The importance of mass spectrometric dereplication in fungal secondary metabolite analysis

Nielsen, Kristian Fog; Larsen, Thomas Ostenfeld

Published in: Frontiers in Microbiology

Link to article, DOI: 10.3389/fmicb.2015.00071

Publication date: 2015

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

Link back to DTU Orbit

The importance of mass spectrometric dereplication in fungal secondary metabolite analysis

Kristian F. Nielsen and Thomas O. Larsen*
Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark

Having entered the Genomic Era, it is now evident that the biosynthetic potential of filamentous fungi is much larger than was thought even a decade ago. Fungi harbor many cryptic gene clusters encoding for the biosynthesis of polyketides, non-ribosomal peptides, and terpenoids – which can all undergo extensive modifications by tailoring enzymes – thus potentially providing a large array of products from a single pathway. Elucidating the full chemical profile of a fungal species is a challenging exercise, even with elemental composition provided by high-resolution mass spectrometry (HRMS) used in combination with chemical databases (e.g., AntiBase) to dereplicate known compounds. This has led to a continuous effort to improve chromatographic separation in conjunction with improvement in HRMS detection. Major improvements have also occurred with 2D chromatography, ion-mobility, MS/MS and MS^n, stable isotope labeling feeding experiments, classic UV/Vis, and especially automated data-mining and metabolomics software approaches as the sheer amount of data generated is now the major challenge. This review will focus on the development and implementation of dereplication strategies and will highlight the importance of each stage of the process from sample preparation to chromatographic separation and finally toward both manual and more targeted methods for automated dereplication of fungal natural products using state-of-the art MS instrumentation.

Keywords: liquid chromatography, metabolomics, mass spectrometry, diode array detection, dereplication

INTRODUCTION
Filamentous fungi are prolific producers of secondary metabolites (SM) of importance to humankind. Useful fungal metabolites include drugs, food colorants, feed additives, industrial chemicals, and biofuels (Bills, 1995; Bode et al., 2002; Firn and Jones, 2003; Butler, 2004). Fungi are also known for their negative consequences as contaminants of food and feed due to the production of mycotoxins which can be cytotoxic, immunotoxic, estrogenic, or carcinogenic (Miller, 2008; Shephard, 2008). Fungi can also cause invasive human infections, especially in immuno-compromised individuals (Larsen et al., 2007).

Having entered the Genomic Era it is now clear that the biosynthetic and metabolic diversity potential of filamentous fungi is still vast, due to the presence of many cryptic SM encoding gene clusters. This is particularly true for non-model organisms. The abundance of cryptic gene clusters has resulted in the use of many strategies to stimulate gene expression aimed toward the discovery of novel bioactive compounds and characterization of their biosynthetic pathways (Bode et al., 2002; Bergmann et al., 2007; Bok et al., 2009; Schroechk et al., 2009; Sarkara et al., 2012; Sorensen et al., 2012; Droce et al., 2013; Sorensen and Søndergaard, 2014). Uncovering the full chemical potential of any micro-organism is a challenging exercise. Firstly, expression of metabolites related to a given gene cluster is highly regulated and may only be expressed under special condition, and furthermore each cluster may be responsible for more than 10 end products and a similar number of stable intermediates also present in detectable concentrations (Degenkolb et al., 2006; Chiang et al., 2010; Nielsen et al., 2011b; Ali et al., 2014; Holm et al., 2014; Petersen et al., 2015). In addition, crosstalk between biosynthetic pathways can result in compounds that are products of more than one gene cluster (Nielsen et al., 2011b; Tsunematsu et al., 2013). Overall this results in an extremely complex pool of diverse small organic molecules to identify by chemical analysis, especially when considering a single species of Aspergillus and Fusarium contains 50–80 gene clusters (Sanchez et al., 2012; Lysoe et al., 2014).

It is common for natural products to have identical elemental composition (up to 130 for some terpenes) making unambiguous identification very challenging. As such, without access to authentic standards, elemental composition alone – obtained by high resolution mass spectrometry (HRMS) – is not enough to unambiguously identify compounds (Nielsen et al., 2011a; El-Elimat et al., 2013). Consequently, fast identification of previously described compounds without reference standards – known as dereplication – is a very challenging task. As reference standards of most secondary metabolite standards are not available, fast dereplication is vital for progress in both drug discovery and pathway elucidation projects (Cordell and Shin, 1999; Dinan, 2005; Zhang, 2005; Bitzer et al., 2007; Feng and Siegel, 2007; Stadler et al., 2009; Zengler et al., 2009; Nielsen et al., 2011a). Dereplication is most often done by Ultra high performance liquid chromatography (UHPLC) coupled to diode array detection (DAD) and HRMS, in combination with database searching. Proper dereplication strategies ensures that time consuming and costly efforts with isolation and subsequent NMR based structure elucidation
can be focused solely on novel compounds (Cordell and Shin, 1999) or that re-isolation of known compounds can be done in an efficient way based on functional groups and thus fewer steps (Månsson et al., 2010). This paper will highlight important issues and recent approaches toward fast and reliable dereplication of fungal NPs primarily based on UHPLC-DAD-HRMS techniques.

