addition of the noted amino acids at the indicated concentrations. The addition of anthranilic acid only resulted in growth of the fungus at the lowest tested concentration. The fungi were kept stationary while being incubated at 25 °C in darkness for 7 days in MM without any added amino acids.

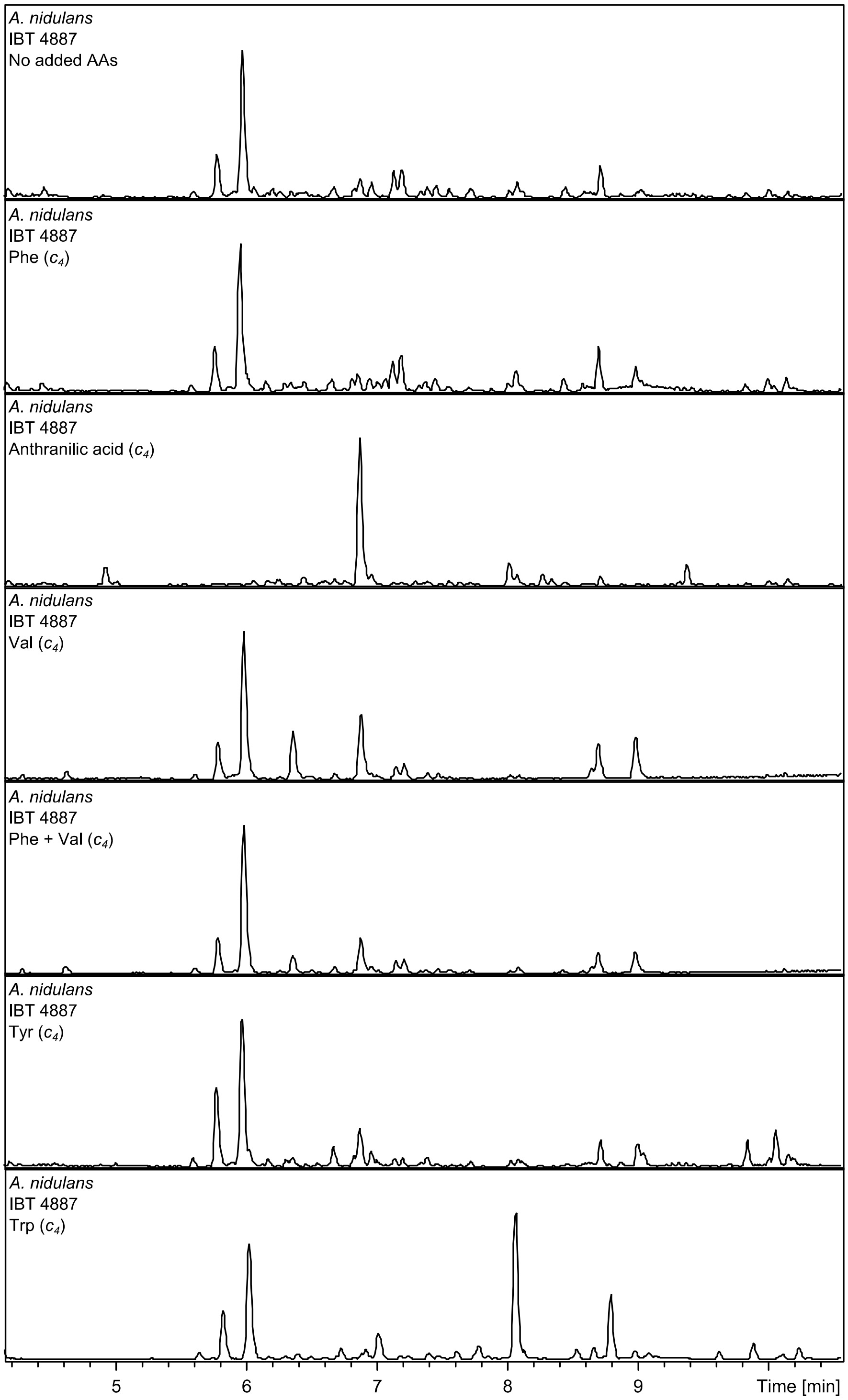


Figure S2. Top is a BPC from *A.* nidulans IBT4887 cultivated without any added amino acids, while the other BPC are from fungi where AAs have been added in the denoted concentration. The chromatograms showed a difference in intensity of several peaks, including peaks at RT 5.8 min (austinol), 6.0 min (dehydroaustinol), and 6.9 min (sterigmatocystin). However, a close inspection of the data showed that no signs of incorporation of labeled AAs in any of the corresponding compounds. The chromatograms have been scaled to the highest signal. The extract from the sample with added Trp showed a strong signal at 8.1 min, which corresponded to a known impurity (tributyrin).

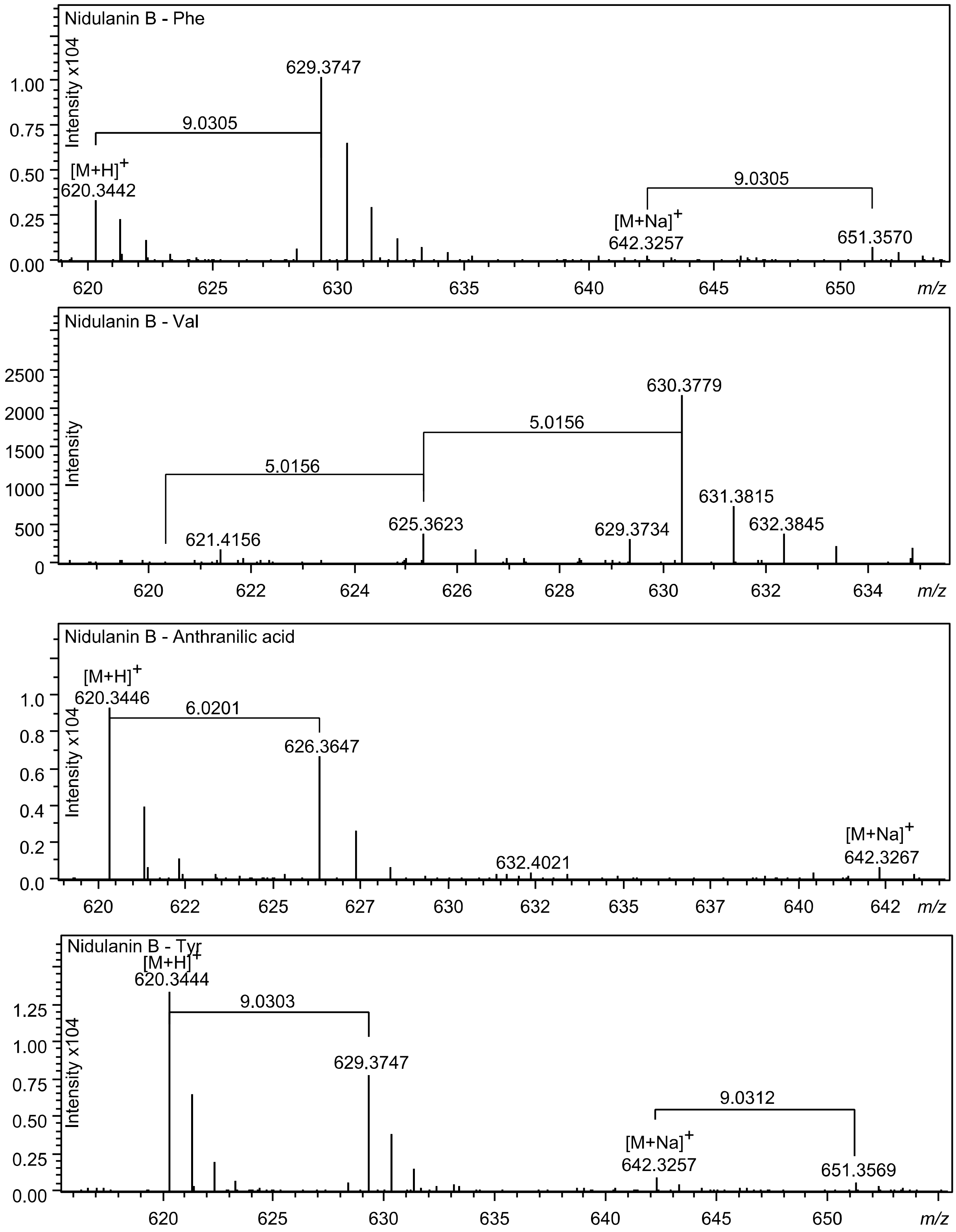


Figure S3. Labeling of nidulanin B. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.70-7.75 min and have been scale to the highest signal.

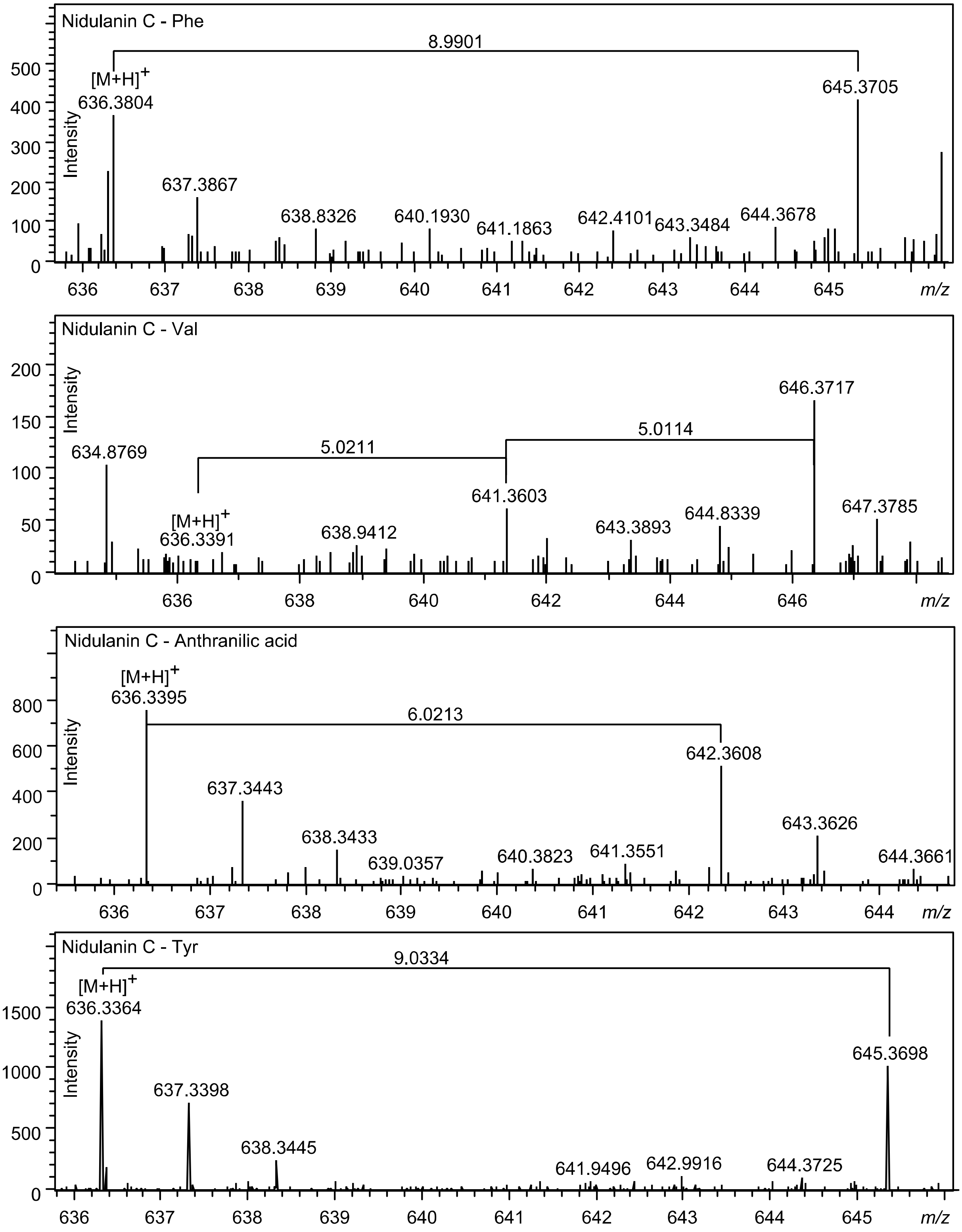


Figure S4. Labeling of nidulanin C. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.27-7.32 min and have been scale to the highest signal.

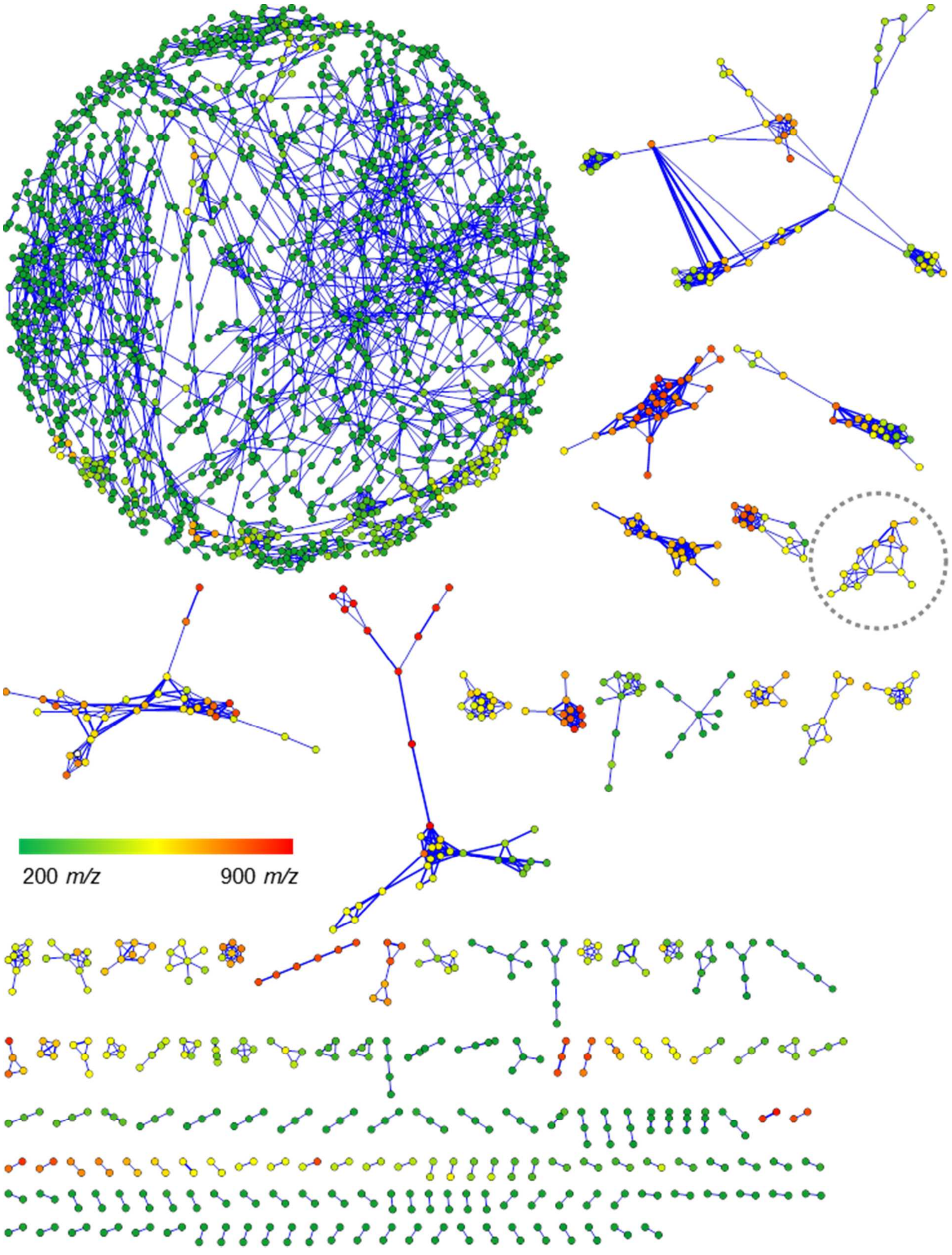


Figure S5. Molecular network generated from analysis of samples from *A. nidulans.* Each circle represents the precursor ion of a given compound where as the color of the circle represents the *m/z*-ratio. The thickness of the blue lines connecting the nodes (circles) indicates the similarity of the MS/MS spectra for the connected nodes, as scored by the networking algorithm. The network was constructed based on samples from experiments with and without addition of stable isotope labeled AAs. The sub-network marked with the dotted ring contains a node corresponding to NA.

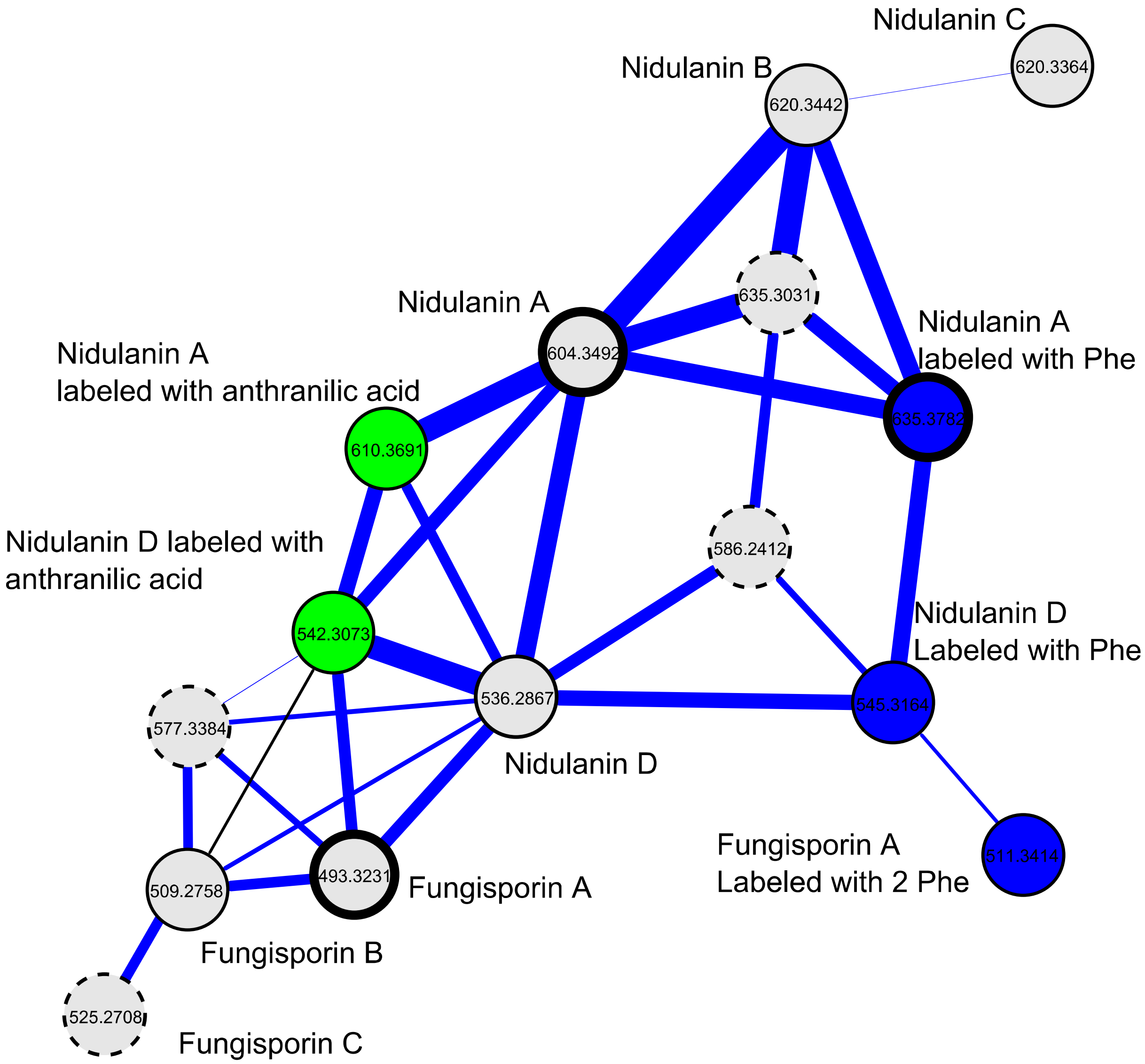


Figure S6 – Sub-cluster containing a node corresponding to nidulanin A and several previously described analogues. The circles represent the consensus MS/MS spectrum for a given parent. The thickness of the blue lines connecting the nodes (circles) indicates the similarity of the MS/MS spectra for the connected nodes, as scored by the networking algorithm. Previously undescribed compounds are marked with a dashed outline.

