



# JOURNAL OF INTEGRATED OMICS

A METHODOLOGICAL JOURNAL

HTTP://WWW.JIOMICS.COM



ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v3i2.139

## Revisiting Protocols for the NMR Analysis of Bacterial Metabolomes

Steven Halouska<sup>1,†</sup>, Bo Zhang<sup>1,†</sup>, Rosmarie Gaupp<sup>2</sup>, Shulei Lei<sup>1</sup>, Emily Snell<sup>1</sup>, Robert J. Fenton<sup>2</sup>, Raul G. Barletta<sup>2</sup>, Greg A. Somerville<sup>2</sup>, Robert Powers<sup>1,\*</sup>

<sup>1</sup>Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304. <sup>2</sup>School of Veterinary Medicine and Biomedical Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588-0905. <sup>†</sup>Equal contribution.

Received: 07 May 2013 Accepted: 09 July 2013 Available Online: 17 July 2013

### ABSTRACT

Over the past decade, metabolomics has emerged as an important technique for systems biology. Measuring all the metabolites in a biological system provides an invaluable source of information to explore various cellular processes, and to investigate the impact of environmental factors and genetic modifications. Nuclear magnetic resonance (NMR) spectroscopy is an important method routinely employed in metabolomics. NMR provides comprehensive structural and quantitative information useful for metabolomics fingerprinting, chemometric analysis, metabolite identification and metabolic pathway construction. A successful metabolomics study relies on proper experimental protocols for the collection, handling, processing and analysis of metabolomics data. Critically, these protocols should eliminate or avoid biologically-irrelevant changes to the metabolome. We provide a comprehensive description of our NMR-based metabolomics procedures optimized for the analysis of bacterial metabolomes. The technical details described within this manuscript should provide a useful guide to reliably apply our NMR-based metabolomics methodology to systems biology studies.

**Keywords:** NMR; metabolomics; chemometrics; Mycobacterium; Staphylococcus; bacteria.

### Abbreviations

NMR, Nuclear Magnetic Resonance; MS, Mass Spectrometry; LC, Liquid Chromatography; GC, Gas Chromatography; CE, Capillary Electrophoresis; PCA, Principal Component Analysis; OPLS-DA, Orthogonal Projection to Latent Structures Discriminant Analysis; HSQC, Heteronuclear Single Quantum Coherence; TOCSY, Total Correlated Spectroscopy; TCA, Tricarboxylic Acid; TSB, Tryptic Soy Broth; MADC, Middlebrook Albumin Dextrose Catalase; CFU, Colony Forming Unit; MIC<sub>50</sub>, Minimum Inhibitory Concentration required to inhibit the growth of 50% of organisms; SOGGY, Solvent Optimized Gradient-Gradient Spectroscopy; HSQC<sub>0</sub>, Time-Zero HSQC; TMSP-d<sub>4</sub>, 3-(trimethylsilyl)propionic acid-2,2,3,3-d<sub>4</sub>.

### 1. Introduction

Metabolomics is the study of small molecules in a biological system that participates in the metabolic reactions responsible for cell growth, survival, and other normal cellular functions [1-3]. Additionally, the metabolome responds to transcriptional and translational alterations associated with genotypical, epigenetic, or environmental perturbations [4-7]. Thus, metabolomics provides an assessment of global perturbations with respect to an altered genome, proteome, or environment [2, 8, 9]. The simultaneous integration of

genomic, transcriptomic and proteomic data has enabled an in-depth analysis of the interplay, interaction, and regulation of DNA, RNA and proteins [10-12]. Along this line, monitoring the bacterial metabolome and integrating the results with other “omics” data has provided valuable insights into bacterial adaptability [13], biofilms [14], evolution [15], pathogenesis [16], and drug resistance [17].

Depending on the organism and growth state, the total number of metabolites within a cell varies between several

\*Corresponding author: Robert Powers, Department of Chemistry, 722 Hamilton Hall, University of Nebraska-Lincoln, Lincoln, NE 68588-0304, Tel: (402) 472-3039; Fax (402) 472-9402. Email Address: rpowers3@unl.edu