THE IMPORTANCE OF SAMPLE PREPARATION

The chemical diversity which is found both between and within the many classes of secondary metabolites (e.g., polyketides (PKs), non-ribosomal peptides (NRPs) and terpenoids), makes it impossible to quantitatively extract all secondary metabolites from a given fungus using a single procedure. Consequently, sample preparation is an important aspect of secondary metabolite profiling and will, no matter the choice of method, lead to bias toward certain types of compounds. To avoid extracting too many polar media components, primary metabolites and sugars an organic extraction is needed using one or more water-immiscible solvents such as ethyl acetate (EtOAc; Nielsen and Larsen Dereplication by LC-MS 2011a; Stadler et al., 2014), dichloromethane (DCM; Abdel-Mawgoud et al., 2008), or 1-butanol (Yokota et al., 2012). The latter is efficient for the extraction of lipopeptides but has a high boiling point (118°C), thus requiring both N2 and heating for evaporation. The pH is vital for an organic extraction, as ionizable molecules will be extracted into the organic phase to a much higher degree in their neutral form than when in a charged state. As almost 50% of all described fungal NPs (Månsson et al., 2010) contains an acidic moiety, a low pH extraction is necessary in most cases but can be supplemented with a neutral extraction (e.g., for stability reasons). Solvents such as ethers (highly flammable and able to form explosive peroxides) as well the carcinogenic and environmental damaging chloroform (CHCl3) and carbon tetrachloride (CCl4) are being phased out. Methanol and ethanol are also efficacious but, due to their high polarity, also results in the extraction of large quantities of salts and polar interfering substances which can quickly clog analytical HPLC columns. On the other hand, methanol/ethanol extracts can also contain highly non-polar waxes, sterols and triglycerides, however, these can usually be flushed out of the column at 80°C with a mixture of acetonitrile-isopropanol for 1 h. We have experienced that injection of 1 μl of crude methanol extracts from marine media clogged an LC-MS electrospray source after analyzing only four extracts.

For extraction with EtOAc an essential process is the centrifugation of the 2-phase system as hard as possible then leaving behind the common interfacial agglomeration containing cells. Drying EtOAc extracts with anhydrous Na2SO4 can also result in cleaner extracts as EtOAc can contain up to 8% water. A major pitfall in sample preparation is the possibility of unwanted chemical reactions taking place. For example, alcohols can form esters with carboxylic acids and lactones (e.g., in homoserine lactones, rubratoxins, statins) under acidic conditions and can also catalyze various intramolecular changes (Rundberget et al., 2004). Acid catalyzed reactions can be especially problematic when evaporating organic extracts, since the lower volatility of acids such as formic and acetic acids will result in upconcentration when using volatile solvents like EtOAc and DCM, especially in the presence of residual water. Examples we have observed include: up to 50% loss of patulin during evaporation from EtOAc (Boonzaaijer et al., 2005); loss of trichothecenes prior to derivatization for GC-MS analysis (Langseth and Rundberget, 1998); up to 80% loss of fumonisins due to binding to the silanol groups of non-derivatized glass with up to 80% lost (Nielsen et al., 2015). We have used a fast extraction procedure to minimize the risk of studying artifact peaks, this procedure involves extraction into acetonitrile-water (1:1; in a ultrasonication bath or a beadbeater for circa 10 min), followed by centrifugation and subsequent transfer of the centrifugate directly to an auto-sampler compatible vial for analysis (Mogensen et al., 2011).

In some cases samples need to be purified on small solid phase extraction (SPE) columns to remove chromatography impairing lipids and phospholipids (Degenkolb et al., 2006; Pucci et al., 2009) or abundant but common secondary metabolites (e.g., as is Stachybotrys; Hinkley and Jarvis, 2000; Andersen et al., 2002). But SPE can also be used to simplify very complex extracts into several fractions. For example, ion exchange SPE can be used to separate an extract into acidic, neutral and basic analytes (Månsson et al., 2010), which may resolve new peaks and simplify the mass spectra. For an orthogonal separation to the common reversed phase analytical separation, we prefer diol or amino-propyl normal phase SPE (Bladt et al., 2013; Petersen et al., 2015) as we have found them to better separate compound classes compared to, e.g., pure silica.

ANALYTICAL SEPARATION

When grown on rich growth media fungi often produce in excess of 100 SMs (Stadler et al., 2007, 2014; Bertrand et al., 2013; Kildgaard et al., 2014; Kildgaard et al., 2014; Wolfender et al., 2015). With such complex mixtures it is practically impossible to acquire enough separation power to provide complete resolution of all individual metabolites and media components and allow for subsequent spectroscopic analysis. On balance, by far the best choice for separation is reverse phase (RP) chromatography, since it’s polarity is well suited to most SMs, especially with the emergence of more polar phases such as pentfluoro phenyl, biphenyl, phenyl (Kildgaard et al., 2014; Wolfender et al., 2015) and also columns with various embedded groups (Euerby and Petersson, 2005). Many vendors have introduced improved the solid core particles of their stationary phase, ensuring no diffusion through the particle center (e.g., Poroshell, Kinexet, Ascentis, and Cortecs). Furthermore, hybrid chemistry particles (e.g., Waters BEH) provide sharper peaks as well as reduced tailing due to secondary interactions at a wide pH range. Performing LC at low pH is statistically preferably as 50% of all secondary metabolites contain an acidic moiety, while ca. 10% have a basic moiety (Månsson et al., 2010). Since sharper peaks are obtained from the non-charged state of SMs, low pH is preferable in RP chromatography. Undoubtedly the most prevalent buffers/acidifiers used are formic and acetic acid (Varga et al., 2013). Although triflurooacetic acid (TFA) is commonly utilized to lower the pH, this has the consequence of significantly suppressing negative ionization MS detection. We have chosen formic acid as, in addition to its lower pKα, it suppresses
microbial growth in solvent reservoirs to a greater extent than acetic acid. We have found that, in general, formic acid appears to give sharper peaks at lower concentrations than acetic acid, and thus lower suppression of the UV signal. Although the use of formic acid to adjust the pH is generally satisfactory, for some compounds it results in poor ionization, for example type B trichotheccenes (Sulyok et al., 2006). Formic acid may also not be optimal if the majority of analysis is performed in negative mode – in which case acetic acid performs better – or if the analysis is performed under non-acidic conditions (Sulyok et al., 2006; Zhang et al., 2012).