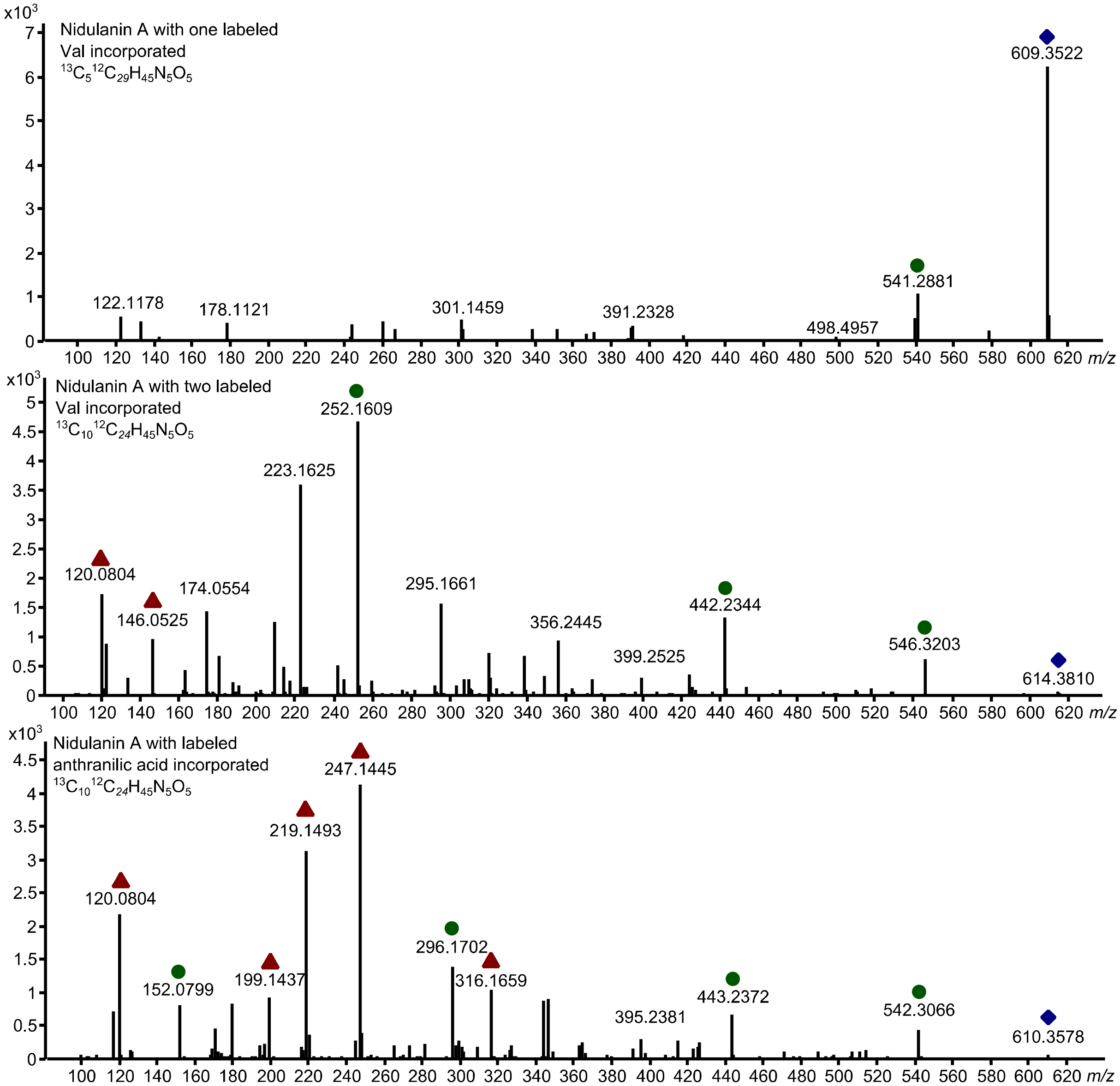
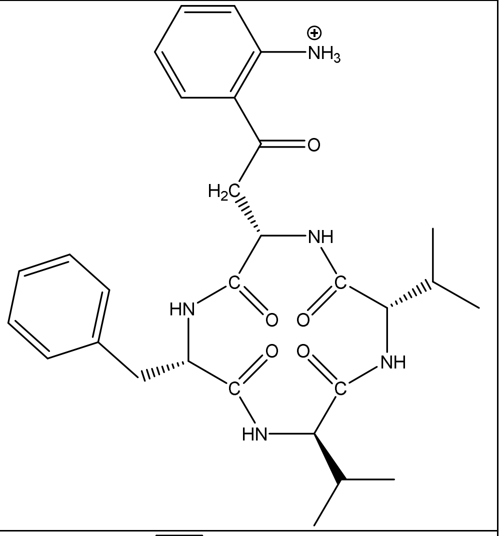
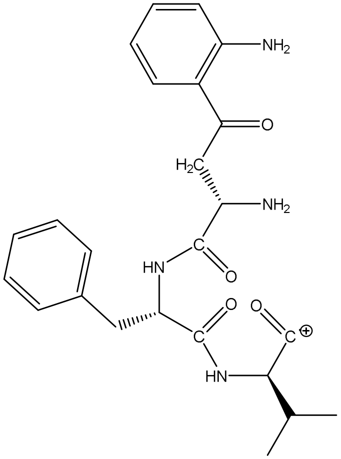
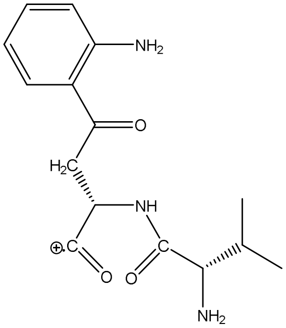
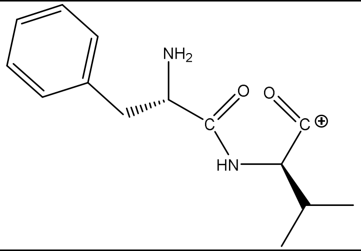
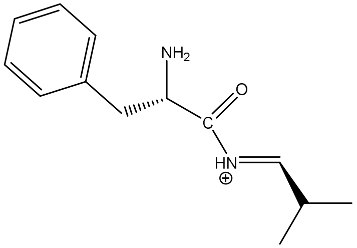
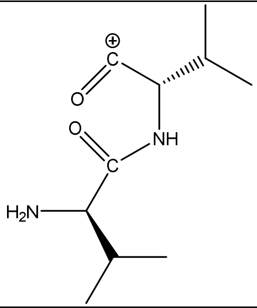
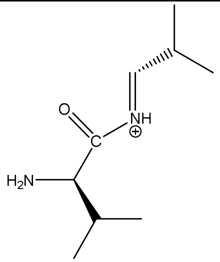
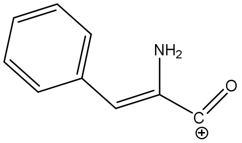
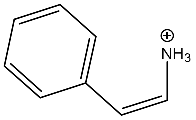


Figure S7. MS/MS spectra obtained from analysis of NA labeled with anthranilic acid as well as one and two Val residues respectively. The blue diamonds denote to the product ion of the compounds, red triangles denote fragments formed by the unlabeled NA, while the green circles denote fragments found in NA that now contain labeled atoms.

Table S1. Fragment ions formed by fragmentation of nidulanin A

Fragment [*m/z*]	Chemical formula	Structure
536	$C_{29}H_{37}N_5O_5$	
437	$C_{24}H_{29}N_4O_4$	
290	$C_{15}H_{20}N_3O_3$	
247	$C_{14}H_{19}N_2O_2$	

219	$C_{13}H_{19}N_2O$	
199	$C_{10}H_{19}N_2O_2$	
171	$C_9H_{18}N_2O$	
146	C_9H_8NO	
120	$C_8H_{10}N$	

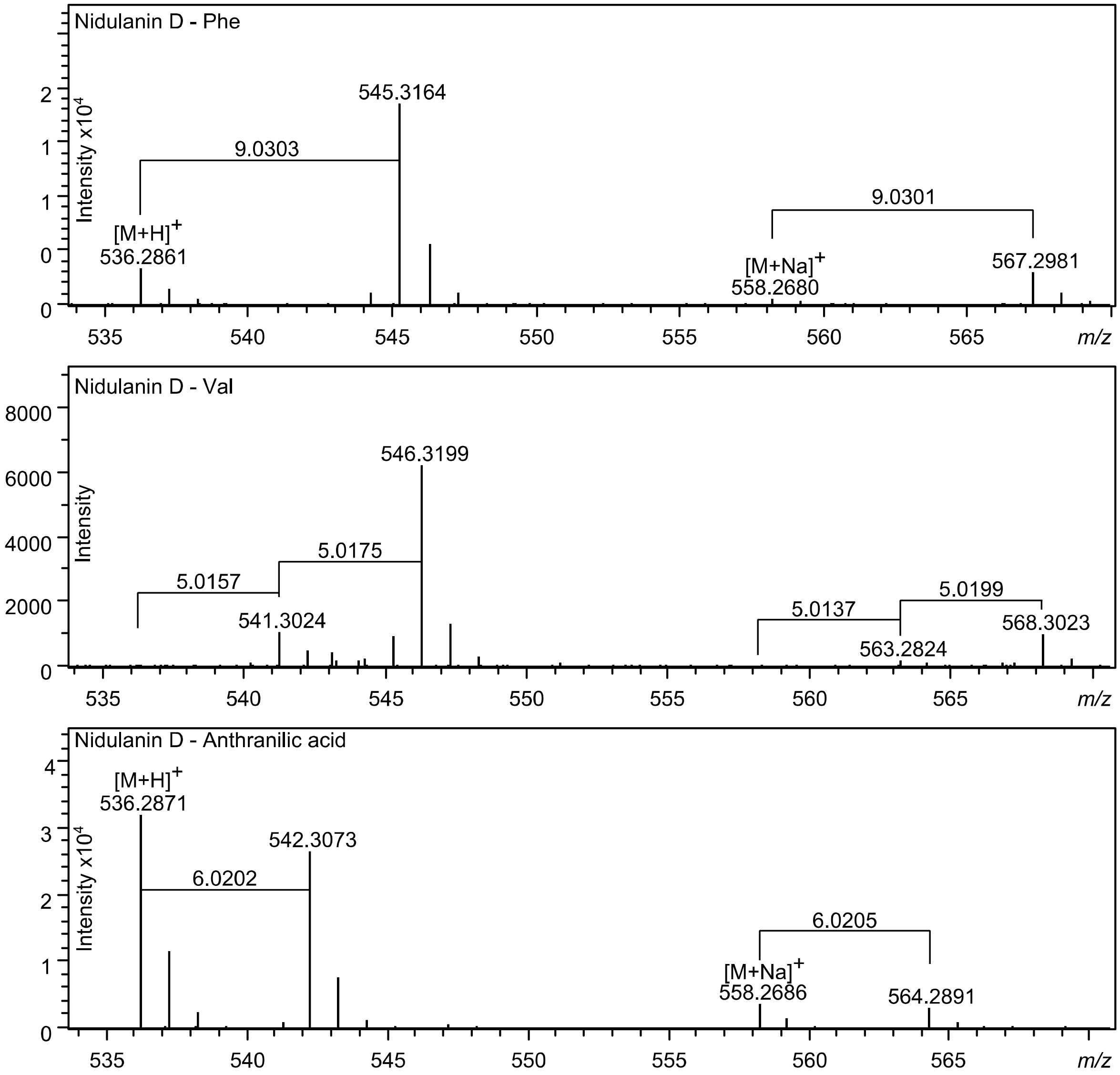


Figure S8. Labeling of nidulanin D. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 6.95-7.00 min and have been scale to the highest signal.

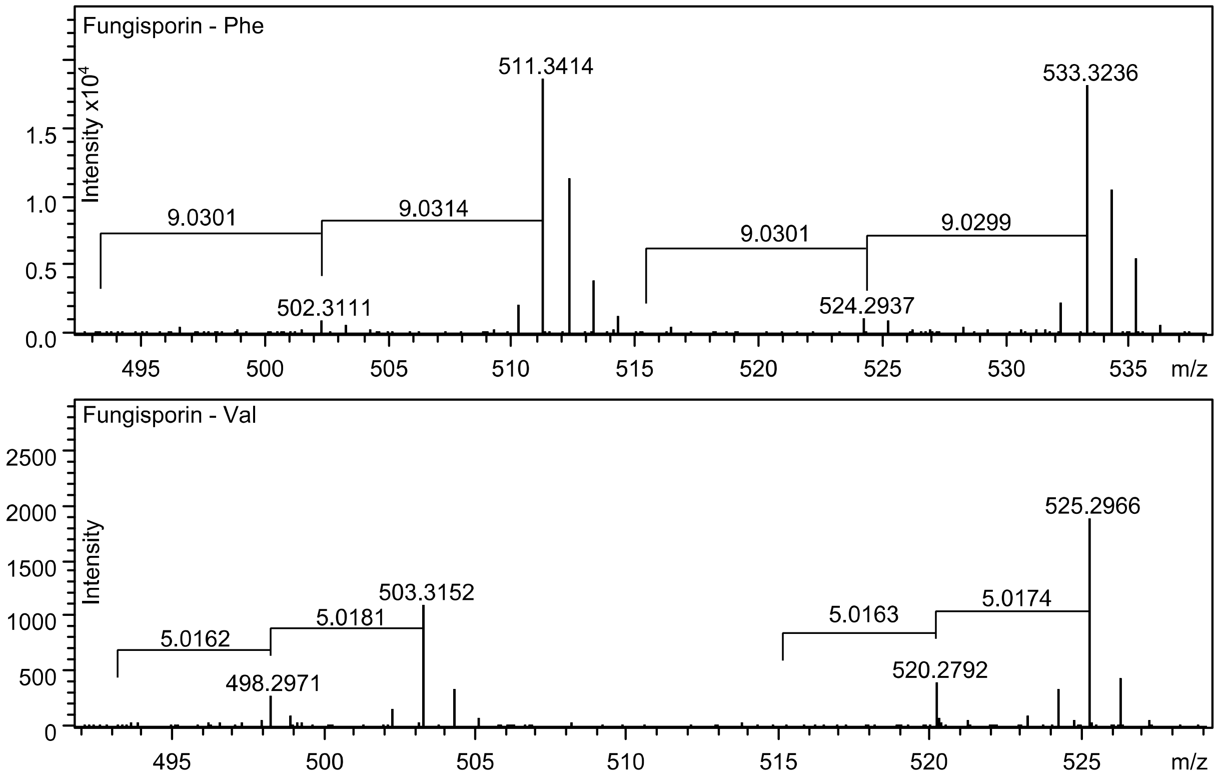
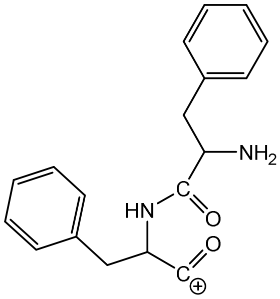
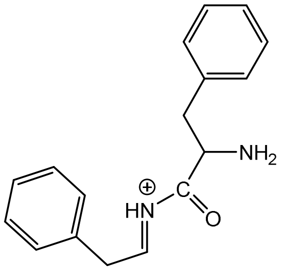
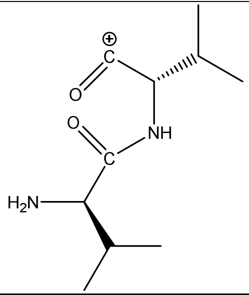
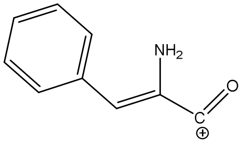
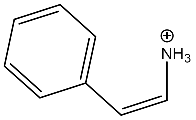


Figure S9. Labeling of fungisporin. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.34-7.40 min and have been scale to the highest signal.

Table S2. Fragment ions formed by fragmentation of nidulanin fungisporin A

Fragment [*m/z*]	Chemical formula	Structure
295	$C_{18}H_{18}N_2O_2$	
267	$C_{17}H_{18}N_2O$	
199	$C_{10}H_{19}N_2O_2$	
171	$C_9H_{18}N_2O$	
120	$C_8H_{10}N$	

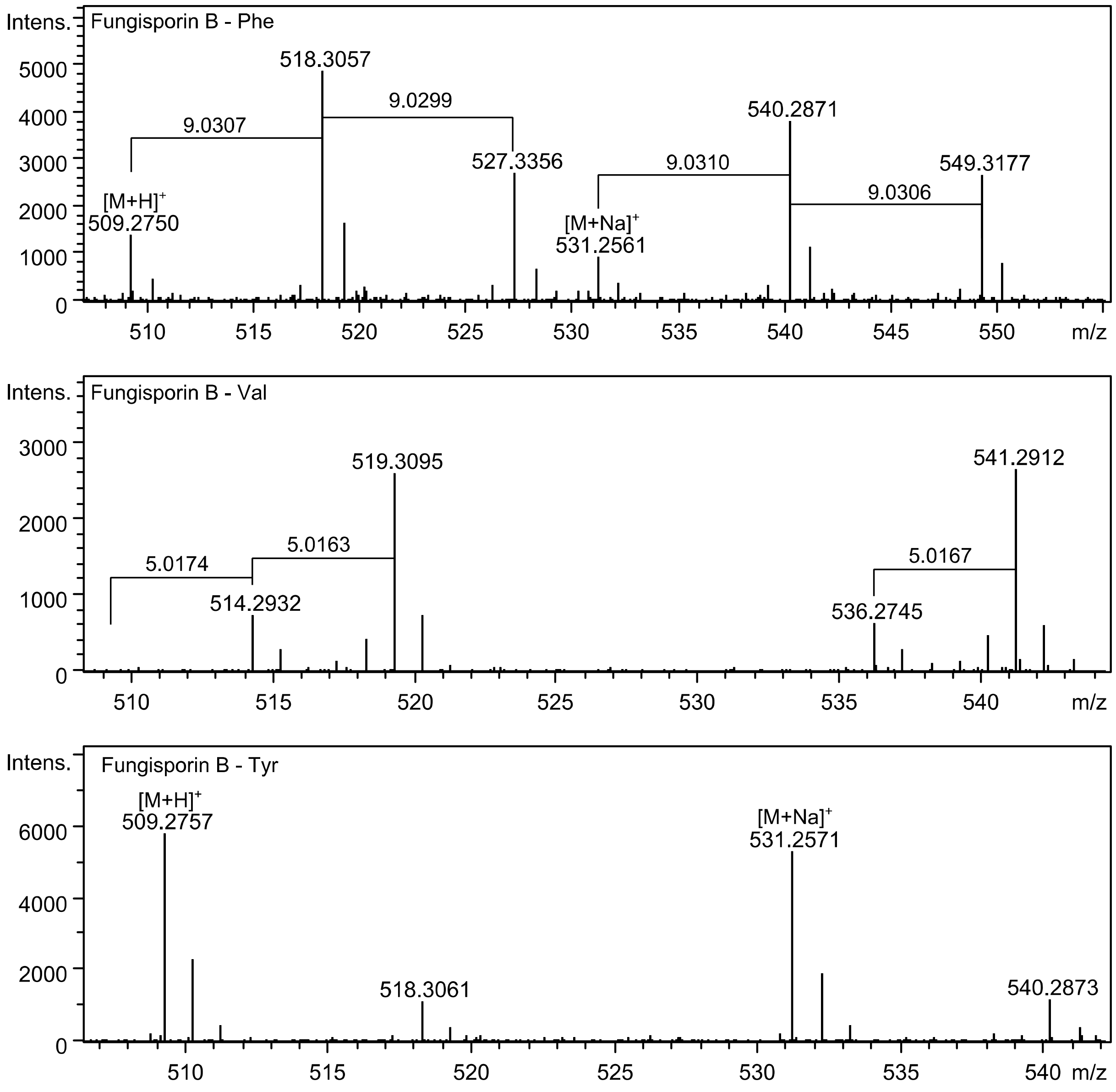


Figure S10. Labeling of fungisporin B. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 6.25-6.29 min and have been scale to the highest signal.

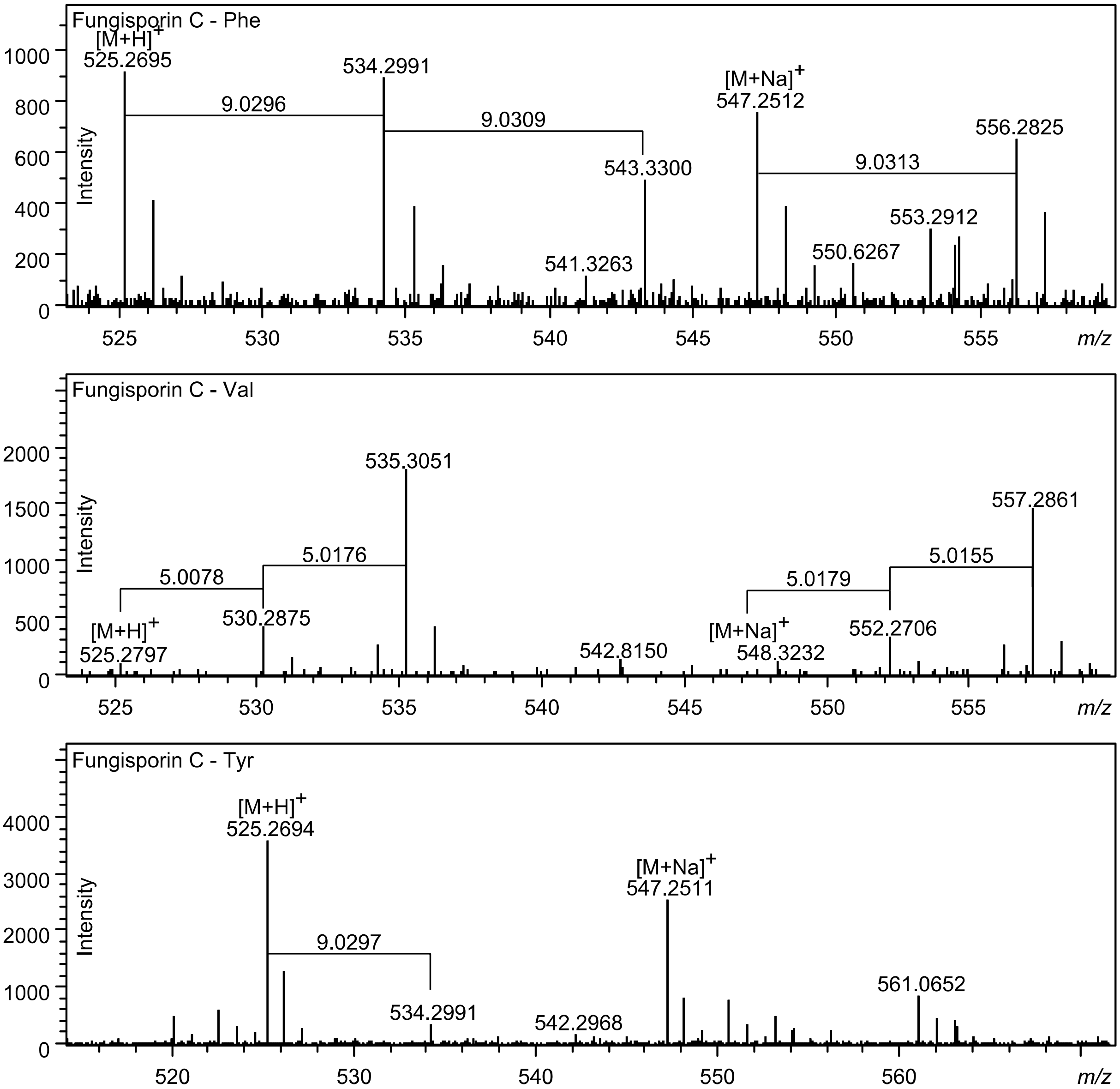


Figure S11. Labeling of fungisporin C. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 5.40-5.45 min and have been scale to the highest signal.