hundred to a few thousand, with a corresponding diversity in physical and chemical properties, such as size, stability, and concentration [18]. In addition to the challenge of the simultaneous study of all the metabolites within a given biological system [19], the selection of an analytical technique will influence which metabolites are observed. NMR and MS are commonly employed for metabolomics, where both instruments can be interfaced with LC, GC, and CE systems to select and emphasize specific components of the metabolome [20-24]. NMR has a number of advantages in analyzing the metabolome that includes minimal sample handling and that it is not reliant on chromatography to purify or separate metabolites. In addition, multiple resonances from a single molecule increase the accuracy of metabolite identification and quantitation. This accuracy can be further enhanced by the application of  $^{13}\text{C}$ - and  $^{15}\text{N}$ -isotope labeling to enhance specific regions of the metabolome [25, 26]. Importantly, the choice of  $^{13}\text{C}$ - or  $^{15}\text{N}$ -labeled metabolite determines the region of the metabolome observed by NMR, providing significant flexibility in experimental design. In contrast to MS, NMR is a relatively insensitive technique and only observes the most abundant ( $\geq 1$  to  $5\ \mu\text{M}$ ) metabolites. In addition, MS has the advantage of detecting a wider-range of the metabolome. However, because of the relatively low molecular-weight range of the metabolome, MS methods generally require chromatography to separate metabolites before analysis [27]. Additionally, variations in ionization and the occurrence of ion suppression in a complex mixture add uncertainty in detecting specific metabolites by MS [28]. Finally, quantitation by MS is typically more challenging than NMR. Taken together, NMR and MS each have strengths and weaknesses but should be viewed as complementary techniques [29].

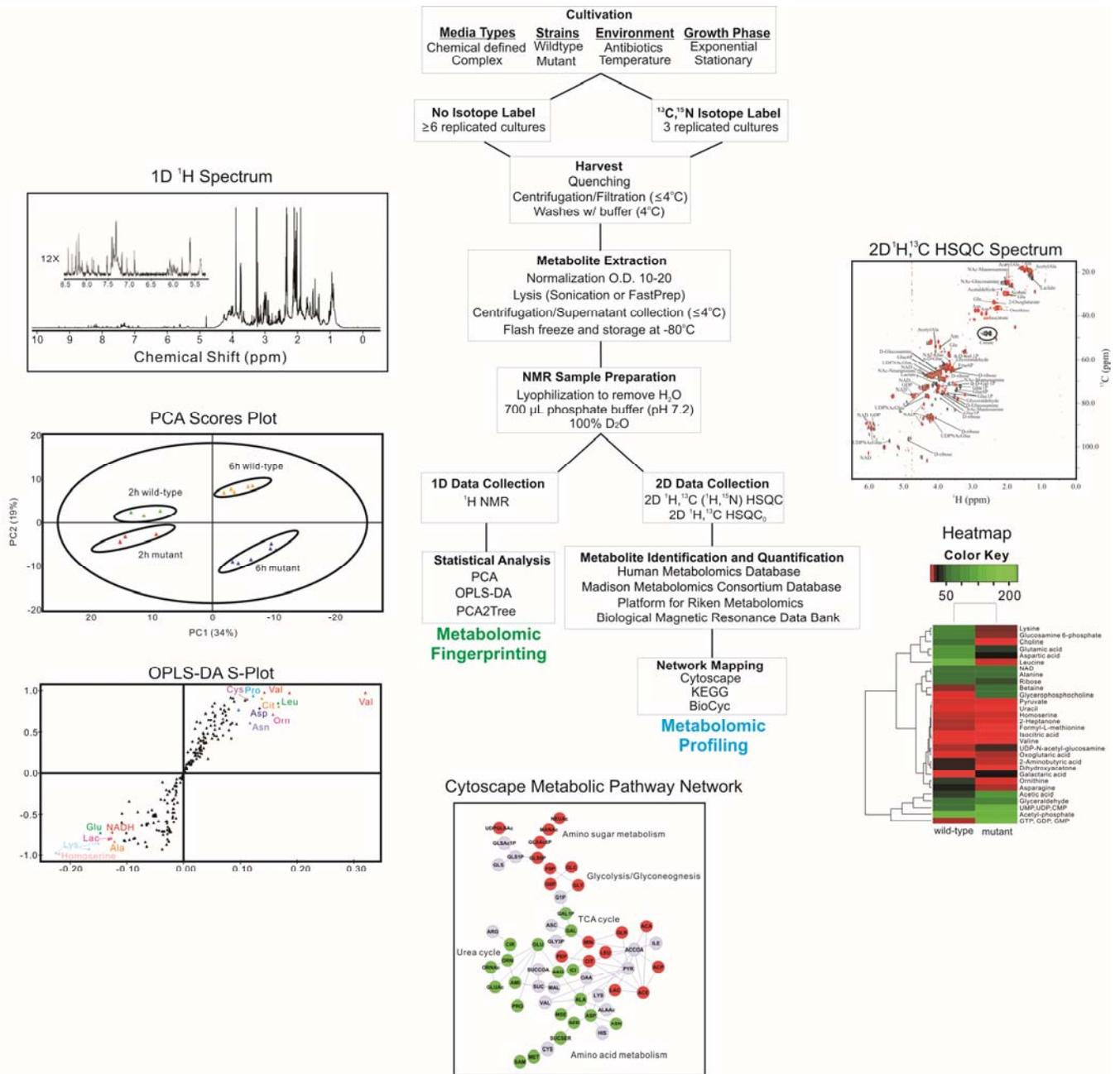
NMR-based metabolomics have been used to study a wide range of biological systems such as tissues [30], biofluids [31], mammalian cell cultures [32], plants [33] and bacteria [34-36]. The overall procedure for an NMR-based metabolomics study includes the following general steps: cell growth and harvesting, metabolite extraction, NMR data collection and analysis, multivariate statistical analysis, metabolite identification and quantification [37]. Typically, one-dimensional (1D)  $^1\text{H}$  NMR spectra are used for a multivariate analysis such as principal component analysis (PCA) or orthogonal projection to latent structures discriminant analysis (OPLS-DA) [38, 39]. Both PCA and OPLS-DA provide global profiles of metabolome changes [40, 41]. Two-dimensional (2D)  $^1\text{H}$ ,  $^{13}\text{C}$  Heteronuclear Single Quantum Coherence (HSQC) or  $^1\text{H}$ ,  $^1\text{H}$  Total Correlated Spectroscopy (TOCSY) NMR experiments are used for the quantitative assessment of metabolite changes resulting from genetic modification or external stimuli [5, 14]. The ability to generate global profiles and quantitative differences coupled with the ease of applying NMR-based metabolomics has contributed to the rapid growth of the NMR metabolomics field. While NMR data acquisition and analysis methods are