Comprehensive 2D LC (Dugo et al., 2009; Chen et al., 2012) have recently been introduced by several of the UHPLC manufacturers and introduced some interesting possibilities for orthogonal separation through the use of a long high-polarity HPLC column (20–30 min run) in the first dimension and a short highly retentive small particle silica C18 in the second dimension (20–30 s runs).

Whereas RP chromatography is usually a suitable choice for the separation of the wide range of middle-polarity secondary metabolites mainly produced by fungi, smaller very polar compounds are usually poorly retained. As a result of the lack of retention this chemical window has been exploited to a much lesser degree due to difficulties with analysis. To better investigate compounds in this high polar and/or ionic window (compounds with LogD < -3) techniques such as HILIC (Sørensen et al., 2007; Creek et al., 2011), high-pH ion-chromatography (IC; van Dam et al., 2002; Kvitvang et al., 2014) or ion-pair chromatography (Lebrun et al., 1989; Werner, 1993; Magdenoska et al., 2013) are needed to sufficiently retain the analytes. Common to all these techniques is that the window of separation is much narrower than for the extremely versatile RP, thus requiring more optimization. A major benefit of these techniques can be found in that many NP chemists only examine organic extractable compounds thereby overlooking many interesting high polar compounds (Cannall, 1998; Dufresne, 1998; Shimizu, 1998; Månsson et al., 2010). Other alternative choices for retention of polar compounds include anion or cation mixed mode columns (Aepfelbacher et al., 2008). Examples such as the Dionex Trinity or the PrimeSep columns can retain anionic and/or cationic analytes strongly while also performing conventional RP chromatography (Lamerhofer et al., 2008).

Though LC is the dominant technique for the separation of compounds from complex mixtures there are several alternatives. For volatile compounds such as terpenes or compounds (e.g., acids) that can be derivatized to heat stable analytes GC-MS can be an efficient choice. GC-MS can be beneficial both due to the superior separation (10–100 higher peak capacity) but also due to the reproducible EI ionization that can be employed. Furthermore, technological improvements such as comprehensive GC × GC are becoming more affordable and dedicated TOFMS detectors for GC are being developed delivering accurate mass (<5 ppm accuracy) as well as GC-APCI interfaces for molecular mass identification.

Most recently, significant technological advances have occurred in the previously neglected field of supercritical fluid chromatography (SFC). Improvements have been made to SFC with better columns and much enhanced reproducibility by both Waters and Agilent. Together with novel column systems, many for enantiomeric separation, we foresee that this technique will become very valuable for analysis of SMs from the mid-polar region to the extremely non-polar (e.g., waxes and sterol esters). The technology can now be used with UV detection and is especially well suited for APCI.

**DIODE ARRAY DETECTION**

Ultra high performance liquid chromatography MS is often combined with the relative inexpensive DAD as UV/Vis spectra provide information on chromophores which can be used for database searching (Hansen et al., 2005; Larsen et al., 2005; Bitzer et al., 2007; de la Cruz et al., 2012). More often DAD is used as second criteria after a MS based search, as neither AntiBase nor Dictionary of Natural Products provides a direct UV/Vis search from UV/Vis max data nor do they provide whole spectra. Many secondary metabolites, such as non-reduced PKs and NRP s contain conjugated chromophore systems. As such many metabolites have distinct UV-spectra (Figure 1) that are, in some cases, indicative of the basic skeleton of the given compound class, for example, quinazolines and γ-naphtopyrones produced by many *Aspergilli* and *Penicillia* (Larsen et al., 1999; Nielsen et al., 2009). Consequently, UV/Vis spectral chromophores can provide a means to differentiate compounds with the same elemental composition and can be highly valuable in dereplication for exclusion or confirmation of candidates during a database search. Though often useful, many compounds lack or contain few conjugated double bonds and only giving rise to one absorption band. Examples of compounds without diagnostic UV/Vis spectra include trichodermin, patulin, and deoxynivalenol (Figure 1). In cases such as these the UV/Vis spectrum is of little or no use. Other compounds, such as mycophenolic acid, contain a highly diagnostic chromophore (Figure 1), with three absorption maxima. Despite this informative fingerprint, compounds like mycophenolic acid also require MS verification due to the existence of at least 55 known compounds containing the 5,7-dihydroxymethylphthalide core of mycophenolic acid and four with the cichorein core, as revealed by a substructure search in AntiBase 2012.

Analysis of UV/Vis spectra can also be very powerful for biosynthetic pathway elucidation studies, since compounds which belong to the same pathway often contain the same chemical scaffold and thereby chromophore system. This phenomenon was recently demonstrated during elucidation of the yanuthone biosynthetic pathway (Holm et al., 2014). Furthermore, DAD can be very useful for detection and discovery of compounds of the same type across species and genera, especially in cases where a certain class of compound has shown great promise as a new potential drug lead. Hansen et al. (2005) have developed an algorithm X-hitting for automated comparison of full UV spectra from LC-DAD analysis against a UV-library of standards as well as spectra across samples. This allows for both the identification of known compounds as well as new compounds with UV spectra similar to known compounds (Hansen et al., 2005; Larsen et al., 2005). Altogether DAD is a cheap and complementary spectroscopic technique easily performed in combination with full scan MS. Interestingly a recent software made in the
Nielsen and Larsen Dereplication by LC-MS

FIGURE 1 | UV/Vis spectra (pH 3) and structures of some common fungal natural products illustration the diversity in absorption range depending on the number of conjugated double bonds present in the molecules.

same R environment1 as the popular XC-MS package (vide infra) for LC-MS, is now also available for LC-DaD (Wehrens et al., 2014).