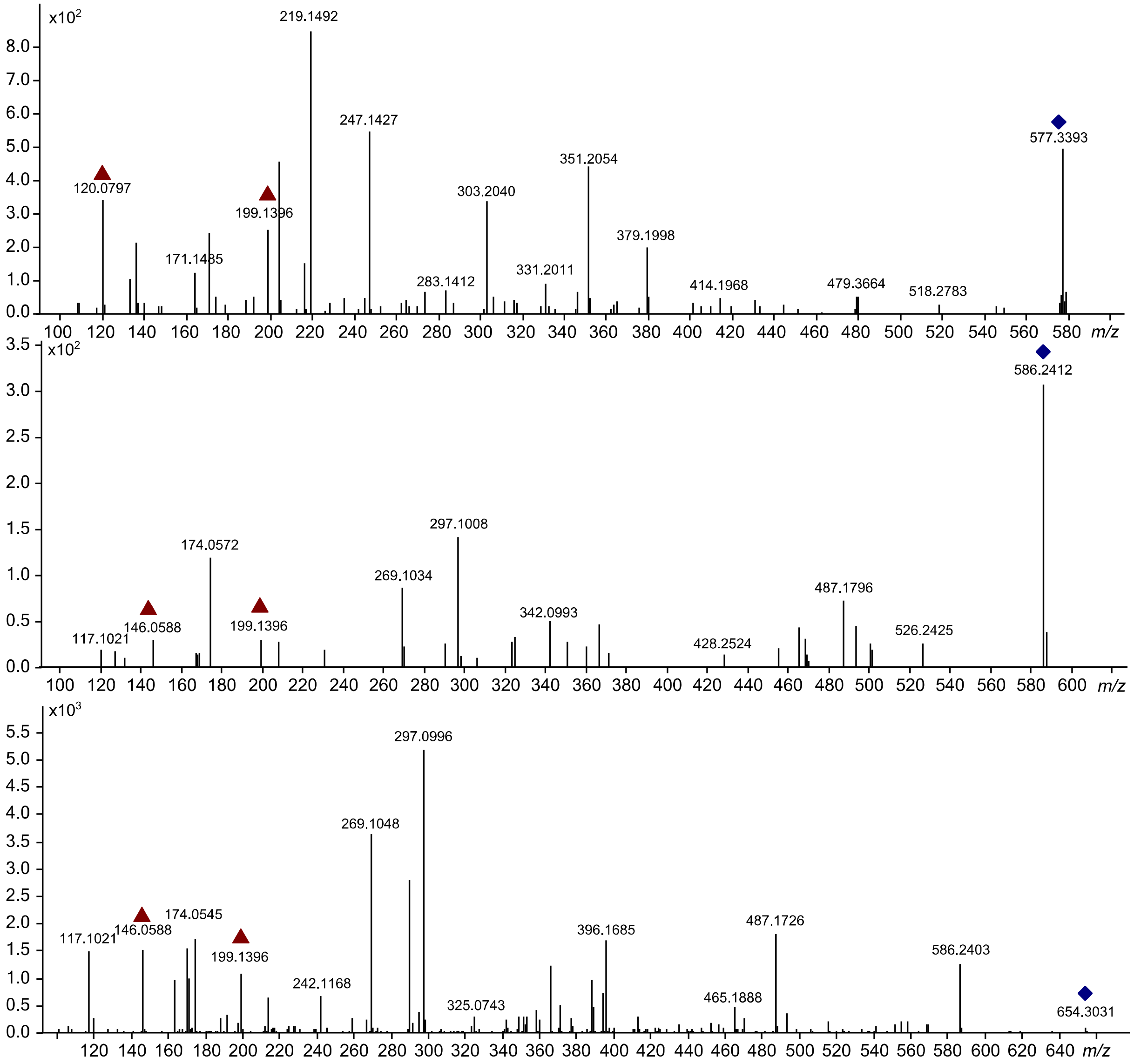


Figure S12 MS/MS spectra obtained from analysis of three unknown. The blue diamonds denote to the product ion of the compounds while the red triangles denote fragments formed by the unlabeled NA. The MS/MS obtained from fragmentation of the ion 654 exhibits many of the same ions as the one obtained from 586. The mass difference between the two ions indicate that they could be a prenylated and unprenylated form of the same compound.

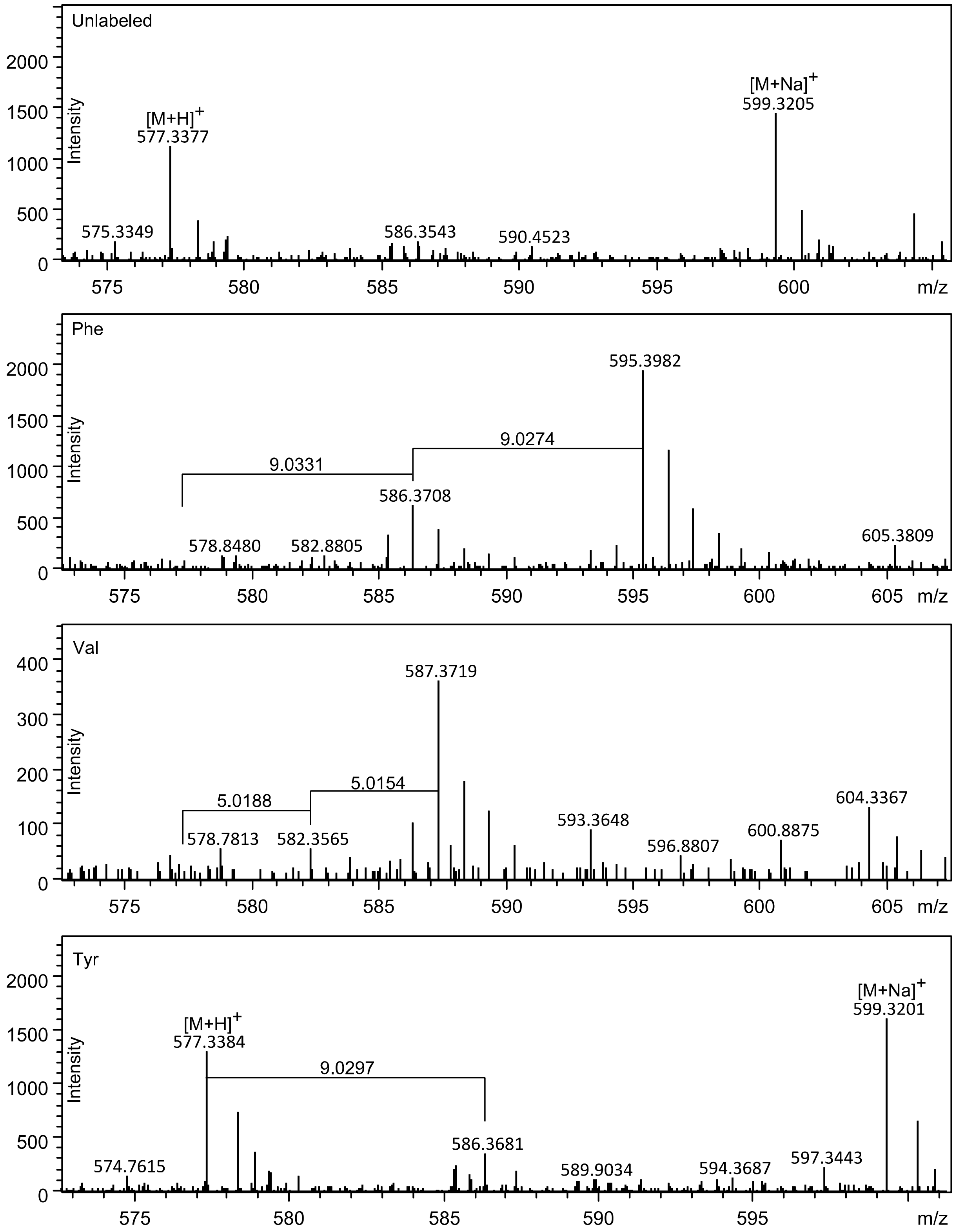


Figure S13. Labeling of new compound with the molecular formula $C_{33}H_{45}N_4O_5$. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.3-7.4 min and have been scale to the highest signal.

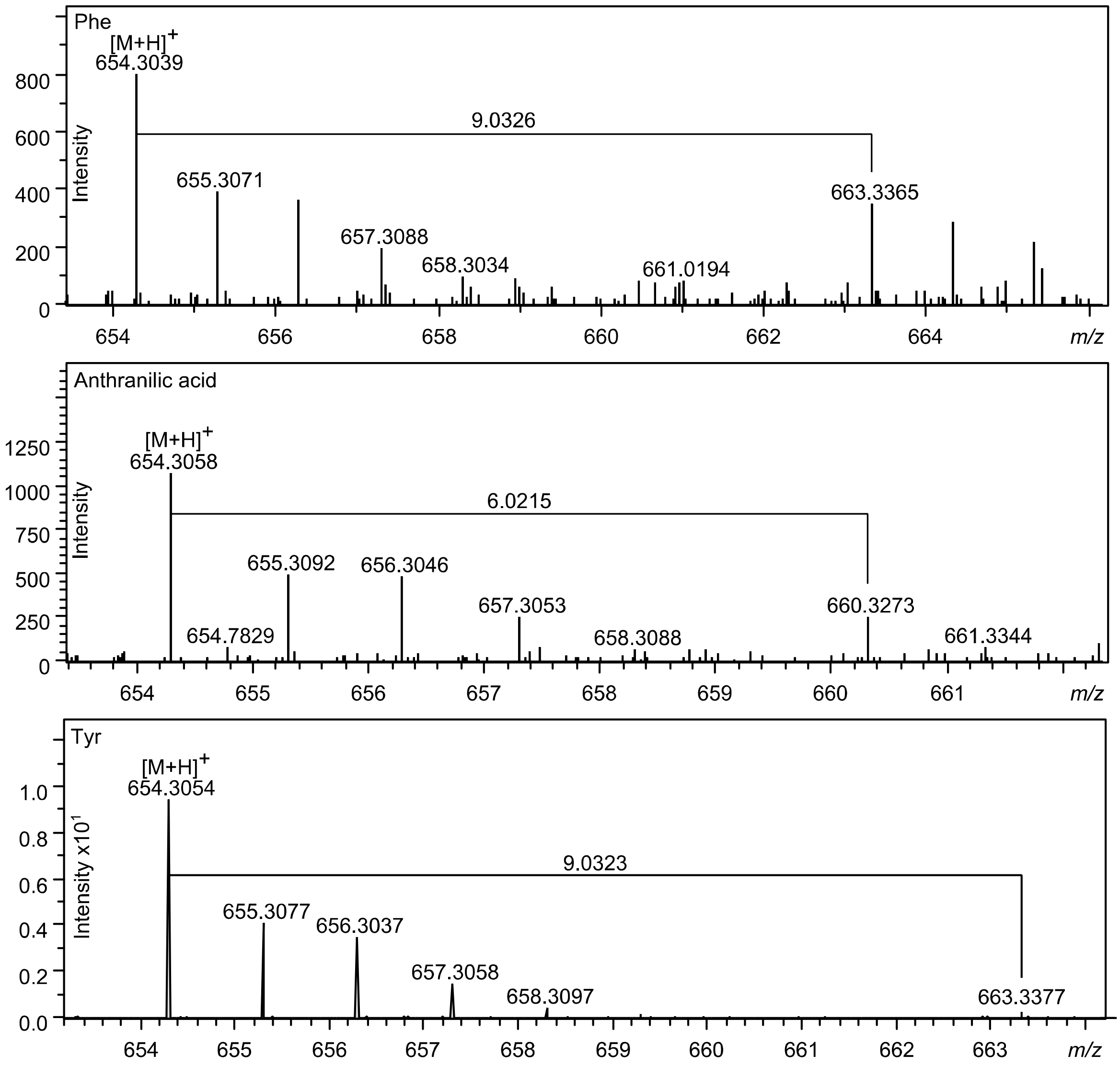


Figure S14. Labeling of new compound with the molecular formula $C_{34}H_{40}N_5O_7$. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.93-7.99 min and have been scale to the highest signal.

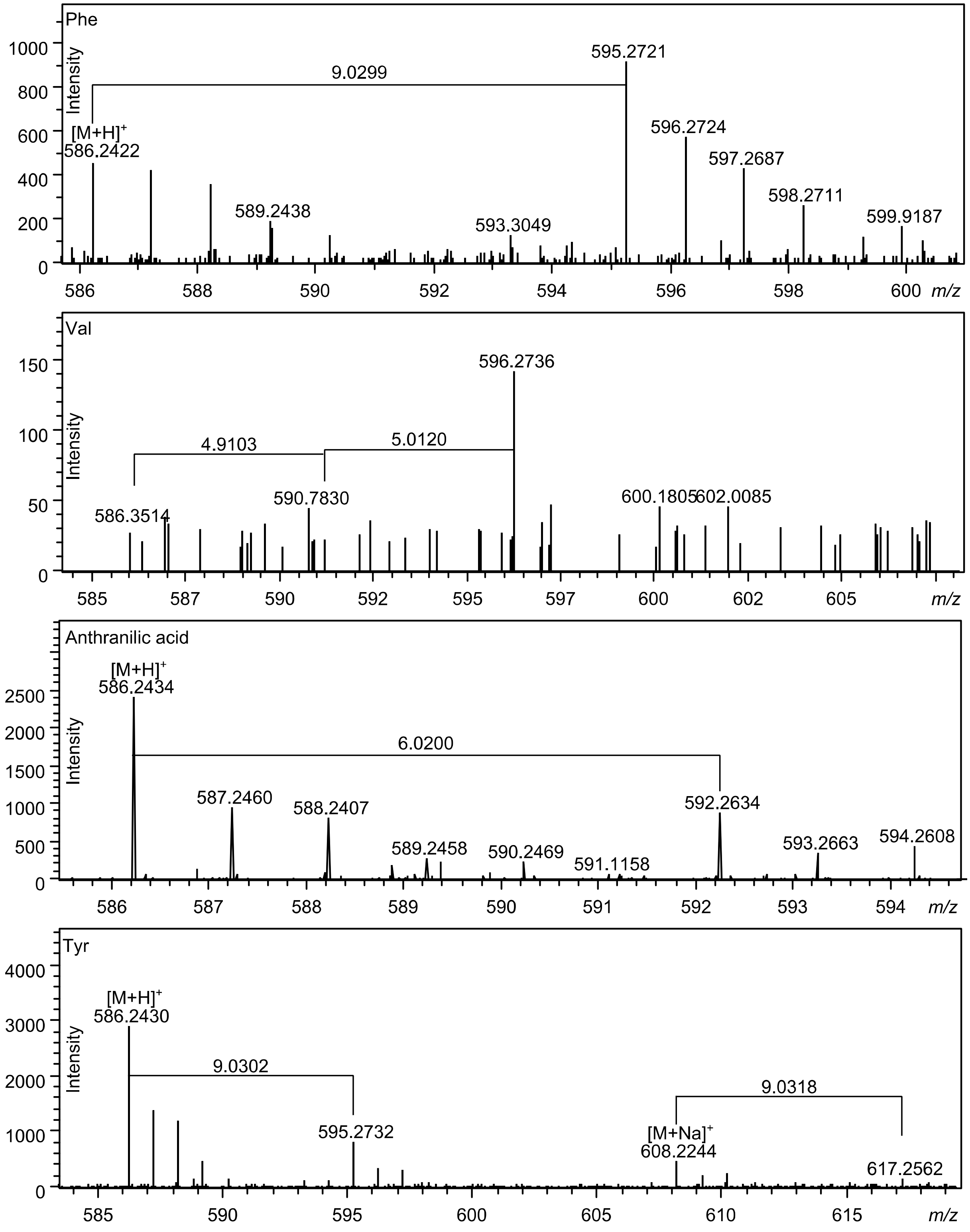


Figure S15. Labeling of new compound with the molecular formula $C_{35}H_{32}N_5O_4$. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 6.50-7.00 min and have been scale to the highest signal.

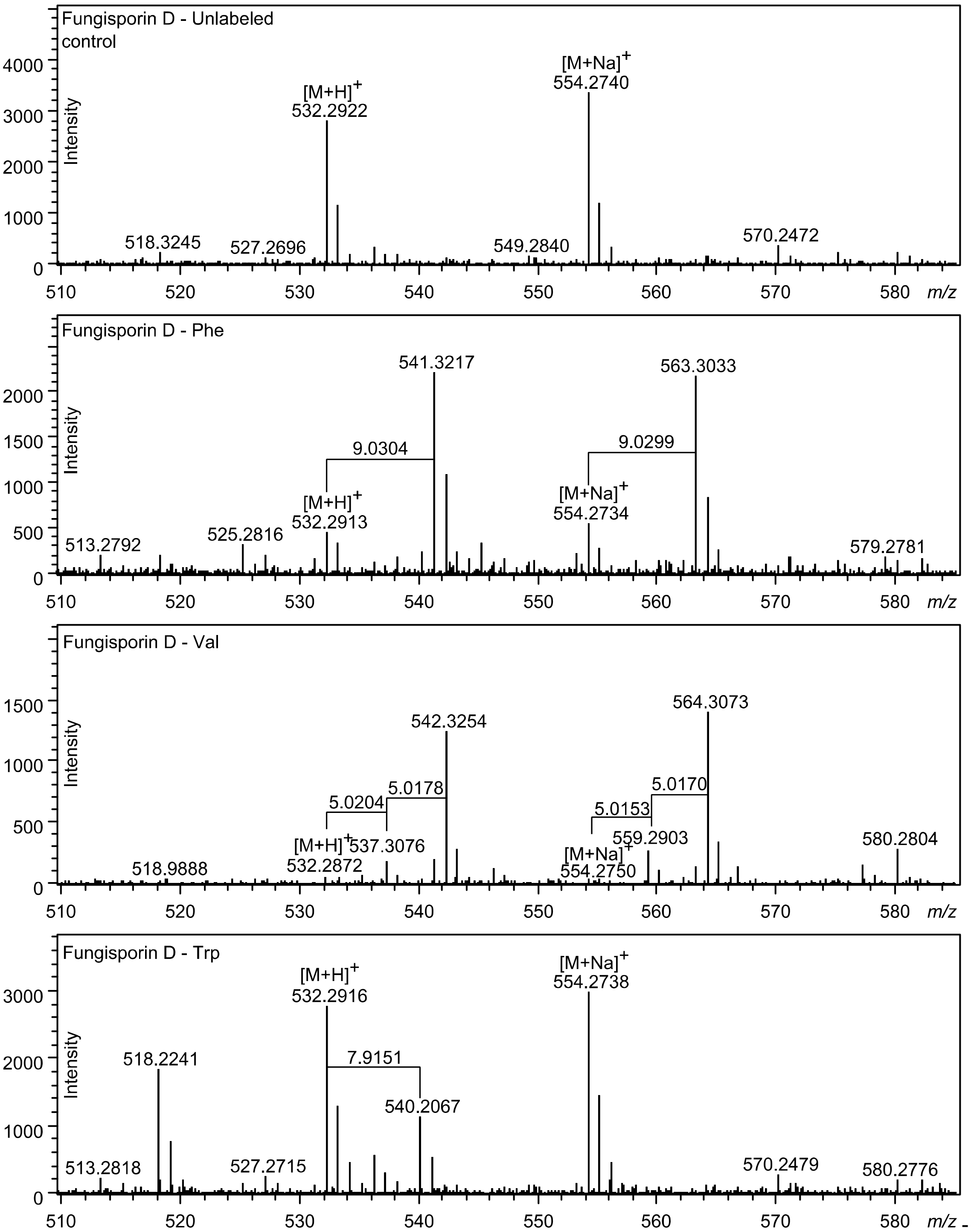


Figure S16. Labeling of fungisporin D. The mass spectra are from *A. nidulans* IBT 4887, cultivated at 25 °C in darkness for 7 days on MM. The mass spectra were extracted at RT 7.15-7.20 min and have been scale to the highest signal.

6.7 Paper 7 – Integrated Metabolomics and Genomic Mining of the Biosynthetic Potential of the Marine Bacterial *Pseudoalteromonas luteoviolacea species*

Maansson, M., Vynne, N. G., Klitgaard, A., Nybo, J. L., Melchiorsen, J., Ziemert, N., Dorrestein, P. C., Andersen, M. R., & Gram, L.

Draft (2014)

Integrated Metabolomic and Genomic Mining of the Biosynthetic Potential of Bacteria

Short title (50 characters): *Integrated Metabolomic and Genomic Mining*

Maria Maansson[1,a], Nikolaj G.Vynne[1], Andreas Klitgaard[1], Jane L. Nybo[1], Jette Melchiorsen[1], Nadine Ziemert[2], Pieter C. Dorrestein[2,3,4], Mikael R. Andersen[1], and Lone Gram[1,b]

[1]Department of Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[2]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093

[3]Departments of Pharmacology and Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093

[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA 92093

[a]Present address: Chr. Hansen A/S, Boege Allé 10-12, DK-2970 Hoersholm, Denmark

[b]Author to whom correspondence should be addressed. Email: gram@bio.dtu.dk

Editorial Board Members (must suggest three): Peter Greenberg (microbial), Jerrold Meinwald (chemistry), Ed de Long (ecology)/John Coffin (comparative genomics)

NAS members (must suggest three): Julian Davies (small molecules, bacteria), Jody Deming (marine microbiologist), Fred W. McLafferty (chemistry, MS)

Reviewers (must suggest five): Staffan Kjelleberg (UNSW), Tilmann Harder (UNSW), Rolf Müller (Department of Pharmaceutical Biotechnology, Saarland University) ……………

Classification: BIOLOGICAL SCIENCES (Microbiology)

Keywords: comparative genomics, untargeted metabolomics, natural products

Abstract (250 words)

There is an urgent need for novel bioactive compounds for control of both acute and chronic diseases. Microorganisms are a rich source of bioactives; however, chemical identification is a major bottleneck. Thus, strategies that can prioritize the most prolific microbial strains and attractive compounds are of highest interest. In this study, we present an integrated approach to evaluate the biosynthetic richness in bacteria and mine the associated chemical diversity. As an example, we subjected 13 strains of *Pseudoalteromonas luteoviolacea* isolated from around the globe to an untargeted metabolomics experiment. The results were correlated to whole-genome sequences of the strains. We found that 30% of all chemical features and 24% of the biosynthetic genes were unique to a single strain, while only 2% of the features and 7% of the genes were shared between all. The list of chemical features was reduced to 50 discriminating features using a genetic algorithm and support vector machines. Features were dereplicated by MS/MS networking to identify molecular families of the same biosynthetic origin, and the associated pathways were probed using comparative genomics. Interestingly, most of the discriminating features were related to antibacterial compounds, including the thiomarinols that were reported from *P. luteoviolacea* here for the first time. Additionally, we used comparative genomics to identify the biosynthetic cluster responsible for the production of the antibiotic indolmycin, a cluster that could not be predicted by antiSMASH. In conclusion, we present an integrative strategy for elucidating the chemical spectrum of a bacterium and link it to biosynthetic genes.