improving, care must be taken to ensure that the methods are appropriate to the task at hand and generate biologically relevant information. As an example, protocols to efficiently extract metabolites without inducing cellular changes are essential for success [32, 42]. In brief, the observed changes in the metabolome should reflect a change in the state of the system instead of how the cells are handled and processed. Similarly, variations in instrument performance, choice of procedures for data collection and processing, and invalidated models from multivariate analysis may induce unintended biases or incorrect interpretation of metabolomics data [43-46].

Since NMR-based metabolomics is a relatively new and still developing technology, improving and enhancing the experimental protocols is necessary to advance the field and ensure continued success. Toward this end, we describe our recently developed and optimized protocols for the application of NMR metabolomics to microbial samples. We present our current methodology and also discuss the challenges associated with each major step of the process: (i) sample preparation, (ii) NMR data collection and processing, (iii) multivariate statistical analysis, (vi) metabolite identification and network generation. Specifically, the overall methodology will be discussed in detail, where a number of key features will also be highlighted, such as automation, bioinformatics, experimental design, and harvesting the metabolome. The focus of our efforts has been to identify and minimize procedural steps that negatively influence the outcome of an NMR-based metabolomics experiment.

## 2. Experimental Design

A general protocol for the analysis of bacterial metabolomes using NMR is shown in Figure 1. The flow diagram illustrates procedures for both a global analysis of metabolome changes (metabolomics fingerprinting); and the identification and quantitation of specific metabolites correlated with the biological process (metabolomics profiling). The overall process consists of the following key steps: bacterial cultivation and harvesting, metabolite extraction, NMR data collection and analysis, multivariate statistical analysis, metabolite identification and quantification. Successful metabolomics sample preparation involves three steps. The first step is the simultaneous growth of all of the bacterial cultures or as many as is practical at a time. The bacteria are grown in a standard medium for fingerprint analysis, whereas the medium is supplemented with a  $^{13}\text{C}$ -labeled metabolite for metabolomics profiling [47, 48]. After the bacteria are grown for a defined time or they have achieved a specified cell density, the second sample preparation step involves harvesting the bacteria, quenching to halt all enzymatic processes, and washing to remove the medium. The third sample preparation step involves lysing the cells and extracting the metabolome. A variety of solvents are routinely employed depending on the solubility of the targeted metabolites (cytosolic



**Figure 1.** A flow chart of our protocol used for the NMR analysis of bacterial metabolomes.

metabolites, lipids, *etc.*). The metabolomics samples are then used to generate a series of NMR spectra, which are used for the multivariate statistical analysis, metabolite identification and quantification. The individual steps of the NMR-based metabolomics protocol will be discussed in detail highlighting challenges associated with each step.