Other (non-MS) detectors which are often used in combination with LC are evaporative light scattering detectors (ELSD; Bitzer et al., 2007; Yang et al., 2014) and the Corona cad detectors. These detectors give more quantitative information about the amount of compounds eluting from the LC unit, in contrast to MS detectors which are highly biased.

MASS SPECTROMETRY

Today all MS instrumentations coupled to LC rely on atmospheric pressure ionization (API) techniques such as electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI), or atmospheric pressure photo ionization (APPI). These are all soft ionization techniques, as such only limited fragmentation of fungal SMs takes place at standard ionization conditions compared to what is seen when using hard ionization such as electron impact (EI) which is used for gas chromatography. Although limited fragmentation is usually observed, the default ion-source settings from many of the LC-MS manufactures, especially for ESI, cause quite extensive fragmentation to occur for very low molecular mass SMs (<300 Da) on most LC-MS systems (Nielsen et al., 2011a; Kiltgaard et al., 2014; Rasmussen et al., 2014). Thus can the task of identifying the molecular mass in an API generated spectra become quite troublesome as a mix of fragment ions adducts and dimeric and double charged ions may be generated (Figure 2; Nielsen et al., 2011a; Kildgaard et al., 2014; Kiltgaard et al., 2014). Consequently, the task of identifying the molecular mass in API generated spectra comes down to interpretation of the adduct patterns that can be seen for the different types of secondary metabolites (Figure 2).

Correct assessment of the molecular mass (M) is often not trivial as in-source fragmentation and adduct formation can lead to incorrect assignment of ions such as [M+Na]+, [M+NH4]+ or [M-H2O+H]+ as [M+H]+ (Nielsen et al., 2011a) or due to the low abundance of the [M+H]+ ion due to poor ionization under the selected conditions. A combination of both ESI+ and ESI− ionization will assign the mass unambiguously, while just one polarity will only assign the mass unambiguously, when several adducts in a given mass spectrum all points toward the same molecular mass (Nielsen and Smedsgaard, 2003; Nielsen et al., 2011a). When the mass, or even better elemental composition, has been determined, this information can be searched against the literature to determine whether the compound is likely known or novel. AntiBase is currently the most suitable database for dereplication of fungal secondary metabolites as it only contains microbial metabolites. The 2014 version contains 43000 recorded compounds, including ~18000 fungal and 21000 bacterial compounds. Dictionary of Natural Products additionally contains a large number of plant metabolites (some of which are produced by endophytic fungi and not the plant themselves), thus this database a valuable addition to AntiBase. The databases SciFinder/CAS and RSC/Chemspider are highly biased toward synthetic compounds, and of less value. Furthermore hits in these databases cannot be sorted by organism type (Lang et al., 2008; Nielsen et al., 2011a). For Marine derived fungi MarineLit (Blunt et al., 2008) may be of interest and now available via RSC/Chemspider2.

In recent years data handling software has become the most crucial component of successful dereplication. Modern UHPLC-MS instruments can provide 50–200 data files per day with MS and tandem MS data, this can take many hours if not days per data file to manually handle, thus mandating more automated dereplication approaches. As with manual dereplication, the indisputable first step in automated dereplication is determination of the elemental composition of the individual compounds in the sample. For this accurate mass is vital. Depending on the mass range and

1http://www.r-project.org/

2http://pubs.rsc.org/marinlit/
FIGURE 2 | Position of ions relative to the monoisotopic mass, of ESI$^+$ (A), ESI$^-$ (B) and EI$^+$ (C) along with adduct ions, and 3 example ESI spectra. In shamixanthone (D) the [M+H]$^+$ ion is absent but the molecular mass can easily be assigned using the Δ40 (39.9925) between the [M+Na]$^+$ and the [M+H$_2$O]$^+$. In (E,F) the big difference of formate adduct intensity on the austinos (same run, elutes 0.4 min in-between) is illustrated. For instrument settings see Kildgaard et al. (2014).

the instrumental accuracy the elemental composition can often be determined unambiguously up to 300 Da at an accuracy of 5 ppm, while 1–2 ppm is required for compounds with masses up to 5–600 Da (Kind and Fiehn, 2007; Nielsen et al., 2011a). When accurate isotope ratio assessment is also reliable, it is possible to eliminate close lying elemental composition candidates with different numbers of carbon atoms (Bitzer et al., 2007; Lehner et al., 2011; Nielsen et al., 2011a; Bueschl et al., 2012; Sleno, 2012; Xian et al., 2012; Nagao et al., 2014).

Today the market is dominated by time of flight (TOF), Orbitrap, and Fourier transform ion cyclotron (FT-ICR) based mass spectrometers. TOF instruments deliver both superior isotopic ratios as well as an unmatched scan speed providing ample time for concurrent MS/MS experiments or the possibility of ultra-fast separation runs. A serious drawback of TOF based instruments is overloading of the detector leading to poor mass accuracy, which was very pronounced on the earlier TDC type detectors. Even for the newest 4 GHz ADC detectors overloading is still a problem and TOF related software needs to limit analysis to scans from the non-overloaded parts of the chromatographic peak (Kildgaard et al., 2014). Overall this results in poorer mass resolution than the two other common types of HRMS instruments. The FT-ICR technology provides an unsurpassed mass accuracy and resolution, for example an FT-ICR can easily resolve the A$+2$ ion of sulfur containing ions into the $^{12}$C$_2$$^{13}$C$_2$$^{32}$S and $^{12}$C$_{2+}$+$^{34}$S isotopomers. Unfortunately FT-ICR instruments suffer from high running costs for magnets, high acquisition price, as well as slow scan capability. The practical limitations of FT-ICR instruments have made Orbitrap type instruments much more popular as they do not require the expensive cooling of the superconducting magnets. In contrast to TOF based instruments Orbitraps do not suffer from the detector overload and needs less frequent mass calibration, though they deliver significantly slower scan speeds. The linear ion-trap (LIT)-orbitrap types can provide MS$^n$ for fragmentation trees due to a dual detector providing both high and low resolution spectra, while Quadrupole (Q)-Exactive provides MS/MS as a QTOF and better quantitative performance than the Lit-Orbitrap, as well as the possibility of positive/negative switching.