Significance Statement (96 words, understandable to general public)

To optimize our search for novel bioactive compounds useful in disease treatment, we here combine untargeted metabolomics and comparative genomics to probe for new bioactive secondary metabolites based on their pattern of distribution. We demonstrate the usefulness of this combined approach in the marine Gram-negative bacterium *Pseudoalteromonas luteoviolacea*, which is a chemically and genetically diverse species. The approach allowed us to identify new antibiotics and their associated biosynthetic pathways. Combining metabolomics and genomics is an efficient mining approach for chemical diversity in a broad range of microorganisms that are prolific producers of secondary metabolites.

Author contributions. M.M., N.G.V. and L.G. designed the research; M.M., N.G.V., and J.M. carried out the experiments; M.M., A.K., N.G.V., N.Z., and M.R.A. analyzed the data; J.L.N., M.R.A., and P.C.D. provided methods and algorithms; M.M., A.K., M.R.A., and L.G. wrote the paper.

Introduction

Microorganisms have remarkable biosynthetic capabilities and can produce secondary metabolites with high structural complexity and important biological activities. Microorganisms have especially been a rich source of antibiotics (1, 2); however, with the rapid spread of antibiotic resistance in human and animal pathogens, there is an urgent need for finding and identifying novel bioactive metabolites. Chemical identification of microbial metabolites is a major bottleneck, and tools that can help prioritize the most prolific microbial strains and attractive compounds are of highest interest.

The search for novel chemical diversity can be done 'upstream', at the genome level, or 'downstream', at the metabolite level. While the historical approach has been downstream identification of target molecules, searching upstream has become highly attractive with the availability of full genome sequences at a low cost (3–6). The analyses are greatly aided by several *in silico* prediction tools (7), including antiSMASH (8, 9) and NaPDoS (10) for secondary metabolite pathway identification. Several studies have explored the general genomic capabilities within a group of related bacteria (11–16), but only few studies have explored the overall biosynthetic potential and pathway diversity (17–19). Ziemert *et al.* (18) compared 75 genomes from three closely related *Salinispora* species and predicted 124 distinct biosynthetic pathways, which by far exceeds the currently 13 known compound classes from these bacteria. The study underlined the discovery potential in looking at multiple strains within a limited phylogenetic space, as a third of the predicted pathways were found only in a single strain.

A large potential is found in combining the upstream approach with the significant advances in analytical methods for downstream approaches. Building on the versatility, accuracy, and high sensitivity that the LC-MS platform has achieved, sophisticated algorithms and software suites have been developed for untargeted metabolomics (20–24). The core of these programs is the feature detection (or *peak picking*), i.e. the identification of all signals caused by true ions (25), and peak alignment, the matching of identical features across a batch of samples. Today, many programs consider not only the parent mass and the retention time, but also the isotopic pattern, ion adducts, charge states, and potential fragments (25) which greatly improves the confidence in these feature detection algorithms (26). This high-quality data can be combined with multivariate analysis tools, which not only aids analysis and interpretation, but also form a perfect basis for integration with genomic information. Recently, molecular networking has been introduced as a powerful tool in small molecule genome mining (27, 28). It builds on an algorithm (29, 30) capable of comparing characteristic fragmentation patterns, thus highlighting molecular families with the same structural features and potentially same biosynthetic origin. This enables the study and comparison of a high number of samples, at the same time aiding dereplication and tentative structural identification or classification (31).

Here, we present an integrated diversity mining approach that links genes, pathways, and chemical features at the very first stage of the discovery process using a combination of publically available prediction tools and machine learning algorithms. We use genomic data to interrogate the chemical data and vice versa in order to quickly get an overview of the biosynthetic capabilities of a group of related organisms and identify unique strains and compounds suitable for further chemical characterization. We demonstrate our approach on a unique group of organisms that is strains of the marine bacterial species *Pseudoalteromonas luteoviolacea* (32, 33). Previous studies in our lab have shown that it is a highly chemically prolific and diverse species with strains producing an antibiotic cocktail of violacein and either pentabromopseudilin or indolmycin (34). We use the integrated approach to evaluate the promise of continued sampling and discovery efforts within this species as demonstrated by the finding of an additional group of antibiotics that is the thiomarinols.

Results

The secondary metabolome and genome of P. luteoviolacea is dominated by unique features

A total of 13 strains of *P. luteoviolacea* were analyzed for their genomic potential and ability to produce secondary metabolites. To obtain a global, unbiased view of the metabolites produced, molecular features were detected by LC-ESI-HRMS in an untargeted metabolomics experiment. On average, more than ~2,000 molecular features were detected in each strain. Merging of ESI$^+$/ESI$^-$ data resulted in a total of 7,190 features from the 13 strains (excluding media components), with more features detected in positive mode (6,736) as compared to negative mode (2,151). To facilitate comparison to genomic data, the features were represented as pan- and core plots commonly used for comparative microbial genomics (35, 36). Here, core metabolome features are shared between all strains, while the pan-metabolome represents the total repertoire of features detected within the collection (Fig 1A). Surprisingly, only 2% of the features were shared between all the strains. In contrast, 30% of all features were unique to single strains. As the number and detection of features in each strain change with the chosen threshold for feature filtering, the pan- and core plots were also made based on the 2,000 and 500 most intense features (Fig. S1). Here, the same trend was observed with 6-10% core features and 20% unique features. Thus, regardless of feature filtering settings, the overall pattern of diversity is the same.

To link the chemical diversity to the genomic diversity, we analyzed the genomes of the 13 strains. The average genome size was around 6 Mb with approximately 5,100 putative protein encoding genes per strain (Table S1). The corresponding pan- and core genomic analysis was performed according to Vesth *et al* (36) (Fig. 1B). A total of 9,979 protein encoding genes were predicted in the pan-genome including 3,322 genes (33%) conserved between all strains, thus on average, the core genome constituted ~65% for each strain. Of the accessory genome, 23% of the total genes (2,329) could only be found in a single strain (singletons/unique genes). Considering only genes predicted to be involved in secondary metabolism, the diversity was even higher (Fig. 1C). On average, 8.6% of the total genes were predicted to be allocated to secondary metabolism (Table S1), which is extremely high compared to other sequenced strains belonging to *Pseudoalteromonas* (37, 38). Similar to the total pan-genome, 24% (386) of the genes putatively involved in secondary metabolism were found in only a single strain; however, only 7% (119) were shared between all 13 strains. Thus, we see approximately a 5-fold higher genetic diversity in secondary metabolism as compared to the full pan-genome.

The high number of unique genes and molecular features, suggest an *open* pan-genome/metabolome (35), in which there is a continuous increase in diversity with continued sampling, which is very attractive for discovery purposes. Both set of data suggest, that 90% of the diversity/genomic potential for secondary metabolism can be covered with 10 strains, but that each new strain holds promise for new compounds and biosynthetic pathways.

Pan-genomic diversity and pathway mapping suggest a highly dynamic accessory genome

To get an overview of the potential evolutionary relationship between the strains and associated pathways, a pan-genomic map was generated illustrating shared orthologs between groups of species (Fig. 2). The method uses a conservative BLAST-based non-greedy pairing of genes, which results in 2,435 genes found to be present as 1:1 orthologs in all strains, which is slightly less than the 3,388 genes found in the method illustrated in Figure 1. In general, we observed two main clades (A and B) based on shared genes, one consisting of six strains and the other of seven. Each clade has 190-220 genes unique for that clade. The method also further reflects the genetic diversity of each strain, as illustrated in Figure 1B-C. Based on this, we generated presence/absence patterns for all genes showing in which other strains that gene has orthologs, a useful starting point for data correlation.

For genetic analysis of biosynthetic pathways in multiple strains, pathways predicted by antiSMASH across the 13 strains were grouped into 37 operational biosynthetic units (OBUs) (18) (Table S2). OBU presences were compared to the pan-genomic map (Fig. 2) to trace biosynthetic pathways. Only ten pathways were conserved in all strains, including a glycosylated lantipeptide (*ripp 1*) and two bacteriocins (*ripp 2* and *ripp 3*). All strains maintained essential pathways likely responsible for production of siderophores (NRPS1 putative catechol-based siderophore) and homoserine lactones (different variations). The violacein pathway *vio* is also conserved in all strains, in addition to an unassigned type III PKS and a hybrid NRPS-PKS pathway. Interestingly, the majority of clusters follow the linearity of Figure 2, suggesting that many of the pathways have been introduced and retained based on a competitive advantage of those clusters. More than 50% of the predicted pathways are restricted to one or two strains, suggesting that many pathways are introduced highly dynamically (in evolutionary scale) and through horizontal gene transfer.

Feature prioritization and dereplication of the pan-metabolome by support vector machine and molecular networking reveals key discriminative metabolites

To explore the diversity within the pan-metabolome and prioritize chemical features for more detailed structural analysis, a two-pronged approach was used: multivariate analysis based on machine learning algorithms and comparative analyses based on the pattern of conservation generated from the pan-genomic diversity map. A classifier based on a combination of a genetic algorithm (GA) and support vector machine (SVM) (39, 40) was used as a feature selection method to filter the most important features from the complex data set, starting with the 500 most intense features and reducing it to the 50 most significant features to distinguish all 13 strains (Table S3). In addition, extracts from all strains were analyzed with LC-ESI-MS/MS to generate a molecular network (Fig. S2 for full figure) (28). The candidates identified by multivariate and comparative analyses were correlated to the molecular network (27, 31) for dereplication and connection of molecular features that likely belong to the same structural class and thus biosynthetic pathway. For example, the *vio* pathway (41) was found in all 13 strains, and the antibiotic violacein was a~~ a~~ discriminating core feature (Table S3). In the molecular network, violacein was found to belong to a molecular family of minimum five related analogues (Fig. S3) likely associated with the *vio* pathway, including proviolacein, and oxyviolacein as well as a novel analogue with two extra hydroxyl groups (Fig. S3).

Some **P. luteoviolacea** *strains have lost the ability to produce polyhalogenated compounds*

The discriminating features do not necessarily reflect the same groupings as the genomic analyses. Therefore, they can be used as a tag for identifying the corresponding biosynthetic pathway through correlation with genomic presence/absence patterns. On the list of descriptive features generated using the SVM (Table S3), there are six highly halogenated features that all seem to be restricted to seven strains: CPMOR-2/DSM6061(T), S2607/S4060-1, NCIMB1944/2ta16, and CPMOR-1. To investigate whether halogenation in general is unique to those strains, a list of features with high mass defect was made, resulting in more than 40 halogenated compounds (Table S4) restricted to the seven strains. Most of them had no match to known compounds, but many match the structural scaffolds of poly-halogenated phenols and pyrrols or hybrids hereof (42) and have expected antibacterial activity (43).

No pathway predicted by antiSMASH had a halogenase incorporated, thus the pattern of presence in these seven strains was used to probe for associated clusters. Indeed, we found an intact group of 11 genes (including two brominases) conserved in the seven aforementioned strains (Fig. S4). The recently characterized *bmp* pathway correspond to ten of these genes (*bmp1-10*) (42) which is responsible for the production of poly-brominated phenols/pyrrols in strain 2ta16, with the 11th gene being a putative multidrug transporter possibly conferring resistance (putatively assigned *bmp11*), an activity not described in *bmp1-10*. Surprisingly, all 11 genes were also found in NCIMB1942/NCIMB2035 where no halogenated compounds were detected. However, in the latter strains, the gene cluster is broken, with four genes located elsewhere in the genome, providing a plausible explanation for the lack of halogenated compounds. Also, *bmp1*, *bmp2*, and *bmp7-10* were found in S4047-1/S4054, which suggest that a common ancestor had an intact *bmp* pathway.

Two of the discriminative features found in the seven strains are two isomeric dimeric bromophenol-bromopyrrole hybrids with eight bromines in total (Fig. S5). The monomers corresponding to the likewise novel ‘tetrabromopseudilin’ is also found in the extract, suggesting that these ‘bis-tetrabromopseudilin’ are true compounds rather than artefacts arising from MS insource chemistry. Full structural characterization of these low proton density compounds lies beyond the scope of this study, but underlines the versatility of the *bmp* pathway and associated chemical diversity.

Identification of the indolmycin cluster shows resistance genes and potential QS control

Strains S4047-1, S4054, and CPMOR-1 are all producing the antibiotic indolmycin, as previously reported (34). Indolmycin was identified by GA/SVM as a discriminating feature for those three strains. In addition to indolmycin, the molecular family consisted of the N/C-demethyl- and N/C-didemethyl indolmycin analogues as well as indolmyceinic acid, a methylated and two hydroxylated analogues. Most of these analogues have not been reported from microbial sources and their tentative structures were verified by their MS/MS fragmentation pattern (Fig. S6).

Like violacein, indolmycin is derived from L-tryptophan, but even though the biosynthetic pathway has been described by feeding studies in *Streptomyces* (44–46), the biosynthetic cluster has never been characterized. The pan-genome was probed for genes with presence/absence patterns matching the distribution of indolmycin and the related analogues, which led to the identification of 11 clustered genes, suggesting these to be the genetic basis for indolmycin biosynthesis (Fig. 3). The identified genes had predicted functions to those expected to be required for the synthesis of indolmycin such as an aromatic aminotransferase (*unk3*), aldoketomutase (*unk4*), SAM methyltransferase (*unk5*), and aminotransferase (*unkX*). Indolmycin has been identified as a competitive inhibitor of bacterial tryptophanyl-tRNA synthetases (47, 48), and the putative cluster seems to incorporate a tryptophanyl tRNA synthase (*unk2*), which in *Streptomyces griseus* has been found to confer resistance to indolmycin (48). Interestingly, the cluster is flanked by *luxI* and *luxR* homologues, suggesting that the indolmycin pathway potentially could be under QS regulation.

Thiomarinols add to the antibiotic cocktail

The strains 2ta16/NCIMB1944 were identified as hotspots for biosynthetic diversity based on Figure 2. This was supported by 313 chemical features unique to these two strains. Based on the GA/SVM, they can be distinguished from the rest of the strains based on a feature with *m/z* 640 RT 9.73 min ($C_{30}H_{44}N_2O_9S_2$), tentatively identified as thiomarinol A. Thiomarinols are hybrid NRPS-PKS compounds based on pseudomonic acid and pyrrothine. One of the gene clusters (hybrid NRPSPKS5) restricted to the pair of 2ta16/NCIMB1944 was found to have high similarity to that of pseudomonic acid (*mup*) (49) and the recently characterized thiomarinol (*tml*) cluster (50), corroborating the finding of the compound class. Thiomarinols have previously reported antibacterial activities from *Pseudoalteromonas* sp SANK 73390 (51, 52).

In the molecular network, it was possible to identify a whole series of thiomarinol and pseudomonic acid analogues (Fig. 4A+D), all restricted to NCIMB1944 and 2ta16. In addition to thiomarinol A-D, pseudomonic acid C amide and its hydroxyl-analogue could be assigned based on the characteristic MS/MS fragmentation pattern (Fig. 4B+C). Besides the known analogues, two novel analogues with formulas $C_{25}H_{43}NO_8$ and $C_{34}H_{51}NO_{11}$ could be identified. Both shared the marinolic acid moiety based on the $C_6H_6O_2$ (*m/z* 110.0368) fragment and the loss of $C_{11}H_{20}O_4$ (*m/z* 216.1362); however, they contained only a single nitrogen and no sulfur, indicating a completely new type of thiomarinol based on neither a holothine nor ornithine 'head' like the known analogues (Fig. 4C).

Discussion

Advances in genomics and metabolomics have significantly increased our ability to generate high-quality data on microbial secondary metabolism at a very high speed. This, in turn, has enabled a completely new approach to drug discovery combining the two 'omics approaches.

Using a combination of comparative metabolomics and genomics, we find a high potential and remarkable intra-species diversity in terms of secondary metabolite production for *P. luteoviolacea*. Overall, 8.6% of the genes are allocated to secondary metabolism and on average 10 NRPS/PKS related OBUs are predicted. This is very high considering the relatively small size of the genomes (~6 Mb) and is comparable to that of recognized prolific species such as *Salinospora arenicola* (10.9% of 5.8 Mb)(13, 18, 53) and *Streptomyces coelicolor* (8% of 8.7 Mb) (54). Our data suggest an open pan-genome which is characteristic for species that are adapted to several types of environments (55), i.e. being both planktonic and associated with marine macro-algal surfaces. The pan-genome is a dynamic descriptor that will change with the number of strains and the specific subset. Nonetheless, our findings correlate with comparative genomic studies of other bacterial species (11, 12, 14, 55). .

We found ~5-fold higher genetic diversity in secondary metabolism compared to the full pan-genome which supports that production of secondary metabolites is a functionally adaptive trait (56, 57). More than half of the 41 predicted pathways are restricted to one or two strains, while only ten pathways were shared between all. This is similar to findings in *Salinispora* (18), where 78% of the pan-genome is associated with one or two strains. Violacein (58, 59), indolmyin (60, 61), and pentabromopseudilin (42) are all examples of cosmopolitan antibiotics found in unrelated species, thus, we hypothesize that *P. luteoviolacea* acquired and retained biosynthetic genes linked to e.g. antibiotic production as part of adapting to a specific niche that it commonly occupies.

Diversity is further supported at the chemical level: Using unbiased global metabolite profiling, we identify >7,000 putative chemical features among the 13 analyzed strains. As the number of chemical features depends on the filtering threshold, this should not be seen as an absolute number of compounds that can be isolated and fully characterized. However, it provides an unbiased estimate of diversity, which in this case does not seem to change with the chosen threshold. Surprisingly, only 2% of the features were shared between all the strains. To the best of our knowledge, there is only one other study in intra-species chemical diversity. Krug *et al* (19, 62) analyzed 98 isolates of *Myxococcus xanthus* in a semi-targeted approach and found 11 out of 51 identified compounds to be shared between all strains and a similar fraction present in only one or two strains. We find almost half of all features and one third of the 500 most intense features could be assigned to one or two strains (thus taking into account the almost clonal strains), which underlines a great potential for unique chemistry within a single species.

The remarkable chemical diversity can be found even within the same sample. Strains S4047, S4054, and S4060 that were all collected from seaweed from the same geographical location (2,9817, -86,6892). Strains S4047 and S4054 share 99% of their gene families (clonal) and 70% of their chemical features, but strain S4060 only share 24% of gene families and 30% of features with the other two. It is also reflected in the biosynthetic pathways, where nine pathways were found in S4060, but not in S4047 and S4054. This is a fascinating ecological conundrum as the accessory metabolites and genes usually are considered to answer the immediate, more localized needs for the strains. Nonetheless, this is not the first report of such an occurrence. Vos *et al.* (63) found 21 genotypes of *M. xanthus* using multilocus sequence typing among 78 strains collected from soil on a centimeter scale. Likewise, significant differences have been found in the chemical profiles of co-occurring strains of *M. xanthus* (19) and *Salinibacter ruber* (64). In contrast, NCIMB1944 and 2ta16 that originate from the Mediterranean Sea (France) and Florida Keys (US), respectively, share 99% of their gene families and 70% of their features. That demonstrates that genomic content can be relatively conserved across bio-geographical locations, suggesting a high selective pressure to conserve those genes despite an overall low degree of chemo-consistency.

In this study, SVM was applied in conjunction with GA to compile a list of 50 chemical features of interest for further structural characterization. Based on SVM, the reduced set of features are the ones that maximize the difference between samples, which in this study is exploited to select features unique to each strain or a subset of strains. GA works as a wrapper to select features to be evaluated in the SVM classifier (65). The intrinsic nature of the GA makes it highly suitable for discovery purposes as it favors diversity in how the subset of features is selected (40). To the best of our knowledge, there are only few examples on the use of SVM in untargeted secondary metabolite profiling (66, 67). The list of discriminating features highlights key metabolites, both in the core- and accessory metabolome. Of the 50 discriminating features, only 15 could be tentatively assigned to known compound classes. In this specific case, the list even reflects the four antibiotic classes identified in this species, underlining the utility of GA/SVM to prioritize not only strains but also compounds before the rate-limiting step of structural identification. The combination with molecular networking further strengthen this approach as it makes it possible to identify structural analogues that likely have similar biological activity.

To the best of our knowledge, this is the first example of direct coupling of genomic and metabolomic data at a global level and at this early stage of the discovery process. By solely using the patterns of presence/absence across the pan-genome in conjunction with synteny, we could identify gene clusters without relying on the functions. This allowed for the identification of the pentabromopseudilin and indolmycin gene clusters. Combined with presence/absence of molecular features, this is an extremely powerful tool for translation back and forth between genome and metabolome. Thus, it is possible to identify specific compounds using genomic queries or to specifically identify a gene cluster based on chemistry. Of course, in order to fully confirm the link between compound and genes, knock-out mutants need to be analyzed, but here, single candidates for clusters could be directly and rapidly identified.