*2.1 Identify the appropriate biological system for a metabolomics study*

NMR-based metabolomics is an important tool in systems biology research. The quantitative and qualitative measure-

ment of metabolites from cytosolic extracts can be extremely valuable for investigating cellular processes, pathogenesis, and the effects of drugs or the environment on bacteria. Unfortunately, the bacterial metabolome is a complex mixture of metabolites and numerous interconnected metabolic and signaling pathways. This high interconnectivity may result in significant metabolite concentration changes far from the origin of the perturbation (inhibited, inactivated or down-regulated protein). Correspondingly, it is easier to observe changes to the metabolome than deduce the primary source of the perturbation after its impact has rippled throughout the metabolome. As an illustration, treating a bacterial cul-

ture with a particular drug would be expected to lead to a global change in the metabolome, but interpreting these changes to identify the therapeutic target is extremely challenging. To address this challenge, the *in vivo* mechanism of action of a potential drug lead may be determined by comparing these metabolome changes to other drugs with known biological targets [49] or to a mutant bacteria where a specific protein target is ablated or modified by genetic inactivation [50, 51]. This example illustrates that the comparative analysis between two or more metabolomes is an effective application of metabolomics. In order to obtain reliable insights into the physiology of bacteria or any other organism, it is essential to identify and establish at least two reference metabolomes (wild-type vs. mutant, drug-resistant vs. drug susceptible, nutrient-rich vs. nutrient-limited, *etc.*) for a comparative analysis. Once the reference conditions are established, bacteria can be exposed to any range of experimental variables such as a drug treatment, environmental stimuli (pH, temperature, nutrient change), or gene knock-out (mutants, RNAi, inhibitor) to determine if similarities exist with the reference metabolome. The similarity between metabolomes infers an overlap in the underlying physiological processes or responses that gave rise to the metabolome changes. We have used this approach to demonstrate the similarity of *Staphylococcus epidermidis* metabolomes resulting from exposure to divergent environmental stressors that are known to facilitate biofilm formation [5, 14, 52]. Our results suggested that the tricarboxylic acid (TCA) cycle acts as a metabolic signaling network to transduce external stresses into internal metabolic signals. This conclusion was only possible because the experimental design was based on comparing the metabolomes of the *S. epidermidis* wild-type strain 1457 and an aconitase mutant strain 1457-*acnA::tetM* with and without the treatment of biofilm stressors. In summary, the successful outcome of a metabolomics study hinges on the experimental design and the proper choice of the cellular metabolomes to be compared.

## 2.2 Minimize unintended bias and biologically irrelevant variations

In addition to the proper choice of bacterial strains to compare in a metabolomics study, the experimental protocols must be optimized to reduce unwanted variation or bias in the collection of cell-free lysates. It is essential to ensure that any metabolome changes are limited to biologically relevant factors and are not caused by the handling or processing of the samples. Thus, the key to metabolomics is establishing an efficient methodology that closely captures the true state of the metabolome [53, 54]. Fundamental to a successful metabolomics experiment is maximizing the uniformity of the preparation, handling, processing, and analysis of each replicate sample [35, 45, 55-58]. In instances where cultivation and/or processing variation is unavoidable (*e.g.*, if multiple incubators are required to accommodate all the replicates), then the

cultures should be randomly distributed between the incubators to minimize bias. Ideally, all of the metabolomics samples should be handled by the same person because subtle differences in individual techniques may influence the outcome. If multiple investigators are required to efficiently handle the samples, each researcher should be assigned a specific set of tasks that are consistently applied to each sample. For example, one investigator lyses all the bacterial cells while another performs the metabolome extraction procedure on every sample.

## 2.3 Optimization of the number of replicates to maximize statistical significance

As with sample cultivation and preparation methods, the NMR spectra generated from metabolomics samples need to accurately represent the state of the system. In other words, the NMR spectra must reflect the actual concentrations and identity of the metabolites present in the biological sample at the time of harvest. If the sample preparation and data acquisition represent the metabolic status at the time of harvest, then multivariate statistical techniques, such as PCA and OPLS-DA, will enhance the identification of similarities or differences in the NMR spectra, and, correspondingly, between the bacterial metabolomes [39]. These multivariate statistical techniques typically involve multiple replicates of 1D  $^1\text{H}$  NMR or 2D  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC spectra for each bacterial class or group (*e.g.*, wild-type, mutant, drug treated, *etc.*). The exact number of replicates is dependent on a number of factors: (i) the variance within a group, (ii) the variance between groups, (iii) the number of variables, and (iv) type of statistical analysis performed [59, 60].

In most metabolomics experiments, the number of biological samples is significantly smaller than the number of variables; in this case, the variables correspond to peaks in the NMR spectra or the detectable metabolites [60]. For this reason, a larger number of replicates ( $\geq 6$ ) per class are required to obtain a statistically significant PCA or OPLS-DA model. While greater numbers of replicates are desirable, there are practical considerations to increasing the number of replicates, including increased experimental time, availability of incubator space, and practical limits on the number of samples that can be simultaneously prepared and processed within a reasonable time frame. The increased time, larger number of samples, and added complexity may be detrimental to maintaining consistency between samples, where metabolite stability may become more of an issue [61]. So the potential benefit in improving the reliability of the PCA or OPLS-DA models may be negated by too large of a sample size if sample consistency is sacrificed. In general, 6 to 10 replicates per class can be routinely handled while providing a statistically significant PCA or OPLS-DA model. Lastly, to increase the sample consistency, the application of an automated sample changer or flow-probe can minimize variability by eliminating human involvement and providing a uniform and consistent protocol for NMR data collection.