Some of the earlier Orbitrap types used an ion-trap for fragmentation and helium as collision gas, which does not provide the same fragment ions as heavier gases. Furthermore, ion-trapping usually has a limited m/z trapping window, a limitation that have been overcome by adding an additional collision cell on newer models (Perry et al., 2008). With respect to MS/MS spectra obtained from QqQ and QTOF instruments, we find the Q-Orbitrap to be superior. A class of lower cost instruments, which can perform MS$^n$, are ion-trap (IT) spectrometers, albeit with low mass accuracy. This is very useful for peptide identification, and for other polymeric molecules made up of known units. In regard to dereplication, the common occurrence of chromatographic peaks indicating an unknown compound renders IT instruments inferior to accurate mass instruments. Triple quadrupole instrumentation have very poor full-scan sensitivity and are cost wise a poor choice for dereplication applications, but highly valuable for environmental analyses where sensitivity is the most important parameter (Sulyok et al., 2006; Sørensen et al., 2009).

For a number of years ion-mobility has been used in a range of configurations, for example, in the Waters Synapt instrument used for differentiating drift times of the fragment ions after MS/MS, however this has limited applications in SM analysis. Recently (May et al., 2014), a much higher resolution ion-mobility interface (positioned between the API source and a QTOF) has
been disclosed and appears extremely promising to resolve co-eluting compounds which often have very different masses and cross sectional areas. Deconvolution of chromatographically co-eluting compounds can occur as the pseudo molecular ions and simple fragment ions (e.g., [M + H-H\_2O]\^+) have almost the same cross-section and thus same drift times. In contrast the co-eluting compounds are likely to have different cross sectional area and different drift times so can then be deconvoluted in the drift time dimension. This will provide a cleaner MS spectrum and more reliable interpretation of the molecular mass. Theoretically, a three dimensional deconvolution using both retention time and drift time can be used, but has to our knowledge not been developed.

**AUTOMATED TARGETED ANALYSIS**

An alternate dereplication strategy is to search LC-MS data files for the masses (preferable accurate masses or elemental compositions) of a list of possible known compounds. This strategy, named aggressive dereplication, has been shown to work well for lists up to 3000 compounds (Kildgaard et al., 2014; Klitgaard et al., 2014), using the software from several manufactures. A major challenge is fragile compounds that easily fragment in the ion-source and thus appear as another elemental composition, leading to erroneous compound identification. The technology can be improved tremendously by introducing pseudo MS/MS in a second scan trace, so the instrument alternates between high and low fragmentation during a run (MS-All, MS-E, All-Ions; Olanpera et al., 2006; Broecker et al., 2011; Guthals et al., 2012; El-Elimat et al., 2013) to obtain compound specific fragment ions known from our own standards, literature data, or from *in-silico* fragmentation (Hill and Mortishire-Smith, 2005; Kangas et al., 2012; Hufsky et al., 2014). The MS-All, MS-E, All-Ions strategy also has the advantage that the [M+Na]\^+ is often enhanced in the fragmentation trace and an accurate confirmation of [M+Na]^+ to [M+H]^+ or [M+H-H\_2O]^+ cuts down on the number of false positives resulting from misidentification of in-source fragment ions (e.g., [M+H-H\_2O]^+ being identified as [M+H]^+ of another compound). Here we have observed that using an algorithm locking the pair like Bruker Target 1.3, Agilent Quant (*vide infra*), or ACD/Spectrus Processor provides 2-10 fold less false positives than Agilent Qual (Kildgaard et al., 2014) or Bruker Target 1.2 (Klitgaard et al., 2014) where one cannot require pairs of [M+Na]^+ and [M+H]^+ or [M+H-H\_2O]^+.

**CONSTRUCTION OF COMPOUND DATABASE**

An important part of targeted analysis is construction of the compound database to be used for searching. Kildgaard et al. (2014) and Klitgaard et al. (2014) used ACD Chemfolder (Advanced Chemistry Development, Toronto, ON, Canada) for construction of a database including: (i) a number of in-house reference standards; (ii) a selection of common fungal compounds from AntiBase; (iii) common impurity compounds known from blank samples or as common media components, and finally (iv) a number of tentatively identified compounds. Compound-database handlers like Chemfolder or Chemfinder have the advantage of the ability to perform substructure searches, which is why these are preferred for a master database. For each compound, major adducts (known or predicted) were registered in the database [41]. If known and/or predicted fragment ions from an alternating broad band fragmentation (MS-E, All-Ions) are included the specificity is greatly increased.

The next step is to create a taxonomically relevant search list (if chemotaxonomic data from properly identified fungal identifications are available). If a species specific search list is not available data from the whole genus may be used depending on the number of compounds described. However, a balance must be found between the number of false positives from compounds with the same elemental composition and with compounds failing to be identified and requiring subsequent manual deconvolution.

Minor in-house programming (e.g., Excel) is required to transfer from a Chemfolder/Chemfinder database to a ready search list for the MS vendor software (Bruker Target analysis (Klitgaard et al., 2014), Agilent Masshunter-find by formula (Kildgaard et al., 2014), or Waters ChromaLynx etc. Different settings for retention time, mass accuracy of peaks, area cut-off, among others can be used to process data-files depending on the nature of the samples.