The combination of metabolomics and genomic data identifies obvious hotspots for chemical diversity among the 13 strains, which permit intelligent strain selection for more detailed chemical analyses. By randomly picking a single strain, worst case, only 38% of the 500 most intense chemical features (and thus most relevant from a drug discovery perspective) are covered (NCIMB2035). However, when maximizing strain orthogonality by selecting the two strains (NCIMB1944 + CPMOR-1) with the highest number of unique genes, pathways, and chemical features, 82% of the diversity can be covered. This is extremely important as the isolation and full structural characterization of these compounds still represent the greatest bottleneck in the discovery process. This study shows that investigation of multiple strains of the same species can be a valuable strategy for detection of new compounds and is imperative to uncovering the full biosynthetic potential of a species.

Material and Methods

Strains, cultivation, and sample preparation for chemical analyses. The 13 strains of *P. luteoviolaceae* included in the study were collected or donated to us as previously described (34, 68). We did attempt to build a larger collection; however, *P. luteoviolaceae* autolyses very easily and in most laboratories it has not been possible to store and revive strains. The strains were cultured in biological duplicates in Marine Broth (MB, Difco 2216) at 25 ºC (200 rpm) for 48h before extraction. See details in SI.

LC-MS and LC-MS/MS data acquisition. LC-MS and MS/MS analyses were performed on an Agilent 6550 iFunnel Q-TOF LC-MS (Agilent Technologies, Santa Clara, CA, US) coupled to an Agilent 1290 Infinity UHPLC system. Separation was performed using a Poroshell 120 phenyl-hexyl column (Agilent, 250 mm × 2.1 mm, 2.7 µm) with a water/ACN gradient and MS data recorded both in positive and negative electrospray (ESI) mode in the *m/z* 100-1,700 Da mass range. Data for molecular networking was collected using a data-dependent LC-MS/MS as reported previously (69) with optimized collision energies and scan speed. See SI for full experimental setup, procedures, and method parameters.

Feature extraction and multivariate analysis. Extraction of chemical features was performed using MassHunter (Agilent Technologies, v. B06.00) and the Molecular Features Extraction (MFE) algorithm and recursive analysis workflow. Feature lists were imported to Genespring – Mass Profiler Professional (MPP) (Agilent Technologies, v. 12.6) and filtered with features resulting from the media removed. The feature lists from ESI$^+$ and ESI$^-$ were merged in a table as generic data and re-imported into MPP. The data was then normalized and aligned resulting in a single list of chemical features for each sample. The list of discriminating features was generated in MPP using genetic algorithm with a population size of 25, 10 generations, and a mutation rate of 1. The GA was evaluated using the SVM with a linear kernel type with and imposed cost of 100 and ratio of 1. The feature list was validated via the leave-one-out method. Further details and settings found in SI. All 50 discriminating features (Table SX) were manually verified to be present in the original datasets. Molecular formulas were predicted from the accurate mass of the molecular ion or related adducts (70) as well as the isotope pattern and matched against AntiMarin (v. 08.13) and Metlin (71) databases to tentatively assign known compounds.

Molecular networking. For molecular networking, raw LC-MS/MS data was converted to .mgf using MSConvert from the ProteoWizard project (72) and analyzed with the algorithm described in Watrous *et al.* (28). The data can be accessed here (provide the public link to MSV munber, make sure all the annotations and molecules discussed here are annotated there, the esiest way to do this is to create a network and then click on addto library in the network viewer). The network corresponding to a cosine value of more than 0.7 was visualized using Cytoscape 2.8.3 (73).

DNA extraction and sequencing. Cultures were grown in MB for xx days and genomic DNA isolated using either the JGI phenol-chloroform extraction protocol or the xxx kit [Jette for extraction protocol]. Library preparation and 150 base paired end sequencing was done at Beijing Genomics Institute (BGI) on the Illumina HiSeq 2000 system. At least 100-fold coverage was obtained for all genome sequences in this study. Genomes were assembled using CLC Genomic Workbench (v. 2.1/2.04) with default settings. All sequences have been deposited in GenBank and assigned the accession numbers provided in table SXX. The genome of strain 2ta16 was downloaded from GenBank.

Genome annotation and analysis. Contigs were analyzed using the CMG-biotools package as described Vesth *et al.* (36). Genes were predicted using Prodigal 2.00. Gene families were constructed by genome-wide and pairwise BLAST comparisons. Genes were considered part of the same gene family with a sequence identify >50% over at least 50% of the length of the longest gene.

A pan- and core-genome plot was constructed according to Friis *et al* [ref]. A pan-genomic dendrogram based on occurrences of gene families was used to sort input order by clustering prior to generating the plot (14).

Putative biosynthetic pathways were predicted from sequences (FASTA) with antiSMASH 2.0 (8, 9), with KS and C domains of PKS and NRPS predicted with NaPDoS (10) using default settings. Pathways were assessed to be similar OBUs when MultiGeneBlast (74) analyses revealed that 80% of the genes in the pathway are present with homologues that show at least 60% amino acid identity. For assessment and assembly of pathways split between different contigs, the sequences of homologues on the same contig were used as scaffold. MultiGeneBlast (74) was used for recursive OBU analysis across all 13 strains, thus proving pseudo-scaffolds for larger pathways, which in turn give higher confidence in the assignments. Partial pathways with the same pattern of conservation were combined in order to avoid overestimation of diversity.

Mapping of genes shared by groups of species. All predicted sets of protein sequences for the 13 strains were compared using the blastp function from the BLAST+ suite (75). These 169 whole-genome blast tables were analyzed to identify bi-directional best hits in all pairwise comparisons. Using custom Python-scripts, this output was analyzed to identify, for all proteins, in which strains orthologs were found. This allowed identification of unique genes, genes shared by clades and sub-clades of species, and genes shared by all 13 strains of *Pseudoalteromonas.* The script also generates a binary 13 digit "barcode" of the presence/absence of gene orthologs across the 13 species for all proteins in the pan-genome.

Acknowledgements. This study was supported by the Danish Research Council for Technology and Production Science with Sapere Aude (#116262). Instrumentation and software used in this study was supported by Agilent Technologies Thought Leader Donation. We acknowledge Farooq Azam and Krystal Rypien of Scripps Institution of Oceanography, UCSD, for supplying strain 2ta16; Antonio Sanchez-Amat of University of Murcia for supplying strains CPMOR-1/CPMOR-2; and Tillman Harder of University of New South Wales for supplying strains H33/H33S. Don D. Nguyen and Laura Sanchez are acknowledged for introduction to molecular networking.

1. Peláez F (2006) The historical delivery of antibiotics from microbial natural products--can history repeat? *Biochem Pharmacol* 71:981–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16290171 [Accessed November 6, 2012].

2. Clardy J, Fischbach M a, Walsh CT (2006) New antibiotics from bacterial natural products. *Nat Biotechnol* 24:1541–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17160060 [Accessed November 5, 2012].

3. Müller R, Wink J (2014) Future potential for anti-infectives from bacteria - how to exploit biodiversity and genomic potential. *Int J Med Microbiol* 304:3–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24119567 [Accessed November 10, 2014].

4. Aigle B et al. (2014) Genome mining of Streptomyces ambofaciens. *J Ind Microbiol Biotechnol* 41:251–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24258629 [Accessed November 10, 2014].

5. Goldman BS et al. (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *Proc Natl Acad Sci U S A* 103:15200–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1622800&tool=pmcentrez&rendertype=abstract.

6. Omura S et al. (2001) Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* 98:12215–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=59794&tool=pmcentrez&rendertype=abstract.

7. Weber T (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int J Med Microbiol*:1–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24631213 [Accessed March 19, 2014].

8. Medema MH et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39:W339–46. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125804&tool=pmcentrez&rendertype=abstract [Accessed November 8, 2012].

9. Blin K et al. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* 41:W204–12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3692088&tool=pmcentrez&rendertype=abstract [Accessed February 23, 2014].

10. Ziemert N et al. (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* 7:e34064. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315503&tool=pmcentrez&rendertype=abstract [Accessed November 5, 2012].

11. Mann R a et al. (2013) Comparative genomics of 12 strains of Erwinia amylovora identifies a pan-genome with a large conserved core. *PLoS One* 8:e55644. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3567147&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

12. Park J et al. (2012) Comparative genomics of the classical Bordetella subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics* 13:545. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3533505&tool=pmcentrez&rendertype=abstract.

13. Penn K, Jensen PR (2012) Comparative genomics reveals evidence of marine adaptation in Salinispora species. *BMC Genomics* 13:86. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3314556&tool=pmcentrez&rendertype=abstract [Accessed September 30, 2013].

14. Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* 60:708–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2974192&tool=pmcentrez&rendertype=abstract [Accessed January 22, 2014].

15. Tagomori K, Iida T, Honda T (2002) Comparison of Genome Structures of Vibrios , Bacteria Possessing Two Chromosomes. 184:4351–4358.

16. Aylward FO et al. (2013) Comparison of 26 sphingomonad genomes reveals diverse environmental adaptations and biodegradative capabilities. *Appl Environ Microbiol* 79:3724–33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3675938&tool=pmcentrez&rendertype=abstract [Accessed April 30, 2014].

17. Penn K et al. (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 3:1193–203. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2749086&tool=pmcentrez&rendertype=abstract [Accessed November 20, 2012].

18. Ziemert N et al. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci* 2014:1–10. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1324161111 [Accessed March 12, 2014].

19. Krug D et al. (2008) Discovering the hidden secondary metabolome of Myxococcus xanthus: a study of intraspecific diversity. *Appl Environ Microbiol* 74:3058–68. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2394937&tool=pmcentrez&rendertype=abstract [Accessed August 7, 2014].

20. Lommen A (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem* 81:3079–86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19301908 [Accessed November 10, 2014].

21. Katajamaa M, Miettinen J, Oresic M (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22:634–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16403790 [Accessed November 10, 2014].

22. Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11:395. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2918584&tool=pmcentrez&rendertype=abstract [Accessed July 31, 2014].

23. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* 84:283–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3658281&tool=pmcentrez&rendertype=abstract [Accessed October 27, 2014].

24. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem* 84:5035–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3703953&tool=pmcentrez&rendertype=abstract.

25. Katajamaa M, Oresic M (2007) Data processing for mass spectrometry-based metabolomics. *J Chromatogr A* 1158:318–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17466315 [Accessed September 24, 2013].

26. Lange E, Tautenhahn R, Neumann S, Gröpl C (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* 9:375.

27. Nguyen DD et al. (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proc Natl Acad Sci*. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1303471110 [Accessed June 25, 2013].

28. Watrous J et al. (2012) Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* 109:E1743–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22586093 [Accessed October 26, 2012].

29. Ng J et al. (2009) Dereplication and de novo sequencing of nonribosomal peptides. *Nat Methods* 6:596–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2754211&tool=pmcentrez&rendertype=abstract [Accessed November 19, 2012].

30. Liu W-T et al. (2009) Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Anal Chem* 81:4200–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2765223&tool=pmcentrez&rendertype=abstract.

31. Yang JY et al. (2013) Molecular Networking as a Dereplication Strategy. *J Nat Prod*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24025162.

32. Bowman JP (2007) Bioactive compound synthetic capacity and ecological significance of marine bacterial genus pseudoalteromonas. *Mar Drugs* 5:220–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2365693&tool=pmcentrez&rendertype=abstract.

33. Holmström C, Kjelleberg S (1999) Marine Pseudoalteromonas species are associated with higher organisms and produce biologically active extracellular agents. *FEMS Microbiol Ecol* 30:285–293. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10568837.

34. Vynne NG, Mansson M, Gram L (2012) Gene sequence based clustering assists in dereplication of Pseudoalteromonas luteoviolacea strains with identical inhibitory activity and antibiotic production. *Mar Drugs* 10:1729–40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3447336&tool=pmcentrez&rendertype=abstract [Accessed December 8, 2012].

35. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. *Curr Opin Genet Dev* 15:589–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16185861 [Accessed April 30, 2014].

36. Vesth T, Lagesen K, Acar Ö, Ussery D (2013) CMG-biotools, a free workbench for basic comparative microbial genomics. *PLoS One* 8:e60120. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3618517&tool=pmcentrez&rendertype=abstract [Accessed February 25, 2014].

37. Thomas T et al. (2008) Analysis of the Pseudoalteromonas tunicata genome reveals properties of a surface-associated life style in the marine environment. *PLoS One* 3:e3252. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18813346.

38. Médigue C et al. (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. *Genome Res* 15:1325–35. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1240074&tool=pmcentrez&rendertype=abstract [Accessed November 27, 2012].

39. Lin X et al. (2012) A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J Chromatogr B Analyt Technol Biomed Life Sci* 910:149–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22682888 [Accessed September 2, 2013].

40. Lin X et al. (2011) A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics* 7:549–558. Available at: http://link.springer.com/10.1007/s11306-011-0274-7 [Accessed August 19, 2013].

41. Zhang X, Enomoto K (2011) Characterization of a gene cluster and its putative promoter region for violacein biosynthesis in Pseudoalteromonas sp. 520P1. *Appl Microbiol Biotechnol* 90:1963–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21472536 [Accessed May 28, 2013].

42. Agarwal V et al. (2014) Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat Chem Biol*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24974229 [Accessed July 14, 2014].

43. Laatsch H (1995) STRUCTURE-ACTIVITY-RELATIONSHIPS OF PHENYLPYRROLES AND BENZOYLPYRROLES. *Chem Pharm Bull* 43:537 – 546.

44. Hornemam U, Hurley LH, Speedie MK, Floss HG (1970) The Biosynthesis. 430:178–179.

45. Woodard RW, Mascaro L, Horhammer R, Eisenstein S, Floss HG (1980) No Title. 276:6314–6318.

46. Speedie K Isolation and Characterization of Tryptophan and Indolepyruvate. 7819–7825.

47. Vecchione JJ, Sello JK (2009) A novel tryptophanyl-tRNA synthetase gene confers high-level resistance to indolmycin. *Antimicrob Agents Chemother* 53:3972–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2737876&tool=pmcentrez&rendertype=abstract [Accessed November 10, 2014].

48. Kitabatake M et al. (2002) Indolmycin resistance of Streptomyces coelicolor A3(2) by induced expression of one of its two tryptophanyl-tRNA synthetases. *J Biol Chem* 277:23882–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11970956 [Accessed November 10, 2014].

49. El-Sayed A, Hothersall J, Cooper S (2003) Characterization of the Mupirocin Biosynthesis Gene Cluster from Pseudomonas fluorescens NCIMB 10586. *Chem Biol* 21:419–430. Available at: http://www.sciencedirect.com/science/article/pii/S1074552103000917 [Accessed November 25, 2014].

50. Fukuda D et al. (2011) A natural plasmid uniquely encodes two biosynthetic pathways creating a potent anti-MRSA antibiotic. *PLoS One* 6:e18031. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3069032&tool=pmcentrez&rendertype=abstract [Accessed January 16, 2013].

51. Journal THE, Antibiotics OF Thiomarinols B and C , New Antimicrobial Antibiotics Produced by a Marine Bacterium derivatives , mophore part (holothin) of 6 possessed an additional. 48:907–909.

52. To C, Editor THE Thiomarinols D , E , F and G , NewHybrid Antimicrobial Antibiotics Produced by Isolation procedures a Marine Bacterium ; Isolation , Structure , thiomarinol TMAwas a major product and isolated from EtOAc extracts of the culture broth1 }. The residue obtai. 50:449–452.

53. Udwary DW et al. (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. *Proc Natl Acad Sci U S A* 104:10376–81. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1965521&tool=pmcentrez&rendertype=abstract.

54. Thomson NR et al. (2002) Complete genome sequence of the model actinomycete Streptomyces. 3.

55. Tettelin H et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 102:13950–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1216834&tool=pmcentrez&rendertype=abstract.

56. Osbourn A (2010) Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet* 26:449–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20739089 [Accessed October 26, 2012].

57. Firn RD (2003) Bioprospecting – why is it so unrewarding ? 207–216.

58. Tobie WC (1935) The Pigment of Bacillus violaceus: I. The Production, Extraction, and Purification of Violacein. *J Bacteriol* 29:223 – 227.

59. Yada S et al. (2008) Isolation and characterization of two groups of novel marine bacteria producing violacein. *Mar Biotechnol (NY)* 10:128–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17968625 [Accessed January 16, 2013].

60. Von Wittenau MS (1963) Chemistry of Indolmycin. *J Am Chem Soc* 85:3425 – 3431.

61. Månsson M et al. (2010) Explorative solid-phase extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification. *J Nat Prod* 73:1126–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20509666.

62. Krug D, Zurek G, Schneider B, Garcia R, Müller R (2008) Efficient mining of myxobacterial metabolite profiles enabled by liquid chromatography-electrospray ionisation-time-of-flight mass spectrometry and compound-based principal component analysis. *Anal Chim Acta* 624:97–106. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18706314 [Accessed September 2, 2010].

63. Vos M, Velicer GJ Genetic Population Structure of the Soil Bacterium Myxococcus xanthus at the Centimeter Scale Genetic Population Structure of the Soil Bacterium Myxococcus xanthus at the Centimeter Scale †. 72.

64. Antón J et al. (2013) High metabolomic microdiversity within co-occurring isolates of the extremely halophilic bacterium Salinibacter ruber. *PLoS One* 8:e64701. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3669384&tool=pmcentrez&rendertype=abstract [Accessed November 10, 2014].

65. Li S, Kang L, Zhao X-M (2014) A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *Biomed Res Int* 2014:362738. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3963368&tool=pmcentrez&rendertype=abstract [Accessed November 10, 2014].

66. Boccard J et al. (2010) Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in Arabidopsis thaliana. *Chemom Intell Lab Syst* 104:20–27. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0169743910000341 [Accessed September 2, 2013].

67. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM (2008) Analysis of metabolomic data using support vector machines. *Anal Chem* 80:7562–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18767870.

68. Gram L, Melchiorsen J, Bruhn JB (2010) Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar Biotechnol (NY)* 12:439–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19823914 [Accessed December 13, 2012].

69. Kildgaard S et al. (2014) Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. *Mar Drugs* 12:3681–705. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4071597&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

70. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO (2011) Dereplication of microbial natural products by LC-DAD-TOFMS. *J Nat Prod* 74:2338–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22026385.

71. Smith C a et al. (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27:747–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16404815.

72. Chambers MC et al. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 30:918–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3471674&tool=pmcentrez&rendertype=abstract [Accessed October 1, 2014].

73. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27:431–2. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3031041&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

74. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 30:1218–23. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3670737&tool=pmcentrez&rendertype=abstract [Accessed February 27, 2014].

75. Camacho C et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

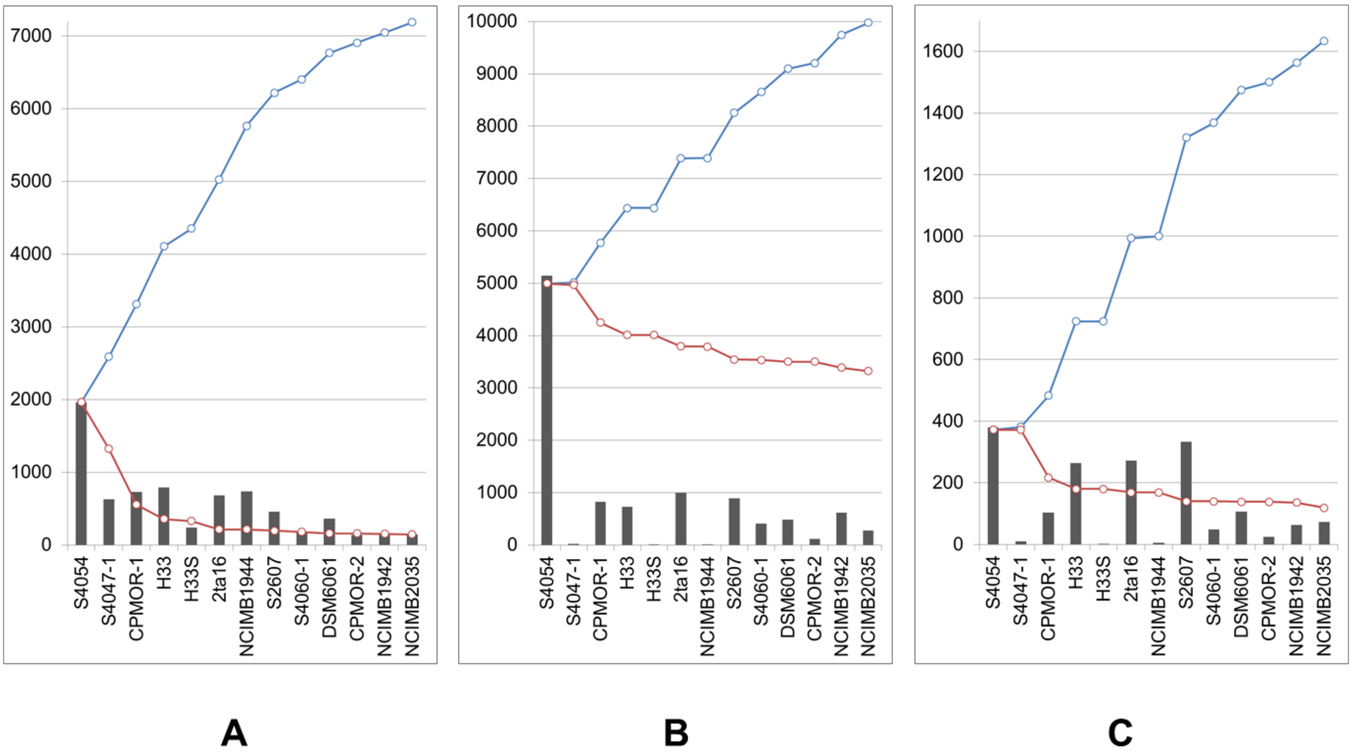


Fig. 1. Pan- and core metabolome and genome plots of 13 *P. luteoviolaceae* strains. A) The pan-metabolome curve (blue) connects the cumulative number of molecular features detected (positive and negative mode merged). The core-metabolome curve (red) connects the conserved number of features. The bars show the number of new molecular features detected in each extract (media components excluded). B) The pan- (blue) and core- (red) genome curves for all predicted genes. C) The pan- (blue) and core- (red) genome curves for genes predicted to be involved in secondary metabolism.