### 3. Sample Preparation

#### 3.1 Techniques for consistent bacterial cultivation

Consistency is critical to metabolomics, where variations in a bacterial metabolome may be introduced by cultivation protocols. To achieve the reproducible cultivation of bacteria requires consideration of three variables: bacterial strain, culture medium, and cultivation conditions. Strain selection is often driven by investigator preference, availability, or cultivability. The choice of culture medium will largely depend on which, if any, isotopically-labeled metabolite is being followed. For example, when using  $^{15}\text{N}$ -arginine, it is impractical to add labeled arginine to a complex medium containing an unknown concentration of unlabeled arginine. In this example, to achieve maximal labeling of the bacteria, it would be best to use a chemically defined medium lacking arginine. Importantly, the culture medium has to be consistently employed throughout a specific metabolomics study of a defined set of bacterial strains. A different culture medium cannot be used for metabolomics fingerprinting and profiling, it cannot vary based on the requirements of the bacterial strain or to accommodate an experimental variable. For example, if a mutant bacterial strain requires the addition of a supplement for viability, then it is also necessary to add the same amount of the supplement to all other bacterial cultures in the study. Simply, any difference in the composition of the culture media will induce changes in the metabolome that will mask or complicate any analysis. Of course, the culture medium needs to be optimized for the specific requirements of each species and bacterial strain and, correspondingly, will likely vary between metabolomic studies. A metabolomics study employing *Staphylococcus epidermidis* will use a different culture medium than a study involving *Mycobacterium smegmatis*.

Bacterial cultures also need to be properly handled in order to avoid inducing biologically irrelevant changes. For example, pre-warming the culture medium prevents temperature shock and minimizes variation between biological replicates. Similarly, randomizing the samples from each group and class also minimizes bias that may occur if all the samples are processed in a predefined order. Importantly, different cell types may require special care or different handling protocols. Cultivation conditions will also vary depending upon the experiment; however, consideration must be given to each of the following: temperature, pH buffering (if used), %  $\text{CO}_2$  (if used), the flask-to-medium ratio, the revolutions per minute of agitation (if used), the use of baffled or non-baffled flasks, and the inoculum dose. In effect, one protocol does not necessarily “fit all” and a general metabolomics protocol needs to be optimized for each experiment and cell type.

#### 3.2 Sample optimization to maximize NMR sensitivity

Identifying an optimal culture size is an important next

step in the design of a metabolomics study. The volume of the bacterial culture should be large enough to provide a sufficient number of cells to maximize the NMR signal-to-noise, but small enough to simplify the handling of numerous replicate samples. An appropriate cell density must be determined empirically for each species and bacterial strain, which will also limit the culture size. Similarly, the growth phase chosen for harvesting bacteria will also contribute to defining the optimal culture size since cell density changes drastically between the lag, exponential and stationary phases. In our experience with staphylococcal and mycobacteria cultures, media volumes between 15 to 50 mL are used to grow cells to an optical density at 600 nm ( $\text{O.D.}_{600\text{nm}}$ ) of 1-2 for bacterial cultures collected during the exponential phase. Conversely, media volumes of between 3 to 5 mL are used to grow cells to an  $\text{O.D.}_{600\text{nm}}$  of 3 to 7 for bacterial cultures collected during the stationary phase (e.g., 6 to 7 for *Staphylococcus epidermidis*, and 3 to 4 for *Mycobacterium smegmatis*). The overall goal is to have an  $\text{O.D.}_{600\text{nm}}$  of 10 to 20 after the bacterial cells have been concentrated to a final volume of 1 mL. This will ensure metabolite concentrations sufficient for detection by NMR. These culture volumes and  $\text{O.D.}_{600\text{nm}}$  values should be viewed as guidelines and targeted goals that may require further optimization for different bacterial strains or species.