In some concurrent large screening studies of black aspergilli, we have used a multi-sample screening tool, Agilent MassHunter Quant (similar packages available from Waters and Thermo), developed for LC-QqQ targeted quantification of pesticides, for example. Despite being outside the original design the software can also handle high-resolution data. Thus a target list can be imported via an XML file and one can convert a Chemfolder/Chemfinder database (e.g., [M+H]^+, [\(^{13}C\_2\)M+H]^+, [M+Na]^+, [\(^{13}C\_2\)M+Na]^+, with RT window and fragment ions) to an XML file and import into MassHunter Quant (Figures 3B–D).

The advantage of the multi-sample screening approach is that whole batches of anywhere from 10 to 100s of samples including blanks, fungal strains grown on multiple media can be screened simultaneously. This not only gives fast processing, but a high chance of identifying a compound not previously detected while concurrently identifying the most prodigious producer of the batch. In many cases several peaks of interest are identified but with a proper selection of qualifier ion and comparisons between species and media and blanks usually only 2–3 peaks in the whole batch require further manual inspection. Here the sample with the most intense peak can be selected automatically as it is most likely to provide the best MS/HRMS and possible UV/Vis data. This is illustrated in Figure 3 where a part of the batch of samples is shown, here screening for campyrene C, using [M+H]^+ (240.1230 ± 20 ppm, Figure 3C). Figure 3D shows the qualifier chromatograms, including [M+Na]^+, showing that the compound has the right elemental composition. The compound was then manually verified (Figures 3E–H) through conclusive MS/HRMS spectra and a correct elution order of the A or B isomers.

Altogether, the target analysis approach makes it possible to easily identify chromatographic peaks which are both likely to represent already known compounds and, even more importantly, also peaks that do not correspond to known compounds. Thus the target approach can quickly support prioritization in relation to which compounds might need to be produced in larger scale for...
FIGURE 3 | Aggressive (A–C) batch processing of 150 samples (12 shown), here sorted by peak area (Response column) of mass of [M+H]+ of campyrone C (240.1230 ± 20 ppm) shown in (C). While (D) shows the overlaid extracted ion chromatograms of the same as well as 12C112C12 [M+H]+ isotopomer and the [M+Na]+ ion. (F–I) Manual verification of data file with highest peak intensity of campyrone C. (E) Shows the highly complex Base Peak chromatogram, while, the extracted ion chromatograms of [M+H]+ of campyrone C and A/B respectively. (G) Shows the MS/MS spectrum at 10 eV of the tentatively identified campyrone C with (H) Showing the structure of campyrone C along with (I) theoretical masses of 3 likely fragmentation of the molecule. For instrument settings see Kildgaard et al. (2014).

PEAK-FINDING

A more classical approach to investigation of a full scan UHPLC-HRMS file is to use a peak finding algorithm based on molecular features. This approach uses deconvolution of the time profiles and adds peaks together and possibly resolves the adduct pattern (Kuhl et al., 2012; Kildgaard et al., 2014). Most vendor software as well as the open source software XC-MS (Smith et al., 2006) with the Camera package (Kuhl et al., 2012) offers this feature although not all can search Chemfolder/Chemfinder Structure Data Format (SDF). However they can all search ChemSpider and similar public databases. This approach is probably best suited to extracts from species where the taxonomy is not well known
Nielsen and Larsen Dereplication by LC-MS and can be used in a true unbiased metabolomics workflow (Cox et al., 2014; Macintyre et al., 2014; Wolfender et al., 2015), and can be especially useful when combined with bioactivity or gene knock-outs.

A major obstacle for this method is that all samples to be compared need to be processed at the same time, as changes in media batches, impurities in solvents, filters, plastic and glassware (often strongly ionizing impurities) strongly influences data analysis. In addition to this there are problems associated with the likelihood of changes in secondary metabolism which are often seen for many organisms, even though great efforts are made to standardize cultivation conditions. Finally analytical separation, cleanliness of ion-sources, changes in LC-MS solvents and many other things also may cause changes in data. One can standardize data between batches but this demands an enormous amount of quality assurance work as it needs to include at least all the variable things mentioned above and is likely to still result in large uncertainties.

Thus, if using a metabolomics approach to find small peaks in extracts, a very strong experimental design, with 4–6 replicates per condition is needed, and ideally all samples need to be processed and analyzed in a single batch. If this strict control is not implemented the peak picking algorithm and unbiased data analysis will likely find the sample preparation day as the most significant parameter, especially when lowering the intensity threshold. In such cases 5,000–30,000 chemical features (Cox et al., 2014; Macintyre et al., 2014; Wolfender et al., 2015) may be detected and statistically some will always be correlated to a given hypothesis. Despite these shortcomings, for extracts with very high media background and very low signal intensity from the secondary metabolites this approach is the most effective. In our view this approach is highly time-consuming and if searching for a bioactive compound, an assay guided approach may be far less time consuming.

The metabolomics workflow is often not easily implementable with the workflow in a natural product laboratory as different species and strains often need substantial optimization in terms of growth and analysis (Figure 4). However, for gene-to-product linking type of projects, it can be the only way to search for low intensity peaks in defined studies, as one cannot add samples from later studies.