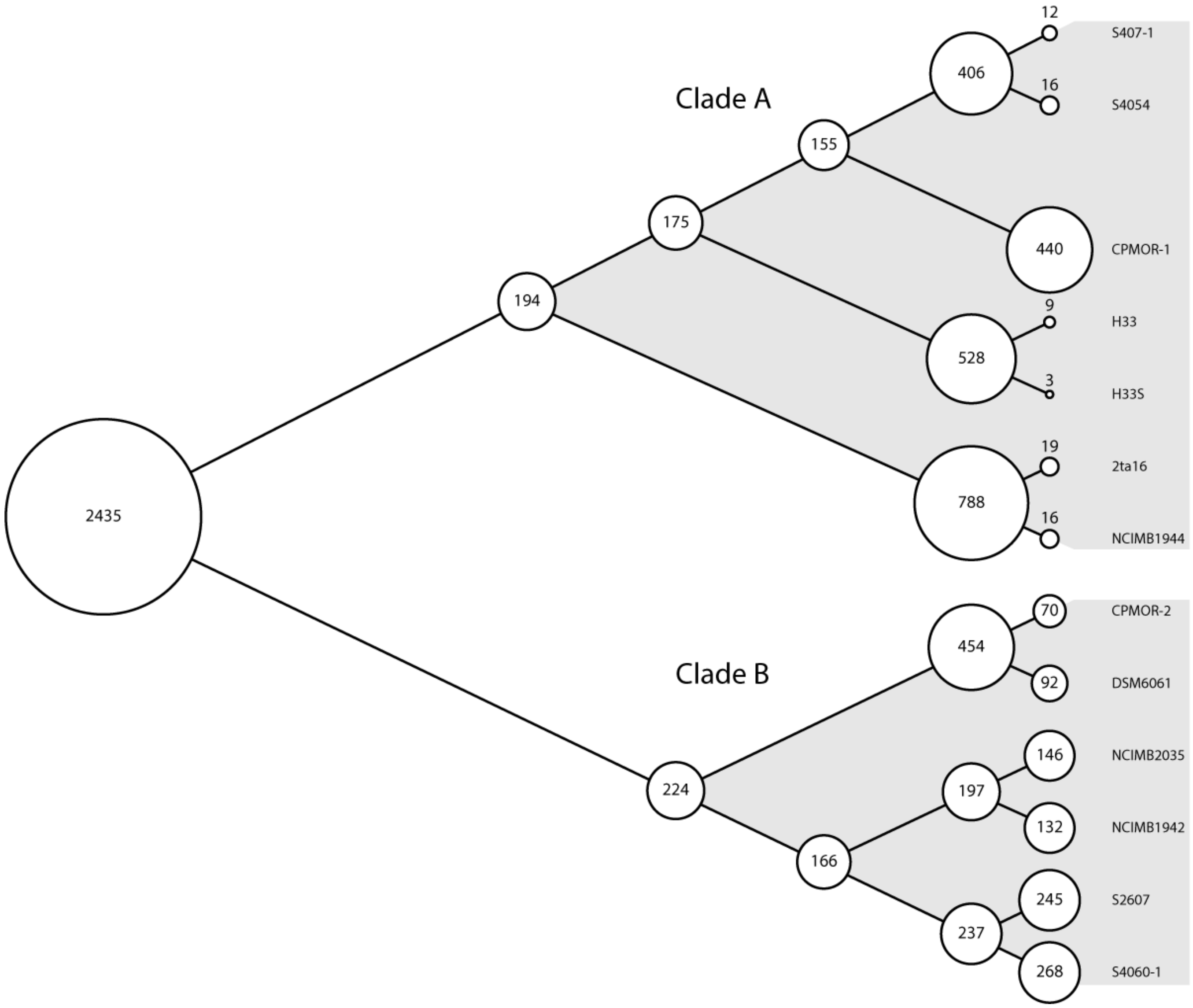


Fig. 2: Tree of shared genes for groups of species with OBUs overlaid. The numbers in the nodes shows the number of mutual 1:1 orthologs found in the species to the right of that circle. The areas of the nodes are proportional to the number of genes. The length of the edges only illustrates connectivity and not phylogenetic distance.

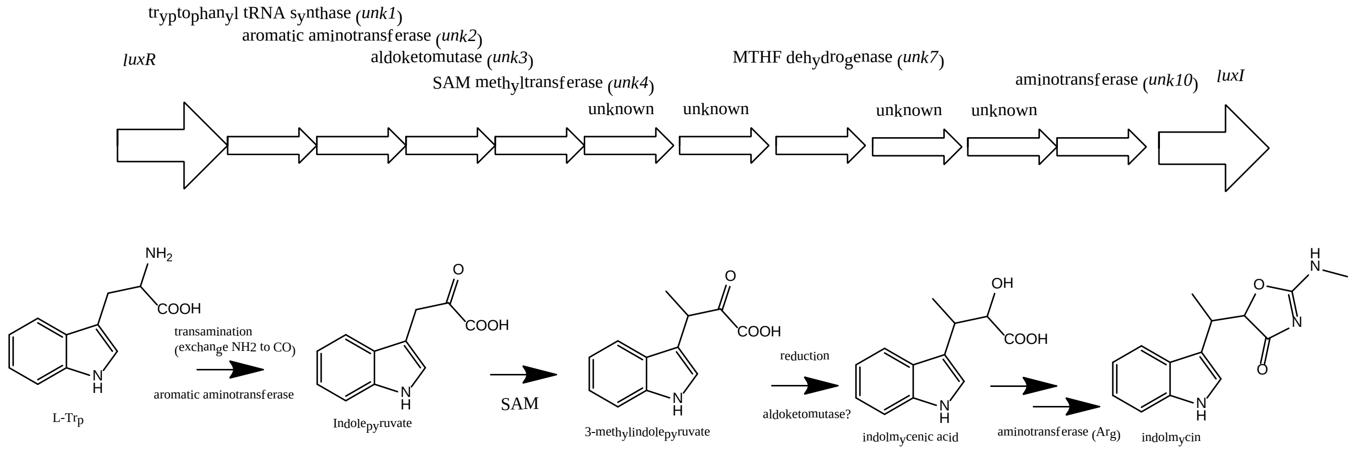
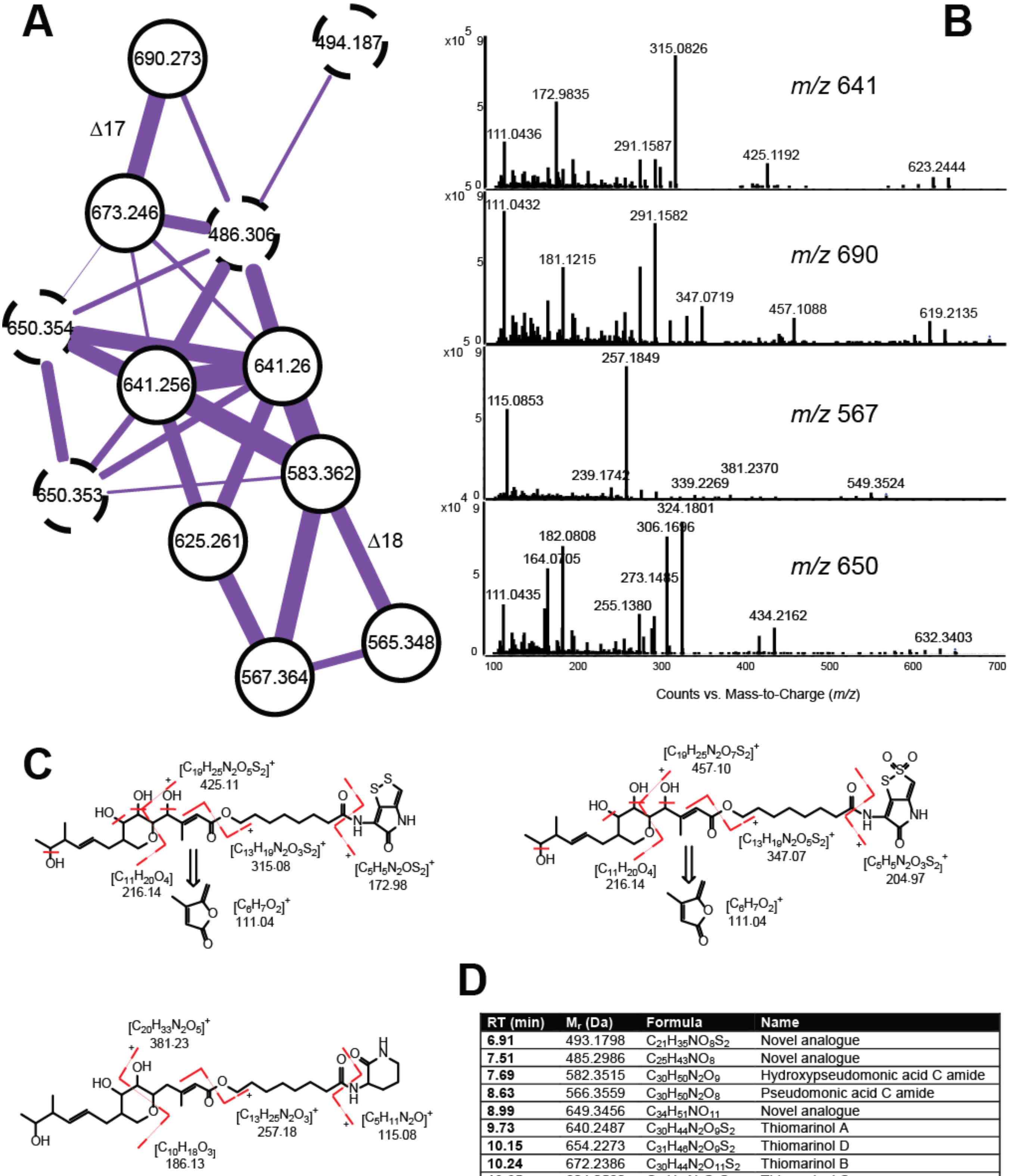


Fig. 3. Putative biosynthetic cluster (A) and proposed biosynthetic scheme (B)(1) for indolmycin. Color-codes for enzyme functions rather than names? ORFs?



RT (min)	M_r (Da)	Formula	Name
6.91	493.1798	$C_{21}H_{35}NO_8S_2$	Novel analogue
7.51	485.2986	$C_{25}H_{43}NO_8$	Novel analogue
7.69	582.3515	$C_{30}H_{50}N_2O_9$	Hydroxypseudomonic acid C amide
8.63	566.3559	$C_{30}H_{50}N_2O_8$	Pseudomonic acid C amide
8.99	649.3456	$C_{34}H_{51}NO_{11}$	Novel analogue
9.73	640.2487	$C_{30}H_{44}N_2O_9S_2$	Thiomarinol A
10.15	654.2273	$C_{31}H_{46}N_2O_9S_2$	Thiomarinol D
10.24	672.2386	$C_{30}H_{44}N_2O_{11}S_2$	Thiomarinol B
10.65	624.2532	$C_{30}H_{44}N_2O_8S_2$	Thiomarinol C

Fig. 4. A) Molecular network of the thiomarinol/pseudomonic acid molecular family. Dashed nodes indicate novel analogues. Mass differences are highlighted for ion adducts only. B) MS/MS spectra representing the four different analogue types. Parent mass *m/z* 641 is thiomarinol A, representing the holothin head type; *m/z* 690 is $[M+NH_4]^+$ of *m/z* 673 thiomarinol B, representing the sulfone head type; *m/z* 567 is pseudomonic acid C amide, representing the non-sulfonated analogues; *m/z* 650 is a novel analogue with a non-sulfonated head. C) Structures and suggested fragmentation of thiomarinol A, B, and pseudomonic acid C amide. D) Table of detected analogues in strains NCIMB1944 and 2ta16.

Supplementary Information for

Integrated Metabolomic and Genomic Mining of the Biosynthetic Potential of the Marine Bacterial *Pseudoalteromonas luteoviolacea* species

Maria Maansson[1,a], Nikolaj G.Vynne[1], Andreas Klitgaard[1], Jane L. Nybo[1], Jette Melchiorsen[1], Nadine Ziemert[2], Pieter C. Dorrestein[2,3,4], Mikael R. Andersen[1], and Lone Gram[1,b]

[1]Department of Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[2]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093

[3]Departments of Pharmacology and Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093

[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA 92093

[a]Present address: Chr. Hansen A/S, Boege Allé 10-12, DK-2970 Hoersholm, Denmark

[b]Author to whom correspondence should be addressed. Email: gram@bio.dtu.dk

Supplementary materials and methods:

Strain cultivation and extraction. The strains were cultured in biological duplicates in 20 mL Marine Broth at 25 ºC (200 rpm) for 48h. Cultures were extracted with 20 mL ethyl acetate (EtOAc) with 0.1% formic acid (FA), ultrasonicated for 10 min, and left on a shaking table (100 rpm) for 30 min. Phases were separated by centrifugation (3000 rcf, 4 ºC, 15 min). The cultures were re-extracted with 10 mL butanol (BuOH). The supernatants were pooled, dried under nitrogen, and re-dissolved in 2 mL methanol (MeOH). Samples for LC-MS/MS and molecular networking were used directly (1 µL injection), while samples for LC-MS and untargeted feature extraction were diluted 20-fold before injection (3 µL injection).

LC-MS and LC-MS/MS data acquisition. LC-MS and MS/MS analyses were performed on an Agilent 6550 iFunnel Q-TOF LC-MS (Agilent Technologies, Santa Clara, CA, US) coupled to an Agilent 1290 Infinity UHPLC system equipped with a Flexible Cube module. Compounds were separated on a Poroshell 120 phenyl-hexyl column (Agilent, 250 mm × 2.1 mm, 2.7 µm) at 60 °C with a water-acetonitrile (AcCN) gradient (both buffered with 20 mM formic acid (FA)) running from 10-100% AcCN over 20 min followed by a 4 min wash (100% AcCN). The gradient was then returned to 10% AcCN for a total gradient time of 26 min. Data was recorded both in positive and negative electrospray (ESI) mode and data was acquired in the *m/z* 100-1,700 Da mass range with a sampling rate of 2 Hz. The instrument was tuned and calibrated using a proprietary Agilent calibration algorithm using the Agilent ESI-L tuning mix solution. During operation, a lock mass solution containing ions *m/z* 119.9881 and 966.0007 in negative and *m/z* 186.2216 and 922.0098 in positive was constantly infused.

Data for molecular networking was collected using a data-dependent ESI$^+$-LC-MS/MS as reported previously (48) with the following modifications. MS1 spectra were recorded in positive electrospray mode from *m/z* 200-1,700 Da followed by MS/MS with a fixed collision energy of 25 V and a speed of 5 scans/sec. Spectra were obtained for the three most intense ions, which were excluded after being detected twice; however, released after 0.5 min for detection of analogues with different retention times.

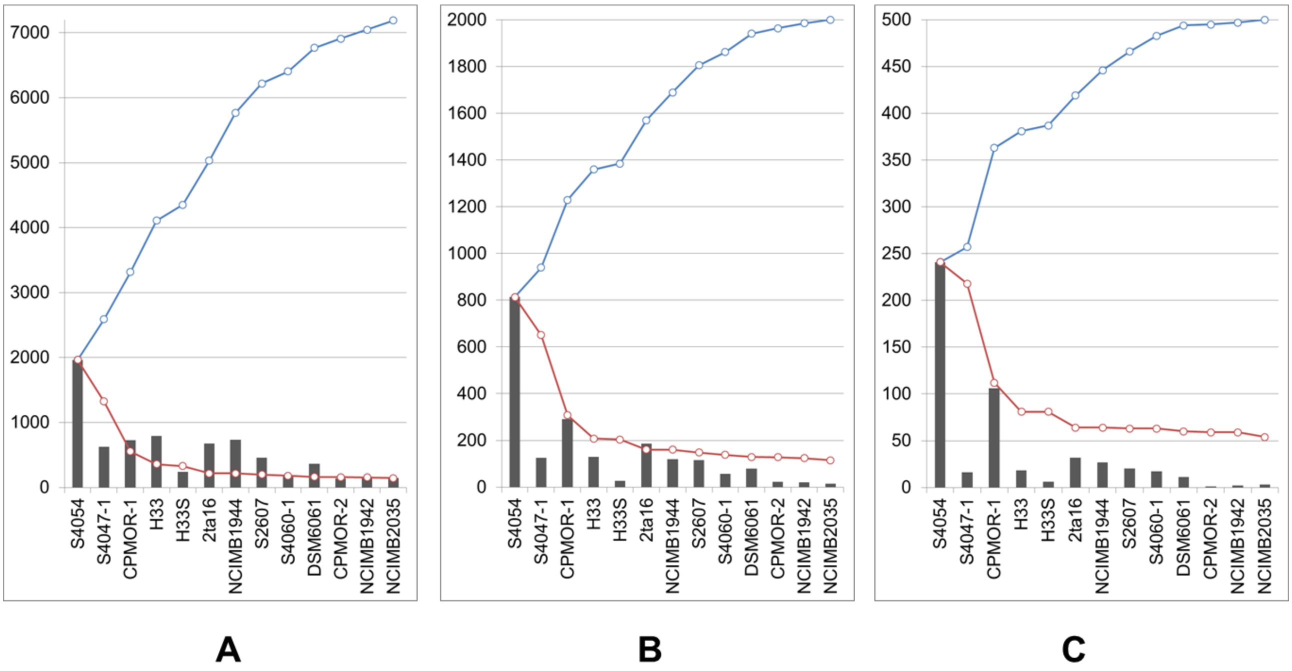
Due to carry-over in the auto sampler of certain compounds (polybrominated), the samples were split in to two groups to minimize carry-over, i.e. non- and positive PBP producers (24). Within the two groups, the samples were randomized using the macro developed by Bertrand *et al.* (47) with blank runs every 5 samples and blank media control samples every 10 samples to assess the extent of the carry-over throughout the batch. Extensive valve cleaning was applied during the run. Likewise, the Flexible Cube solvents were 20% dichloromethane in 2-propanol (v/v%) and 30% water in 2-propanol to maximize removal of problematic compounds.

Feature extraction and multivariate analysis. To deconvolute the raw total ion current spectra, the data-analysis program MassHunter (Agilent Technologies, v. B06.00) was used. Chemical features were extracted from the LC-MS data using the Molecular Features Extraction (MFE) algorithm and the recursive analysis workflow. Features were extracted from RT 2.00-21.00 min, with a minimum intensity of 5,000 counts and aligned considering adducts ($[M+H]^+$, $[M+Na]^+$, $[M-H]^-$, $[M+Cl/Br]^-$, $[M+CH_3COO]^-$) and neutral losses ($[M-H_2O]^+$). The isotopes of the chemical features were detected using a tolerance of 0.0025 *m/z* + 7 ppm error, and were limited to a charge state of 1, while compounds with an interminable charge were excluded. Feature alignment, binning, and alignment was performed using the following tolerances (Δ*m/z* 0.0025 ± 7 ppm), mass window set (±0.2 min, 15 ppm), and a MFE quality score of minimum 98. Only features present in both replicate samples were considered. For the recursive feature extraction, chromatograms were smoothed using a Gaussian function (3 point function width and 1.5 point Gaussian width) and a cut-off intensity of 3,500 counts was used. The threshold used for the MFE and recursive analysis was purposely set low to allow for the detection of numerous features to ensure correct alignment of peaks, after which the aligned feature list could be filtered based on a higher threshold.

Feature lists were imported to Genespring – Mass Profiler Professional (MPP) (Agilent Technologies, v. 12.6), and filtered for features with raw intensities lower than 100,000 (ESI^+ data) and 60,000 counts (ESI^- data). Media components or other interfering signals were defined as peaks present in the medium blank and these were manually excluded from the analysis. Features present in all samples (including the blank), but having more than a 10x fold change between sample and medium blank were treated as potential carry-over and included on the ‘true compound’ feature list. The lists from ESI^+ and ESI^- were merged in an Excel table as generic data and reimported into MPP, where features within RT ±0.15 min and 15 ppm mass tolerance were aligned. Intensities were normalized (quantile) and Z-transformed due to differences in intensities in ESI^+ and ESI^-. A total number of 8,699 features were aligned. By only taken in to account features present in both replicates, the number of features was reduced to 7,190. The list of discriminating features was generated in MPP using genetic algorithm with a population size of 25, 10 generations, and a mutation rate of 1. The GA was evaluated using the SVM with a linear kernel type with and imposed cost of 100 and ratio of 1. The feature list was validated via the leave-one-out method.

Mass defect screening. A list of all halogen containing compounds described from *Pseudoalteromonas* and *Alteromonas* was extracted from AntiMarin (v. 08.13). Based on this, the minimum mass defect from any metabolite was found to be 0.0937 Da, whilst the lowest mass defect increase per 100 Da was found to be 0.0263 Da. Chemical features were extracted using the same settings as for the MFE analysis, and then filtered for compounds with a mass defect of -0.0937 Da with -0.02 Da per 100 Da at a tolerance of +/- 0.0100 Da. Likewise, the listed was validated by the isotope patterns of the filtered features.

Figure S1. Filtered pan- and core-metabolome plots



Total	7190
Unique	2140
Core components	145
Core %	2%
Unique %	30%

Total	2000
Unique	408
Core components	115
Core %	6%
Unique %	20%

Total	500
Unique	93
Core components	54
Core %	11%
Unique %	19%

Fig. S1. A) The pan-metabolome curve (blue) connects the cumulative number of the total number of molecular features detected (positive and negative mode merged). The core-genome curve (red) connects the conserved number of features. The bars show the number of new molecular features detected in each extract (media components excluded). B) The pan- (blue) and core-(red) metabolome curves of the 2,000 most intense features. C) The pan- (blue) and core-(red) metabolome curves of the 500 most intense features.