Ideally, each bacterial culture should contain the same number of cells and be at the same growth phase when harvested. In reality, differences in cultivation conditions, media, and/or bacterial strains may substantially affect growth rates and/or growth yields. The two more common approaches to compensate for different bacterial growth rates are: collect the bacteria when they have reached the same cell density, but at different times to account for the different growth rates; and harvest the bacteria at the same time but harvest equivalent cell numbers. The number of bacterial cells in a given culture is estimated by measuring the culture turbidity using a standard optical density method. As examples, in staphylococci, the exponential and stationary growth phases were typically analyzed at the 2 h and 6 h time points, respectively [5]. For our mycobacterial experiments, a consistent growth phase was achieved by harvesting bacteria at a uniform  $\text{O.D.}_{600\text{nm}}$  of 1.2. In practice, it is extremely difficult to harvest every bacterial culture with an identical  $\text{O.D.}_{600\text{nm}}$  value. To correct for this variability, all the bacterial cultures are normalized to the same  $\text{O.D.}_{600\text{nm}}$  value. Simply, the cultures are suspended into a phosphate buffer until the  $\text{O.D.}_{600\text{nm}}$  values are equal. Alternatively, the bacterial cell cultures can be normalized based on colony-forming units (CFU), if OD-CFU calibration curves are available, or total protein concentration.

#### 3.3 Sample preparation protocols to maximize isotope labeling efficiency

Metabolomics profiling requires  $^{13}\text{C}$ - or  $^{15}\text{N}$ -labeled metabolites and defines the choice of culture media. In our la-

laboratories, we typically label staphylococci using  $^{13}\text{C}$ -glucose in the complex medium tryptic soy broth (TSB) that is devoid of unlabeled glucose [6, 7]. This medium allows for maximal biomass generation, while assuring that nearly all (~99%; 1.1% is due to naturally occurring  $^{13}\text{C}$ ) of the  $^{13}\text{C}$ -labeled metabolites in the metabolome were derived from glucose. Similarly, we have labeled mycobacteria using  $^{13}\text{C}$ -glucose or  $^{13}\text{C}$ -glycerol in Middlebrook 7H9 Albumin Dextrose Complex (MADC; Becton-Dickinson) media. We have also supplemented culture media with  $^{13}\text{C}$ -alanine,  $^{13}\text{C}$ -aspartate,  $^{13}\text{C}$ -glutamate,  $^{13}\text{C}$ -proline and  $^{13}\text{C}$ -pyruvate as a more targeted approach to the analysis of the metabolome. These metabolites are associated with a limited number of metabolic pathways. The analysis of the metabolome can be further focused by using a targeted metabolite where only one or a few specific carbons in the metabolite are  $^{13}\text{C}$  labeled. Only the metabolic pathways involving the specific  $^{13}\text{C}$ -labeled carbon will be observable by NMR. The concentration of the  $^{15}\text{N}$ -, or  $^{13}\text{C}$ -labeled metabolite needs to be high enough ( $\geq 1$  to  $5\ \mu\text{M}$ ) to be detected by NMR. In our experience with staphylococcal and mycobacterial cultures, the volumes range from 25 mL to 100 mL and the culture media should be supplemented with approximately 2.5 to 4 g/L of  $^{13}\text{C}_6$ -glucose or ~10-15 mg/L of a targeted metabolite like  $^{13}\text{C}$ -D-alanine in order to acquire a 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC spectrum with acceptable signal-to-noise.

#### 3.4 Protocols for determining an optimal drug dosage or administering other stress treatments

To ensure consistency, the experimental variable such as a drug treatment, environmental stimuli, or gene knockout needs to be uniformly applied to the “treatment” class. An additional consideration for treatment of cultures, is that the impact on the metabolome should be strong enough to detect [49]. In other words, a particular drug dosage needs to be large enough to affect the cellular metabolome relative to untreated cells, but should not induce cell death. In our experience, a drug concentration that inhibits bacterial growth by 50% relative to untreated cultures is a desirable target [49, 50]. The availability from the literature of a minimal inhibitory concentration for the strain (MIC), or otherwise for the population isolates ( $\text{MIC}_{50}$ ), provides a good starting point for optimizing a drug dosage, but the actual dosage must be determined empirically for each set of cultivation conditions. In our experience, literature MIC or  $\text{MIC}_{50}$  values tend to be too low for cultivation conditions used for metabolomics. We typically test drug concentration ranges at between 1 to 24 times the reported MIC or  $\text{MIC}_{50}$  values in order to identify an optimal drug dosage. Importantly, this also implies that drugs with a range of biological activity will require different drug concentrations in a metabolomics study; hence, the use of the 50% inhibition of growth is used as a metric as opposed to drug concentration. Typically, in our experiments the drug treatments are normally administered during the exponential phase and the bacteria are allowed to