API MS/MS LIBRARIES
As mentioned above, UHPLC-HRMS is often not enough for a positive identification and a library based identification – as is possible for El+ fragmentation – would be highly desirable. If matching LC-MS generated MS spectra to a library they must be reproducible to give reliable matches. This is not possible with full scan MS due to the differences in in-source fragmentation settings and adduct formation and spectra of the same compound can be very different even on the same instrument. However MS/MS spectra of the pseudomolecular ions, in general, produces the same fragment ions, although the fragmentation energy to provide the same spectrum may be very different and also will vary between the harder argon toward the softer nitrogen (Baumann et al., 2000; Fredenhagen et al., 2005; Lee et al., 2005; Pavlic et al., 2006; Oberacher et al., 2012; El-Elimat et al., 2013; Kildgaard et al., 2014).

Thus an additional scan type, MS/MS, can be used simultaneously with full-scan, when using a QTOF, IT-Orbi-Trap, Q-Orbi-Trap, as these can perform auto-MS/MS, or data dependent MS/MS. While TOF and Orbi-trap (Excactive) instruments cannot isolate an ion for MS/MS, they can perform broad band excitation (All-Ions or MS-E) fragmentation at some point in the ion-optics. A corollary is that fragmentation in an ion-trap often provides very different fragments than a collision cell as fragment ions that are outside the excitation window in the ion-trap are no longer accelerated and thus fragmented. To remedy this shortcoming some Orbi-Trap type spectrometers come with an additional hexa- or octapole collision cell.

Compared to forensic science and toxicology (Broecker et al., 2011) only few in-house MS/MS libraries with fungal metabolite...
spectroscopic data are available thus far (Kildgaard et al., 2014). The major reasons are firstly; the lack of requirement to publish MS/HRMS spectra and secondly; the lack of standardization of fragmentation energies between instrument manufacturers. Although only containing relatively few microbial natural products, the two libraries that can be helpful, to identify lipids, medium polar primary metabolites, vitamins, and other compounds are Massbank (Horai et al., 2010) and Metlin (Smith et al., 2005; ~10,000 compounds with spectra).

In addition it is unreliable to use in silico predictors for prediction of fragmentation of NPs, since they often represent highly condensed and complex ring structures, only leaving room for verification of some fragments from a structure in a spectrum (Hill and Mortishire-Smith, 2005; Hufsky et al., 2014). Despite of these challenges we anticipate that the use of MS/HRMS spectral libraries will become much more pronounced in the near future, especially since the different vendors can see a huge advantage in being able to supply their customers with dedicated spectral libraries.

PRECURSOR SELECTION
The use of MS/MS on one of the pseudomolecular ions does indeed already provide very reproducible MS/MS spectra within one brand of instrument (Fredenhagen et al., 2005) as recently demonstrated in our laboratory (Fredenhagen et al., 2005; Kildgaard et al., 2014). Thus, good reproducibility can be achieved if the collision energy is calibrated and the nature of metal ion adducts are known, since [M+Na]⁺ and [M+K]⁺ fragment far less and with different mechanisms than [M+H]⁺, [M+NH₄]⁺ (Oberacher et al., 2009, 2011; Wurtinger and Oberacher, 2012; Kildgaard et al., 2014). For robustness, many of the known adduct ions should be included as it increases the identification confidence especially between related compounds (Kildgaard et al., 2014).

FRAGMENTATION
Importantly, different energies for MS/MS are needed as the stability of compounds varies significantly (Kildgaard et al., 2014). This can be acquired in two ways: (i) as several distinct energies (Smith et al., 2005), or ii) by ramping the collision energy, thus acquiring an average spectrum. For fungal metabolites we have found the spectral quality (Kildgaard et al., 2014) to be much higher when using distinct energies, especially when using accurate mass of the fragment ions. In such cases only 3–7 fragment ions are needed, as long as they are distributed equally over the mass range (neither in the very low range, with common fragments shared with many other compounds, nor close to the molecular ion where the losses may not differentiate from related compounds). Agilent Technologies and Metlin have chosen to acquire spectra at the three different fragmentation energies 10, 20, and 40 eV (Kildgaard et al., 2014), which we also found efficient for microbial metabolites, except for large peptides where 60 eV had to be included, and we suggest a possibility to change the window for these when going to larger masses where much more energy is generally needed.

FIGURE 5 | UHPLC-DAD-HRMS-MS/HRMS analysis of an Aspergillus tubingensis extract. (A) shows the BasePeak chromatogram. Peaks automatically assigned from Auto-MS/MS data searched in the MS/HRMS database are colored. (B) shows the UV/Vis chromatogram (200–640 nm), while (C) shows the MS/HRMS spectrum of nigerazin A (20 eV) or B used to tentatively ID this peak. (D) and (E) shows the MS/HRMS (20 eV) and UV/Vis spectra of asnipyrone B, respectively, with the UV/Vis spectrum matching literature data perfectly, and also eluting before asnipyrone A (having an extra methyl group). For instrument settings see Kildgaard et al. (2014).
Different algorithms have been used to search experimental MS/MS spectra contained in small in-house MS/MS libraries for tentative identification of fungal SMs. For example the NIST (National Institute of Standards and Technology) algorithm, developed for full scan EI$^+$ spectra and the Mass Frontier software for MS$^3$ spectra were compared by Fredenhagen et al. (2005) to search low resolution MS/MS data, with the latter found to be superior. Similarly El-Elimat et al. (2013) used ACD-IntelliXtract, which allows inclusion of accurate masses of the fragments, but does not use the parent ion data as search entry [64]. We use the Agilent search algorithm that is an integral part of the Agilent MassHunter software for fast and automated search in our in-house MS/HRMS library of more than 1300 compounds for unambiguous identification of especially fungal metabolites belonging species in the genera Aspergillus, Penicillium, and Fusarium (Figure 5).