Table S1. Overall genomic features of the 13 *P. luteoviolacea* strains

Strain #	Contig sequence total (Mb)	Genome contig count	# predicted protein coding genes	# unique genes	% genes allocated to secondary metabolism	# OBU*	Accession number
S4054	6.1	219	5146	19	7.4	13 (21)	
S4047-1	6.1	180	5130	8	7.6	13 (21)	
CPMOR-1	6	105	5186	560	6.2	6 (13)	
H33	6.1	151	5160	7	10.2	20 (29)	
H33S	6.1	143	5149	4	9.7	18 (26)	
2ta16	6.4	175	5355	49	9.3	10 (16)	PRJNA210324
NCIMB1944	6.4	107	5323	11	9.4	12 (18)	
S2607	5.9	73	4997	291	11.3	18 (25)	
S4060-1	6	74	5092	371	9.2	12 (18)	
DSM6061(T)	5.9	168	5003	119	8.0	10 (19)	
CPMOR-2	5.9	155	4948	115	8.7	10 (20)	
NCIMB1942	5.5	227	4831	553	7.2	7 (13)	
NCIMB2035			4926		7.1	9 (16)	

Table S1. Overall descriptive features of all 13 draft genomes. Total genes predicted using Prodigal 2.00, while antiSMASH 2.0 (4, 5) was used to predict the number of genes allocated to secondary metabolism. *The total number of OBUs (in parentheses) and number of PKS/NRPS pathways were calculated based on antiSMASH and NaPDoS (6) predictions and recursive analysis by MultiGeneBlast (77).

Table S2. Overview of predicted Operational Biosynthetic Units (OBUs)

Cluster	Predicted compound	S2607	S4060-1	CPMOR-2	DSM6061(T)	2ta16	NCIMB1944	NCIMB1942	NCIMB2035	H33	H33S	S4054	S4047-1	CPMOR-1
ripp1	Glycosylated lantipeptide	1	1	1	1	1	1	1	1	1	1	1	1	1
Homoserinelactone		1	1	1	1	1	1	1	1	1	1	1	1	1
NRPS1 + NRPSPKS4 *		1	1	1	1	1	1	1	1	1	1	1	1	1
ripp2	Bacteriocin	1	1	1	1	1	1	1	1	1	1	1	1	1
PKS2	Type III	1	1	1	1	1	1	1	1	1	1	1	1	1
vio	Violacein	1	1	1	1	1	1	1	1	1	1	1	1	1
NRPSPKS2		1	1	1	1	1	1	1	1	1	1	1	1	1
ripp3	Bacteriocin	1	1	1	1	1	1	1	1	1	1	1	1	1
NRPSPKS3	Trans AT	1	1	1	1	1	1	1	1	1	1	1	1	1
ripp4 **		1	1	1	1	1	1	1	1	1	1	1	1	1
NRPS3		0	0	0	0	1	1	0	0	1	1	1	1	1
NRPS7	Siderophore	1	1	1	1	0	0	1	1	0	0	0	0	0
other2 ** + other3 **		1	1	1	1	0	0	1	1	0	0	0	0	0
PKS3	Type2 PKS	1	1	1	1	0	0	1	1	0	0	0	0	0
NRPS2	Dipeptide	1	1	(0)	(0)	1	1	(0)	0	(0)	(0)	(0)	(0)	(0)
NRPS8	Pentapeptide	0	0	0	0	0	0	0	0	1	1	1	1	0
other 5		0	0	0	0	0	0	0	0	0	0	1	1	1
PKS1	Trans AT PKS	0	0	0	0	1	1	0	0	0	0	0	0	0
NRPSPKS1		0	0	0	0	1	1	0	0	0	0	0	0	0
other1		0	0	0	0	1	1	0	0	0	0	0	0	0
tml	Thiomarinol	0	0	0	0	1	1	0	0	0	0	0	0	0
NRPS5		0	0	1	1	0	0	0	0	0	0	0	0	0
NRPS6*		0	0	1	1	0	0	0	0	0	0	0	0	0
NRPSPKS6		0	0	1	1	0	0	0	0	0	0	0	0	0
NRPSPKS7		0	0	0	0	0	0	0	0	1	1	0	0	0
other4		0	0	0	0	0	0	0	0	1	1	0	0	0
NRPS9	TypeIII/NRPS	0	0	0	0	0	0	0	0	1	1	0	0	0
NRPS11 *		0	0	0	0	0	0	0	0	1	1	0	0	0
NRPS10		0	0	0	0	0	0	1	1	0	0	0	0	0
NRPS13 * + NRPS15 *		1	1	0	0	0	0	0	0	0	0	0	0	0
NRPSPKS10		0	0	0	0	0	0	0	0	0	0	1	1	0
NRPSPKS11 *		0	0	0	0	0	0	0	0	0	0	1	1	0
NRPS4		0	0	0	0	0	0	0	0	0	0	0	0	1
NRPS11	Lipopeptide	1	0	0	0	0	0	0	0	0	0	0	0	0
NRPS12 + NRPSPKS8 *	Lipopeptide	1	0	0	0	0	0	0	0	0	0	0	0	0
NRPSPKS9 *		1	0	0	0	0	0	0	0	0	0	0	0	0
NRPS14 * + NRPS16 *		0	1	0	0	0	0	0	0	0	0	0	0	0
Total no. of pathways		**18**	**16**	**16**	**16**	**16**	**16**	**14**	**14**	**16**	**16**	**15**	**15**	**13**
NRPS/PKS		**11**	**9**	**9**	**9**	**9**	**9**	**7**	**7**	**9**	**9**	**8**	**8**	**6**

Table S2. Pathway (OBU) distributions among the 13 *Pseudoalteromonas luteoviolacea* strains and their tentative functionality as predicted by antiSMASH. * Marks partial pathways on split contigs. Partial pathways with the same pattern of conservation are combined in order ot avoid overestimation of diversity; ** Gene cluster?

Table S3. 50 discriminating molecular features identified by GA/SVM

RT (min)	M_r (Da)	Formula	Δ *m/z* (ppm)	Detected ion	Tentative ID (comments)	S2607	S4060-1	CPMOR-2	DSM6061(T)	2ta16	NCIMB1944	NCIMB1942	NCIMB2035	H33	H33S	S4054	S4047-1	CPMOR-1
2.31	203.0614	-	-	-	Potential noise	0	0	0	0	0	0	0	1	0	0	1	1	0
2.64	218.1019	$C_{12}H_{14}N_2O_2$	- 0.17	$[M+H]^+$	Cyclo(Ala-Phe) isomer	1	1	0	0	1	1	0	0	1	1	1	1	0
2.88	218.1071	$C_{12}H_{14}N_2O_2$	- 0.17	$[M+H]^+$	Cyclo(Ala-Phe) isomer	1	1	1	0	1	1	1	0	1	1	1	1	0
3.15	175.0633	$C_{10}H_9NO_2$	+ 1.99	$[M+H]^+$	Indole-3-acetic acid	1	0	0	0	0	1	0	0	0	0	0	0	0
3.50	138.0325	C_6H_4O	- 6.99	$[M+COOH]^-$	Poor mass accuracy low mass range ESI^-	0	1	1	1	0	1	0	0	1	1	1	1	1
3.55	224.1166	$C_{11}H_{16}N_2O_3$	- 0.94	$[M+H]^+$	Aminochelin	0	0	0	0	1	1	0	0	1	0	0	0	0
3.69	449.1767	$C_{17}H_{31}N_5O_5S_2$	+ 0.30	$[M+Na]^+$		0	1	0	0	0	0	0	0	0	0	0	0	0
3.77	289.1063	$C_{14}H_{15}N_3O_4$	- 1.14	$[M+Na]^+$		0	0	0	0	0	0	0	0	0	0	1	1	1
4.67	554.2444	$C_{22}H_{42}N_4O_8S_2$	- 0.04	$[M+Na]^+$	Pantethine	1	1	1	1	1	1	1	1	1	1	1	1	1
5.38	448.1450	$C_{17}H_{28}N_4O_6S_2$	+ 0.00	$[M+Na]^+$		0	1	0	0	0	0	0	0	0	0	0	0	0
6.03	261.1113	$C_{13}H_{15}N_3O_3$	- 1.83	$[M+H]^+$	Cyclo(L-Asn-L-Phe)	0	0	0	0	0	0	0	0	0	0	1	1	1
6.31	438.1903	$C_{23}H_{26}N_4O_5$	- 0.48	$[M+Na]^+$		0	0	0	0	0	0	0	0	0	0	0	0	1
6.40	219.0895	$C_{12}H_{13}NO_3$	- 0.73	$[M+H]^+$	Methyl indole-3-lactate	0	0	0	0	0	0	0	0	0	0	0	1	1
6.61	872.5081	$C_{45}H_{72}N_6O_9S$	+ 0.00	$[M+2H]^+$		1	1	1	1	1	1	1	0	1	1	1	1	1
7.01	257.1164	$C_{14}H_{15}N_3O_2$	- 1.61	$[M+K]^+$	Indolmycin	0	0	0	0	0	0	0	0	0	0	1	1	1
7.08	283.0936	-	-	-	Potential noise	0	0	0	0	0	0	0	0	0	0	1	1	1
7.27	362.1577	$C_{16}H_{26}O_9$	- 0.08	$[M+Na]^+$		1	1	1	1	1	1	1	1	1	1	1	1	1
7.51	485.2989	$C_{24}H_{37}N_8O_3$	+ 0.33	$[M+Na]^+$		0	0	0	0	1	1	0	0	0	0	0	0	0
8.15	398.1550	$C_{15}H_{22}N_6O_7$	+ 0.86	$[M+H]^+$		1	1	1	1	1	1	1	1	1	1	1	1	1
8.35	343.0957	$C_{20}H_{13}N_3O_3$	- 0.33	$[M+H]^+$	Violacein	1	1	1	1	1	1	1	1	1	1	1	1	1
8.61	448.1945	$C_{20}H_{32}O_{11}$	- 0.23	$[M+Na]^+$		1	1	1	1	1	1	1	1	1	1	1	1	1
8.62	448.1945	$C_{20}H_{32}O_{11}$	- 0.23	$[M+NH_4]^+$		1	1	1	1	1	1	1	1	1	1	1	1	1
9.33	229.1678	$C_{12}H_{23}NO_3$	- 1.73	$[M+H]^+$	Dimethyl-2-oxodecanoylhydroxamic acid	1	1	1	0	0	1	0	0	0	1	0	0	0
9.34	656.2437	$C_{30}H_{44}N_2O_{10}S_2$	- 0.36	$[M+H]^+$		0	0	0	0	0	1	0	0	0	0	0	0	0

Continued on next page

Table S3 continued. 50 discriminating molecular features identified by GA/SVM

RT (min)	M_r (Da)	Formula	Δ *m/z* (ppm)	Detected ion	Tentative ID (comments)	S2607	S4060-1	CPMOR-2	DSM6061(T)	2ta16	NCIMB1944	NCIMB1942	NCIMB2035	H33	H33S	S4054	S4047-1	CPMOR-1
9.34	268.2033	-	-	-	Potential noise	1	1	1	0	0	1	0	0	0	0	0	0	0
9.73	640.2488	$C_{30}H_{44}N_2O_9S_2$	- 0.49	$[M+Na]^+$	Thiomarinol	0	0	0	0	1	1	0	0	0	0	0	0	0
10.38	290.1916	$C_{15}H_{30}O_3S$	- 1.32	$[M+H]^+$		0	0	0	1	0	1	0	1	1	1	1	1	0
10.64	620.2680	$C_{28}H_{44}O_{15}$	- 0.23	$[M+NH_4]^+$		1	1	1	1	0	1	1	1	1	1	1	1	0
10.66	681.2832	$C_{34}H_{43}N_5O_8S$	- 0.28	$[M+H]^+$		0	0	0	0	1	1	0	0	1	1	1	1	1
11.42	706.3061	$C_{33}H_{46}N_4O_{13}$	+ 0.43	$[M+NH_4]^+$		1	1	0	1	0	0	1	1	1	1	1	0	1
11.45	119.0734	-	-	-	Potential noise	0	0	0	0	0	0	0	0	0	0	0	0	1
11.87	320.1624	$C_{18}H_{24}O_5$	- 0.47	$[M-H_2O]^+$	Pseudoalteromone A	1	1	1	1	1	1	1	1	1	1	1	1	1
12.17	414.2743	$C_{22}H_{34}N_6O_2$	+ 0.80	$[M+H]^+$		0	0	1	1	1	1	0	0	1	0	0	0	0
12.17	785.5934	$C_{40}H_{83}N_9S_3$	- 0.12	$[M+H]^+$		0	0	1	1	1	1	0	0	1	0	0	1	0
12.46	361.2253	$C_{21}H_{31}NO_4$	+ 0.24	$[M+H]^+$		0	0	0	0	0	0	0	0	0	0	0	0	1
12.91	268.2038	$C_{16}H_{28}O_3$	+ 1.21	$[M+H]^+$	(7E)-9-Ketohexadec-7-enoic acid	1	0	1	1	0	0	1	1	1	1	1	1	0
14.86	431.2308	$C_{24}H_{33}NO_6$	+ 0.73	$[M+Na]^+$		1	1	1	0	0	1	0	0	0	1	0	0	0
15.44	1150.1522	$C_{10}H_2Br_9Cl_4N_7O_5$	+ 23.08	Ambiguous	Excellent isotope match	1	1	1	1	1	1	0	0	0	0	0	0	1
16.11	497.7101	$C_{12}H_6Br_4O_2$	+ 2.54	$[M-H]^-$	6,6'-Bis-(2,4-dibromophenole); MC21-A	0	0	0	1	0	1	0	0	0	0	0	0	0
16.59	658.6366	$C_{19}H_6Br_5NO$	- 2.70	$[M-H]^-$		1	1	1	0	0	1	0	0	0	0	0	0	0
16.77	745.7626	$C_{21}H_{23}Br_5N_2O_3$	- 0.23	$[M-H]^-$		1	1	1	0	0	1	0	0	0	0	0	0	0
17.14	566.4295	$C_{32}H_{58}N_2O_6$	+ 0.00	$[M+Na]^+$		1	1	1	1	0	1	1	1	1	0	1	1	0
17.45	602.4506	$C_{32}H_{62}N_2O_8$	+ 0.18	$[M+H]^+$		1	1	1	1	1	1	1	1	1	1	1	1	0
17.46	590.4268	$C_{30}H_{54}N_8O_4$	- 0.40	$[M+H]^+$		1	1	1	1	1	1	1	1	1	1	1	1	0
17.58	628.4663	$C_{34}H_{64}N_2O_8$	+ 0.34	$[M+H]^+$		1	1	1	1	1	1	1	1	1	1	1	1	0
17.60	676.5754	$C_{41}H_{76}N_2O_5$	- 0.16	$[M+H]^+$	Potential ornithine lipids	1	1	1	1	1	1	1	1	1	1	1	1	0
17.69	664.5754	$C_{40}H_{76}N_2O_5$	- 0.09	$[M+H]^+$	Potential ornithine lipids	1	1	1	1	1	1	1	1	1	1	1	1	0
18.39	939.4053	$C_{20}H_8Br_8N_2O_2$	+ 0.00	$[M-H]^-$		1	1	1	1	1	1	0	0	0	0	0	0	1
18.66	939.4053	$C_{20}H_8Br_8N_2O_2$	18.88	$[M-H]^-$	Excellent isotope match	1	1	1	1	1	1	0	0	0	0	0	0	1
20.33	658.5649	$C_{41}H_{74}N_2O_4$	- 0.39	$[M+Na]^+$		1	1	1	1	1	1	1	1	1	1	1	1	0

Table S3. The 50 descriminating molecular features identified with GA/SVM from the 500 most intense features. Molecular formulas are determined with MassHunter function ‘Generate formulas’, also considering the isotope pattern of the peak. All tentative IDs are based on hits in AntiMarin or Metlin, and the candidates are evaluated based on accurate mass, isotope pattern (in particular for the halogenated compounds), relative retention time, and fragmentation pattern (for Metlin hits).

Figure S2. Full molecular network

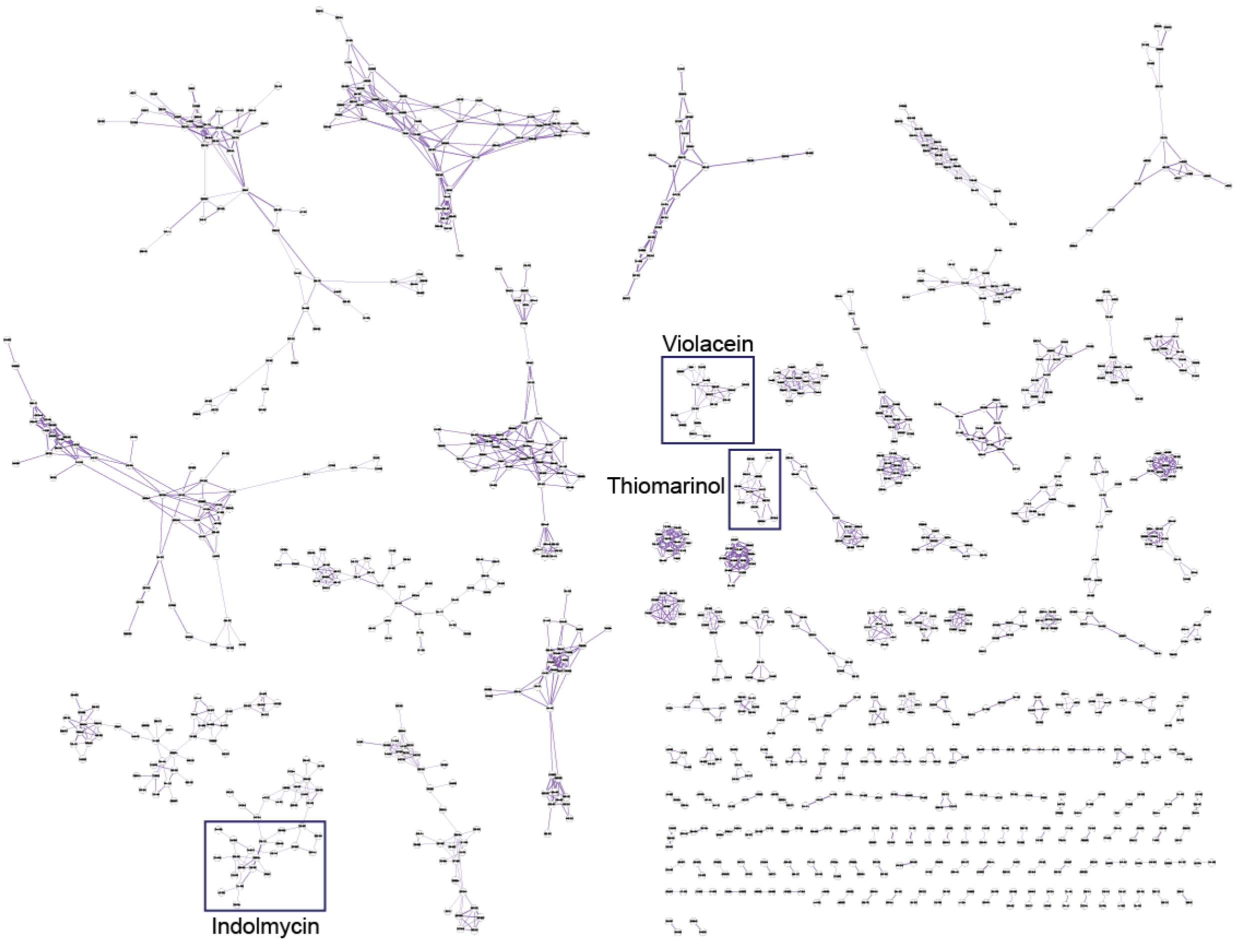


Fig. S2. Molecular network of 13 strains of *P. luteoviolacea* based on LC-ESI$^+$-MS/MS. Spectra originating from blank media samples are excluded from the analysis. Highlighted are the three gene cluster family-molecular family pairs identified in this study, those are violacein, indolmycin, and thiomarinol.

Figure S3. Network of the violacein molecular family

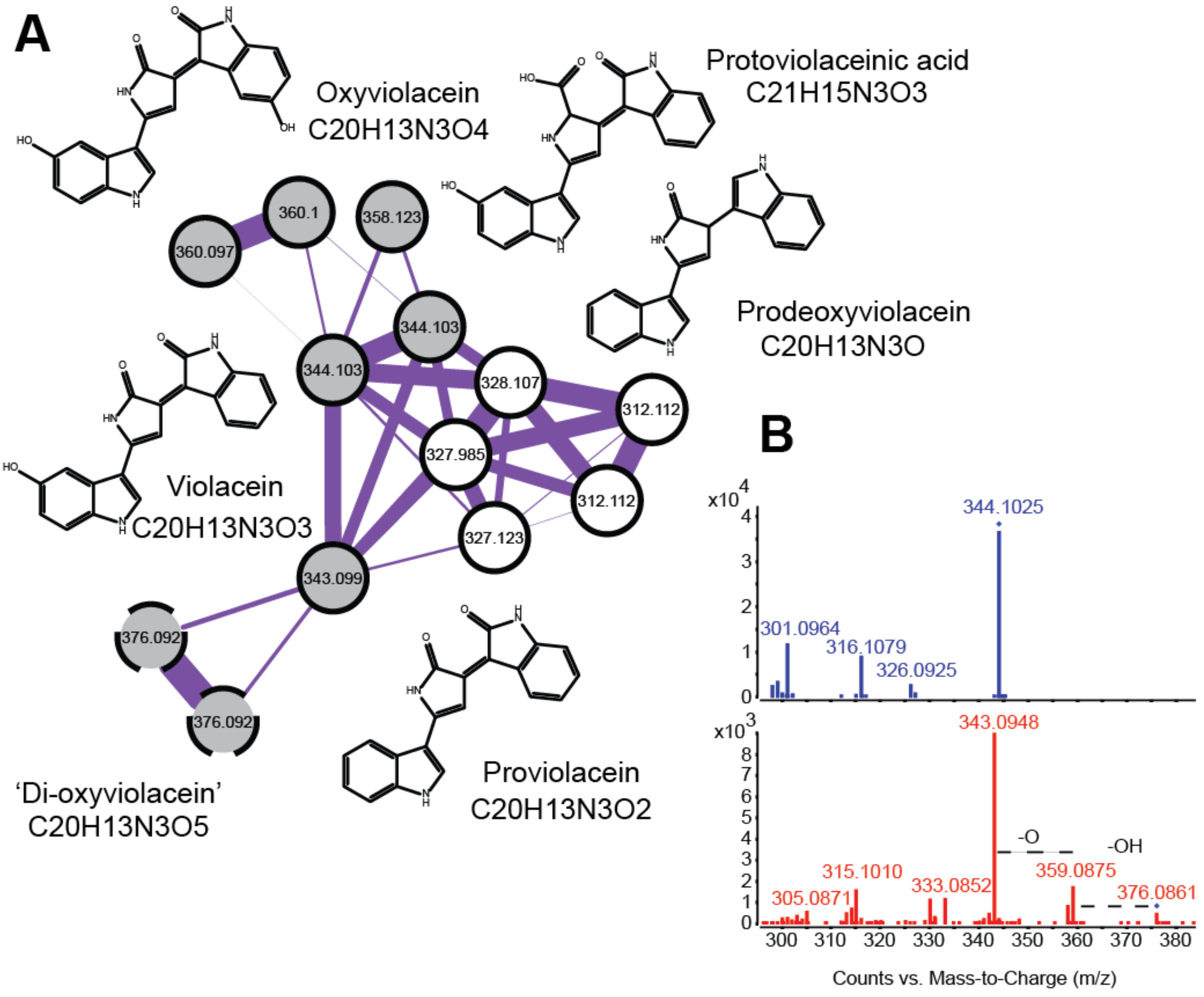


Fig. S3. A) Molecular network of the violacein MF. Grey nodes are shared between all strains, while white nodes are shared but multiple, but not all strains. Dashed nodes indicate a novel analogue. B) Selected zoom of MS/MS spectra of violacein (top) with parent mass $[M+H]^+$ 344 Da and the novel analogue (bottom) with an extra hydroxyl group $[M+H]^+$ 376 Da.