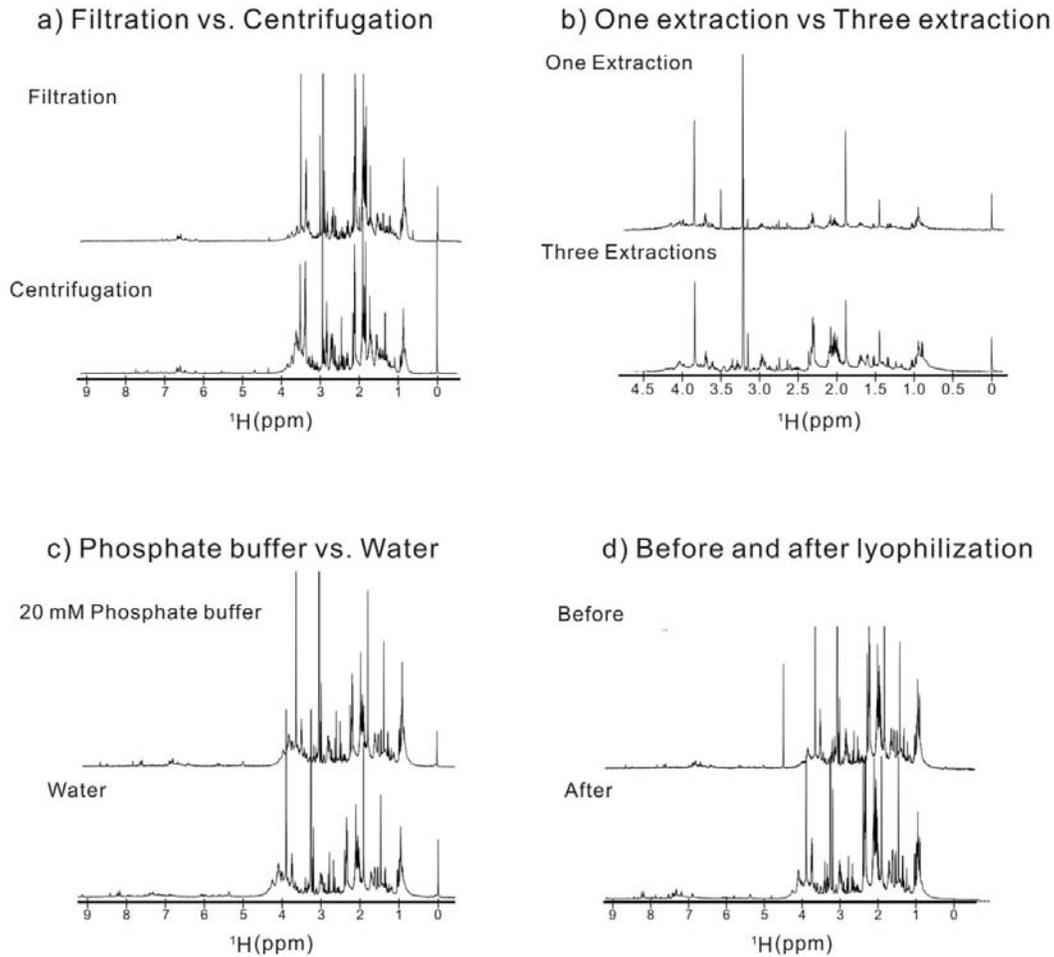
grow for at least one generation before harvesting. In our experience, this provides a sufficient amount of time for the drug to affect cell physiology and induce a perturbation in the metabolome. Administering a drug at an earlier time point can be problematic because of the inability to harvest enough bacteria.

#### 3.5 Quenching, washing and harvesting the bacterial cells

Speed is critical to harvesting bacteria without inducing a change to the metabolome. Changes occur quickly because of different metabolite turnover-rates, varying stabilities, and the induction of stress responses, among other factors [61-63]. As bacteria are being harvested, the environment is changing dramatically: (i) the bacteria are either adhered to the surface of filter paper or at the bottom of a centrifuge tube under anaerobic conditions, (ii) the temperature is changed from  $37^\circ\text{C}$  to  $\sim 0^\circ\text{C}$ , and (iii) the growth media is replaced with a phosphate buffer. To prevent perturbations to the metabolome caused by handling of the cell samples, the bacteria need to be rapidly quenched in order to stop all cellular processes from responding to these changes. Quenching efficiency has been widely discussed in the literature [42, 64-66]. Importantly, the quenching technique employed also defines the washing protocol and the order that quenching, washing and cell separation takes place. Our quenching techniques consist of filtered cells being quickly submerged into liquid nitrogen or the cells and media being directly added to  $-60^\circ\text{C}$  cold ethanol or methanol solution while being vortexed. The media and ethanol/methanol volumes are at an equal 1:1 ratio. After centrifugation, the supernatant is decanted and disposed of, and the cell pellet is ready for washing. Unfortunately, there is a possibility of cell leakage and loss of metabolites when the cells are directly added to the cold ethanol or methanol solution.