The software allows background subtraction and merging of spectra over chromatographic peaks into a single spectrum prior to automatic searching against the library. Importantly, searching of MS/HRMS spectra against a given in-house library in MassHunter allows for both forward and reverse scoring using the parent mass for matching of peaks in the unknown spectrum against the library spectra or vice versa. Often both forward and reverse scoring are needed for correct identification as shown in our work for the identification of patulin (Kildgaard et al., 2014). Part of our library (277 microbial compounds) can be downloaded as PCDL format from the homepage of the Technical University of Denmark (Nielsen and Frisvad, 2014).

Manual interpretation of MS/HRMS spectra might even be used to predict the structure of unknown compounds based on careful inspection of the obtained fragmentation patterns. This approach is especially applicable for NRPs due to their sequential composition of amino acids (Figure 6).

For ion-traps users (especially IT-Orbi trap) the so called MS-trees offers an additional dimension to identification of a substructure of an analyte molecule by MS/MS/MS (MS$^3$; Riddor et al., 2012; Rojas-Cherto et al., 2012), as fragment ions from conjugation with the carbon can occur. (C) shows other novel emericellamides G ($R_1 = H$, $R_2 = H$) presumably with the fatty chain $2 \times \text{CH}_2$ longer. (H), emericellamide B; (G), emericellamide H (mericellamide E, with $R_1 = \text{CH}_3$, $R_2 = \text{CH}_3$); (F), emericellamide F; (E), emericellamide A; and (D), emericellamide C. For instrument settings see Kildgaard et al. (2014).
the first MS/MS event subsequently can be further fragmented and matched in a MS/MS library for assigning parts of a given molecular structure.

**MS NETWORKING**

In recent years the Dorrestein and Bandeira labs (Guthals et al., 2012; Watrous et al., 2012) have been key drivers toward development of new networking MS/MS approaches. Here MS/MS spectra of compounds in a given sample are compared pairwise and structurally related compounds are clustered based on the presence of similar fragments and neutral loss. In this way several compounds belonging to the same pathways have very convincingly been linked together including compounds which belong to both known and novel biosynthetic pathways (Watrous et al., 2012; Yang et al., 2013). A major current drawback is the lack of back-integration of raw data for evaluation of the corresponding full scan data and retention time, making detailed analysis extremely time consuming. Also source code is not available in an open-source format as, e.g., XCMS and it is thus impossible to troubleshoot data in cases where the analysis fails.

**ISOTOPE LABELING FOR STRUCTURE AND BIOSYNTHESIS INVESTIGATION**

The availability of LC-MS instruments and increased availability of stable isotope (13C, 2H, 15N, 34S) labeled substrates has made more reliable determination of the elemental composition of a given compound more robust through determination of the number of carbon (Bueschl et al., 2012), nitrogen (Bode et al., 2012) or sulfur atoms in the molecule (Brock et al., 2014). However, the technology is also promising for incorporation of precursors involved in PK and NRP biosynthetic studies (McIntyre et al., 1982, 1989; Townsend and Christensen, 1983; Steyn et al., 1984; Simpson, 1998; Baran et al., 2010; Bode et al., 2012; Fuchs et al., 2012; Proschak et al., 2013; Brock et al., 2014; Huang et al., 2014), as recently exploited for proving that yanuthone production in A. niger is produced via 6-methylsalisylic acid and not shikimic acid (Holm et al., 2014; Petersen et al., 2015).

**SURFACE TECHNIQUES**

During the last 7 years surface desorption techniques have been introduced for secondary metabolite screening. These techniques can both ionize and detect surface excreted compounds and compounds in agar, for example, the identification of compounds present between colonies of different organisms for studying interactions. Current methods can be separated into two different sub families: desorption electrospray ionization mass spectrometry (DESI), direct analysis in real time (DART) and variants of them, and they are all soft ionization techniques (Watrous et al., 2012; Hsu et al., 2013; Bouslimani et al., 2014). DESI (Esquenazi et al., 2009) is based on charged droplets of organic solvent and a gas jet hitting the surface and then sampling ions close by (Gur-dak et al., 2014), whereas DART uses a gas ionized by a plasma. Both have a poor spatial resolution (mm range) and can only be applied for major structures as DESI imprint of exudate droplets (Figueroa et al., 2014). Another major field of surface desorption techniques are the matrix assisted laser desorption techniques (MALDI). MALDI has a much higher spatial resolution (down to several 100 μm), but still not low enough to provide cellular resolution of fungi (Watrous and Dorrestein, 2011; Watrous et al., 2012; Bouslimani et al., 2014).

**CONCLUSION AND PERSPECTIVES**

The recent advancements in separation sciences, high resolution mass spectrometry, as well as data mining tools have dramatically improved our possibilities for dereplication of fungal natural products in complex mixtures. At the same time these advances call for more automated methods for analysis since the amounts of data that can be generated on a single instrument in a short time is truly immense. Among other techniques, dereplication based on auto-MS/MS has proven to be very robust and effective on a given instrument. However, the use of MS/MS settings between manufacturers and scoring algorithms are calling for standarized methods for both generation and comparison of MS/MS patterns.

The ultimate solution for future dereplication would be the construction of an open natural product database, where all relevant chemical and taxonomic information have been merged together. In particular this should include chromatographic and spectroscopic data (HR MS/MS, UV/Vis, NMR) as well as valid taxonomic information about the source organism, in many cases maybe even including linking biosynthetic origin of a given natural product to the related gene cluster.

**ACKNOWLEDGMENTS**

We are grateful to Agilent Technologies for the Thought Leader Donation of the UHPLC-QTOFMS system, and linguistic corrections by Christopher Phippen. Funding is acknowledged from the Novo Nordic Foundation and from PharmaSea (EEC, Grant Agreement No 312184).

**REFERENCES**


Copyright © 2015 Nielsen and Larsen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.