Table S4. Halogenated molecular features found by mass defect screening

RT (min)	M_r (Da)	Δ *m/z* (ppm)	Formula	Tentative ID (comments)	S2607	S4060-1	CPMOR-2	DSM6061(T)	2ta16	NCIMB1944	NCIMB1942	NCIMB2035	H33	H33S	S4054	S4047-1	CPMOR-1
5.56	215.9430	- 3.67	$C_7H_5BrO_3$	3-Bromo-4,5-dihydroxybenzaldehyde 2-Bromo-4,5-dihydroxybenzaldehyde 3-Bromo-4-hydroxybenzoic acid	1	1	1	1	1	1	0	0	0	0	0	0	1
6.43	369.8314	+0.51	$C_8H_9Br_3N_2$		1	1	1	0	0	0	0	0	0	0	0	0	0
7.04	293.8531	- 1.29	$C_7H_4Br_2O_3$	5,6-dibromoprotocatechualdehyde 2,3-Dibromo-4,5-dihydroxybenzaldehyde	0	0	1	1	1	1	0	0	0	0	0	0	1
8.91	386.8218	-0.13	$C_7H_8Br_3N_3O$		1	1	1	0	0	0	0	0	0	0	0	0	0
9.22	293.8532	- 1.63	$C_7H_4Br_2O_3$	5,6-dibromoprotocatechualdehyde	1	1	1	1	1	1	0	0	0	0	0	0	1
9.32	265.8586	- 2.99	$C_6H_4Br_2O_2$		1	1	1	1	1	1	0	0	0	0	0	0	0
9.90	449.8215	- 0.19	$C_{12}H_9Br_3N_2O_2$		0	1	1	0	0	0	0	0	0	0	0	0	0
10.06	397.8266	-0.25	$C_9H_9Br_3N_2O$		0	0	0	0	0	0	0	0	0	0	0	0	1
11.02	779.7992	+ 1.89	$C_{30}H_{16}Br_4N_4O_2$		1	1	1	1	1	1	0	0	0	0	0	0	1
12.33	436.7895	+ 0.65	$C_{11}H_6Br_3NO_3$		1	1	1	1	0	1	0	0	0	0	0	0	0
12.33	468.8158	+ 0.42	$C_{12}H_{10}Br_3NO_4$		1	1	1	1	0	1	0	0	0	0	0	0	0
13.24	567.7626	+ 1.11	$C_{15}H_{12}Br_4N_2O_2$		0	0	0	0	0	0	0	0	0	0	0	0	1
14.23	521.6220	- 0.31	$C_8H_3Br_5N_2$	Pentabromo-bipyrrole*	1	1	1	1	0	1	0	0	0	0	0	0	1
14.23	419.8001	- 1.15	$C_{12}H_7Br_3O_2$	Corallinaether 3,5,5'-tribromo-[1,1'-biphenyl]-2,2'-diol	1	1	1	1	1	1	0	0	0	0	0	0	1
14.36	659.7885	+ 1.43	$C_{21}H_{16}Br_4N_2O_3$		0	0	0	0	0	0	0	0	0	0	0	0	1
14.42	470.7109	- 0.92	$C_{10}H_5Br_4NO$	Tetrabromopseudilin*	1	1	1	1	1	1	0	0	0	0	0	0	1
14.73	511.6341	- 2.43	$C_8H_2Br_4Cl_2N_2$	Tetrabromo-dichloro-bipyrrole*	1	0	1	0	0	0	0	0	0	0	0	0	1
14.75	467.6837	- 0.71	$C_8H_2Br_3Cl_3N_2$	Tribromo-di-chloro-bipyrrole*	1	1	1	1	1	1	0	0	0	0	0	0	1
14.86	555.5841	- 3.17	$C_8H_2Br_5ClN_2$	Pentabromo-chloro-bipyrrole*	1	1	1	0	0	0	0	0	0	0	0	0	1
14.95	599.5334	- 2.63	$C_8H_2Br_6N_2$	Hexabromo-2'2-bipyrrole	1	1	1	0	0	0	0	0	0	0	0	0	1
15.20	460.7220	+ 0.02	$C_{10}H_4Br_3Cl_2NO$	Tribromo-dichloro-phenol-pyrrole*	1	0	1	1	0	1	0	0	0	0	0	0	1
15.26	497.7106	- 0.94	$C_{12}H_6Br_4O_2$	2-(2',4'-dibromophenoxy)-3,5-dibromophen	1	1	1	1	1	1	0	0	0	0	0	0	1
15.33	504.6727	- 2.39	$C_{10}H_4Br_4ClNO$	Tetrabromo-6-chloropseudiline	1	1	1	1	1	1	0	0	0	0	0	0	1
15.44	1152.1421	-	-	Poor isotope match **	1	1	1	1	1	1	0	0	0	0	0	0	1

Table S4 continued. Halogenated molecular features found by mass defect screening

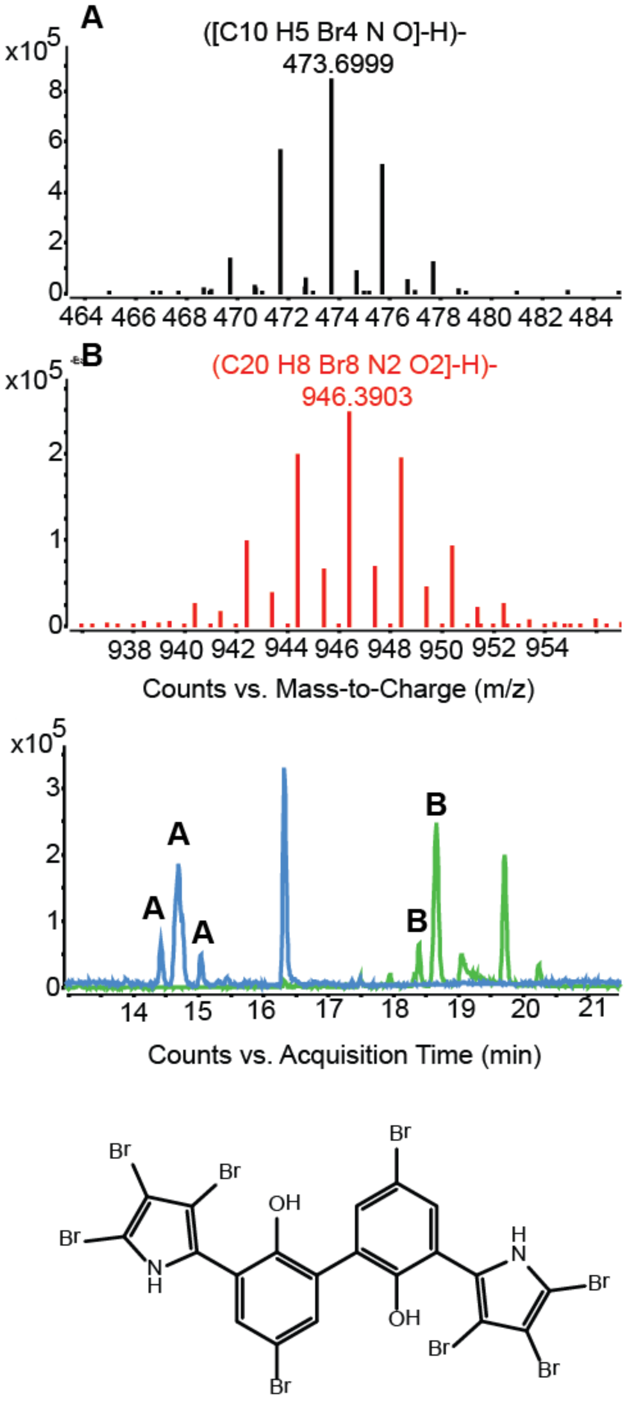
RT (min)	M_r (Da)	Δ *m/z* (ppm)	Formula	Tentative ID (comments)	S2607	S4060-1	CPMOR-2	DSM6061(T)	2ta16	NCIMB1944	NCIMB1942	NCIMB2035	H33	H33S	S4054	S4047-1	CPMOR-1
15.44	548.6216	- 1.13	$C_{10}H_4Br_5NO$	Pentabromopseudilin	1	1	1	1	1	1	0	0	0	0	0	0	1
16.11	497.7093	- 0.94	$C_{12}H_6Br_4O_2$	2-(2',4'-dibromophenoxy)-3,5-dibromophen	0	0	0	1	0	1	0	0	0	0	0	0	0
16.32	468.6955	- 1.46	$C_{10}H_3Br_4NO$	2,3,5,7-terabromobenzofuro[3,2-b]pyrrol	1	1	1	1	1	1	0	0	0	0	0	0	1
16.59	657.6420	- 0.94	$C_{20}H_6Br_5O$		1	1	1	0	0	1	0	0	0	0	0	0	0
16.76	745.7611	+ 1.95	$C_{21}H_{23}Br_5N_2O_3$		1	1	1	0	0	1	0	0	0	0	0	0	0
16.93	779.7453	- 0.71	$C_{20}H_{17}Br_5N_8O$		1	0	1	0	0	1	0	0	0	0	0	0	0
17.66	797.5000	- 0.18	$C_{16}H_9Br_7N_2O$		1	1	1	0	0	0	0	0	0	0	0	0	0
18.19	992.3117	+ 0.66	$C_{10}HBr_7Cl_3N_5O_9$	Poor isotope match**	1	1	1	0	0	0	0	0	0	0	0	0	1
18.25	921.5107	+ 2.73	$C_{19}H_{15}Br_7NO_7$		1	1	1	1	1	1	0	0	0	0	0	0	1
18.29	859.4774	+ 2.00	$C_{20}H_7Br_7N_2O_2$		1	1	1	1	1	1	0	0	0	0	0	0	1
18.39	939.4000	+ 5.62	$C_{20}H_8Br_8N_2O_2$	Bis-tetrabromopseudilin*	1	1	1	1	1	1	0	0	0	0	0	0	1
18.63	1019.3123	+ 0.56	$C_{17}HBr_7Cl_2O_{12}$		1	1	1	1	1	1	0	0	0	0	0	0	1
18.66	939.4028	+ 2.64	$C_{20}H_8Br_8N_2O_2$	Bis-tetrabromopseudilin*	1	1	1	1	1	1	0	0	0	0	0	0	1
18.78	999.4254	+ 1.01	$C_{22}H_{12}Br_8N_2O_4$		1	1	1	1	1	1	0	0	0	0	0	0	1
19.05	969.4156	+ 0.25	$C_{21}H_{10}Br_8N_2O_3$		1	1	1	1	1	1	0	0	0	0	0	0	1
20.29	1330.1637	+ 0.71	$C_{17}H_2Br_7Cl_6N_3O_{20}$		1	1	1	0	0	1	0	0	0	0	0	0	0

Table S4. List of halogenated molecular features identified by mass defect screening in MassHunter. The expected mass defect (0.0937 Da with -0.02 Da per 100 Da +/- 0.0100 Da) was determined from known halogenated compounds from *Pseudoalteromonas* in AntiMarin. The isotope pattern was used to confirm the presence of halogenations and used to calculate the molecular formula. Tentative IDs are based on hits in AntiMarin and evaluated based on accurate mass and isotope pattern. Compound marked * have no hit but belong to a known class of isomeric compounds. ** Peaks have a poor isotope match resulting in ambiguous determination of the formula.

Figure S4. *bmp* pathway

Fig. S4. *bmp* pathway + flanking putative resistance gene…

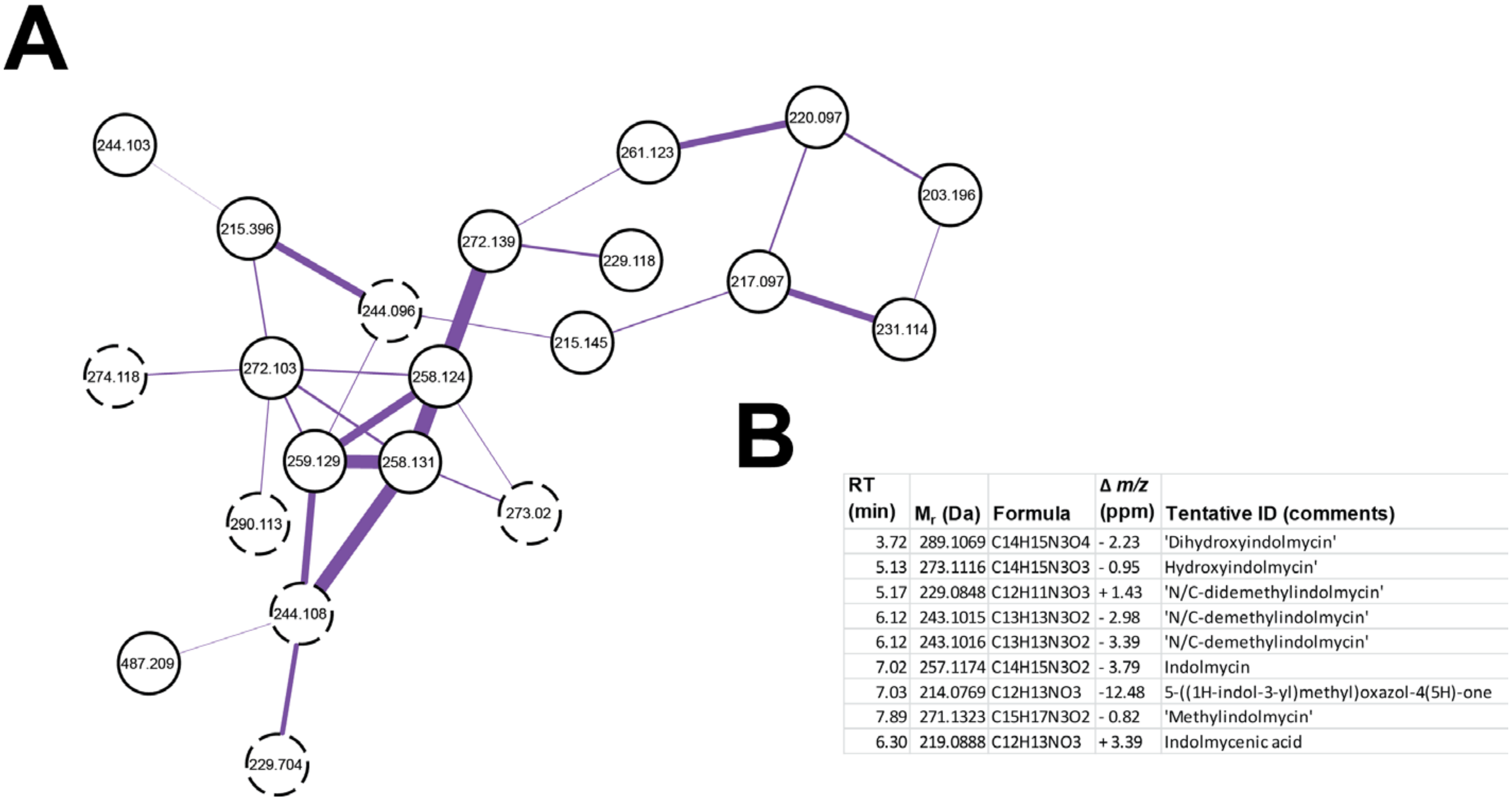
Figure S5. Tentative identification of dimeric halogenated compounds



B

Fig. S5. Isotope patterns of A) $C_{10}H_5Br_4NO$ (RT 14.42, 14.xx, and 14.xx min) and B) $C_{20}H_8Br_8N_2O_2$ (RT 18.39 + 18.66 min) detected in ESI$^-$ (top) and the corresponding EIC (bottom) and putative structure of a 'bis-tetrabromopseudilin'.

Figure S6. Network of the indolmycin molecular family



RT (min)	M_r (Da)	Formula	Δ *m/z* (ppm)	Tentative ID (comments)
3.72	289.1069	C14H15N3O4	- 2.23	'Dihydroxyindolmycin'
5.13	273.1116	C14H15N3O3	- 0.95	Hydroxyindolmycin'
5.17	229.0848	C12H11N3O3	+ 1.43	'N/C-didemethylindolmycin'
6.12	243.1015	C13H13N3O2	- 2.98	'N/C-demethylindolmycin'
6.12	243.1016	C13H13N3O2	- 3.39	'N/C-demethylindolmycin'
7.02	257.1174	C14H15N3O2	- 3.79	Indolmycin
7.03	214.0769	C12H13NO3	-12.48	5-((1H-indol-3-yl)methyl)oxazol-4(5H)-one
7.89	271.1323	C15H17N3O2	- 0.82	'Methylindolmycin'
6.30	219.0888	C12H13NO3	+ 3.39	Indolmycenic acid

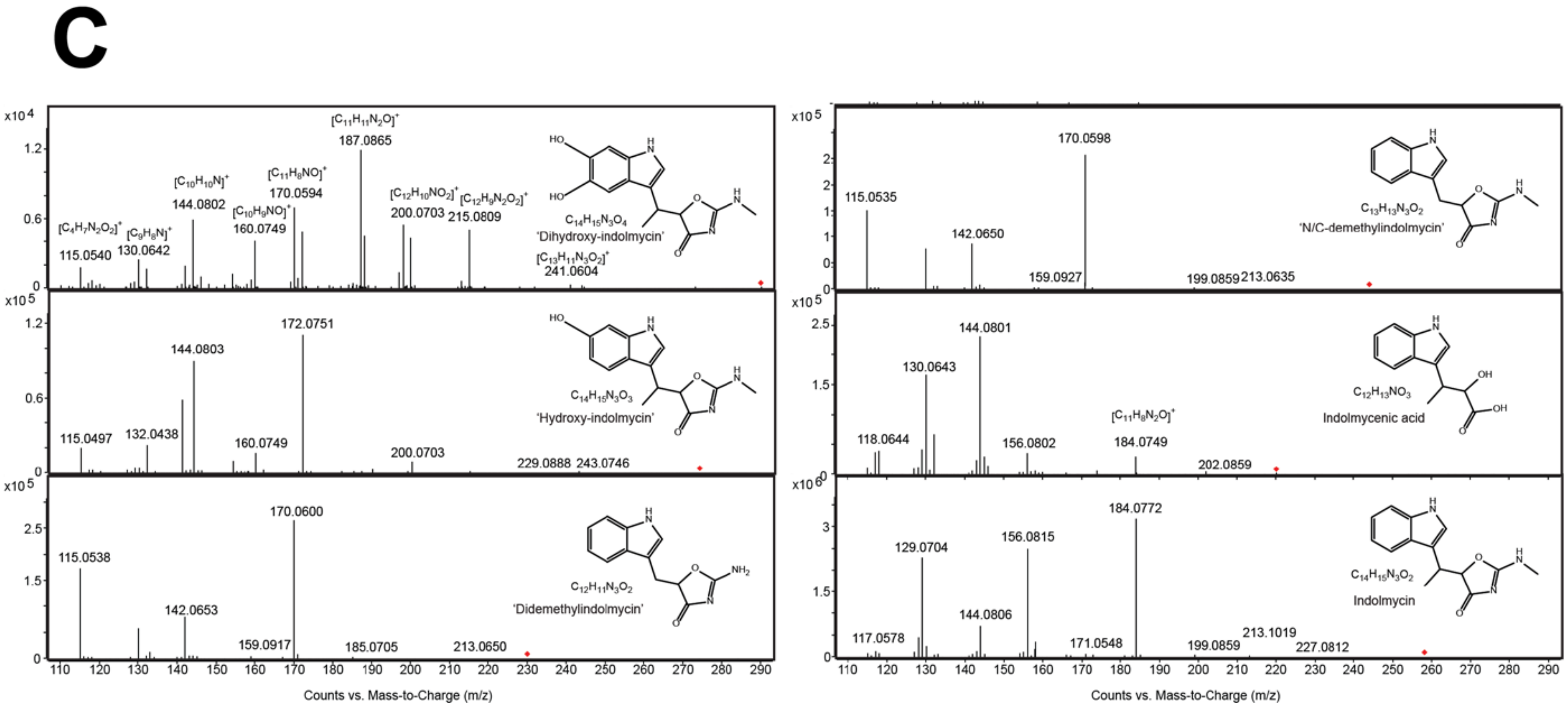


Fig. S6. A) Molecular network of the indolmycin molecular family. Dashed nodes indicate a novel analogue. B) Tentatively identified indolmycin analogues in strains S4047-1, S4054, and CPMOR-1. C) MS/MS spectra of selected analogues with assigned fragments.