Before intracellular metabolites can be analyzed, the bacteria need to be rapidly separated from the culture media. Filtration and centrifugation are both routinely used in our laboratory to separate bacterial cells from the media. Filtration has a definitive advantage because it is significantly faster than centrifugation, but challenges in removing and collecting intact cells from filter paper may lead to sample variability. Conversely, the variability between metabolome replicates is expected to be reduced with centrifugation because of the ease in handling the cells. Nevertheless, our experience with washing bacterial cells using either filtration or centrifugation has resulted in essentially identical metabolomics fingerprints (Figure 2a); thus, any undesirable variation within a group likely occurs during sample preparation.

The use of centrifugation or filtration also determines the quenching protocol [36]. Harvesting bacteria using centrifugation requires quenching the bacteria using the direct addition to  $-60^\circ\text{C}$  cold ethanol or methanol. The bacteria, culture media, and quenching solution are in a properly sized conical centrifuge tube that is centrifuged for



**Figure 2.** Illustrations of the impact of (a) filtration and centrifugation, (b) number of extraction steps, (c) type of wash buffer, and (d) lyophilization on the composition of the metabolome.

8 minutes at 4,284 g (bucket rotor) and  $\leq 4^{\circ}\text{C}$ . Following centrifugation, the culture media and quenching solution are decanted and the bacteria are suspended in 30 mL of an ice cold wash. We routinely wash bacteria with either a phosphate buffer (20 mM, pH 7.2), or phosphate buffered saline (PBS; 6 mM phosphate buffer, pH 7.4, 137 mM NaCl and 2.7 mM KCl) to remove residual media and avoid contamination of the metabolome. The buffered wash eliminates any impact on the cells from a pH or osmotic change that may lead to cell leakage and loss of metabolites. The bacteria are centrifuged again, the wash is decanted off and the process is repeated. After two washes, the cell pellet is suspended in 1 mL of the ice cold wash and transferred to a 2 mL vial for cell lysing. Additional washings provide an insignificant benefit in removing media contaminants, but results in an undesirable increase in time. The cells are kept on ice throughout this entire process.

Harvesting bacteria by vacuum filtration collects the bacteria on sterile filter paper (0.45  $\mu\text{m}$  pore size; Millipore), while simultaneously removing the media. The number of

bacteria that can be harvested onto a filter must be empirically determined to prevent a filter blockage. Under proper conditions, removing the media should take less than a minute, and should never exceed two minutes. If this cannot be achieved, then the bacteria need to be harvested using centrifugation. After filtration, the filter paper containing the cells is then quickly placed into a 50 mL conical centrifuge tube and submerged into liquid nitrogen to freeze and quench the cells. The conical vial is then warmed by placing it into a bucket of ice for ~1 to 2 minutes. This prevents freezing of the 1 mL of wash that is added to the conical vial. The cells are gently removed from the filter paper with the wash and then transferred to a 1.5 mL microcentrifuge tube. The cells are centrifuged and washed once (1 mL).

### 3.6 Cell lysing and metabolite extraction

The cells need to be lysed in order to extract the cellular metabolome. Cells can be lysed by chemical or physical

means, but the use of chemicals runs the added risk of contaminating the metabolome. Thus, the FAST-Prep bead beating method of lysing cells is our preferred approach. Each sample is placed into a 2 mL micro-centrifuge tube with small glass beads (Lysing Matrix B; MP Biomedical) and 1 mL of extraction buffer. The cells are crushed by bead beating for 40 to 60 seconds in the FAST-Prep instrument at a speed of 6.0 m/s. This process is repeated after keeping the crushed cells on ice for 5 minutes. The sample is then centrifuged for 2 minutes at 17,000 g to pellet the cell debris. The supernatant with the extracted metabolites is collected. The cell debris is washed 1 to 3 times with 1 mL of the extraction buffer to maximize the metabolome yield (Figure 2b). Also, double distilled water or a phosphate buffer are routinely used as the extraction buffer, since both approaches provide similar results (Figure 2c). All extracts per sample are combined for lyophilization, where the sample is then dissolved in 700  $\mu$ L of a phosphate buffer in D<sub>2</sub>O at pH 7.2 (uncorrected). Lyophilization may negatively impact some volatile metabolites, but, in general, no effect is observed (Figure 2d). A major concern during the extraction step is maximizing the overall yield while minimizing any perturbation to the metabolome. In our experience, the cell lysing and metabolite extraction process will require approximately 45 minutes for 30 cultures. The metabolomics samples can be stored in a -80°C freezer or directly lyophilized over-night.

#### 4. NMR Spectroscopy

##### 4.1 One-dimensional <sup>1</sup>H NMR methodology

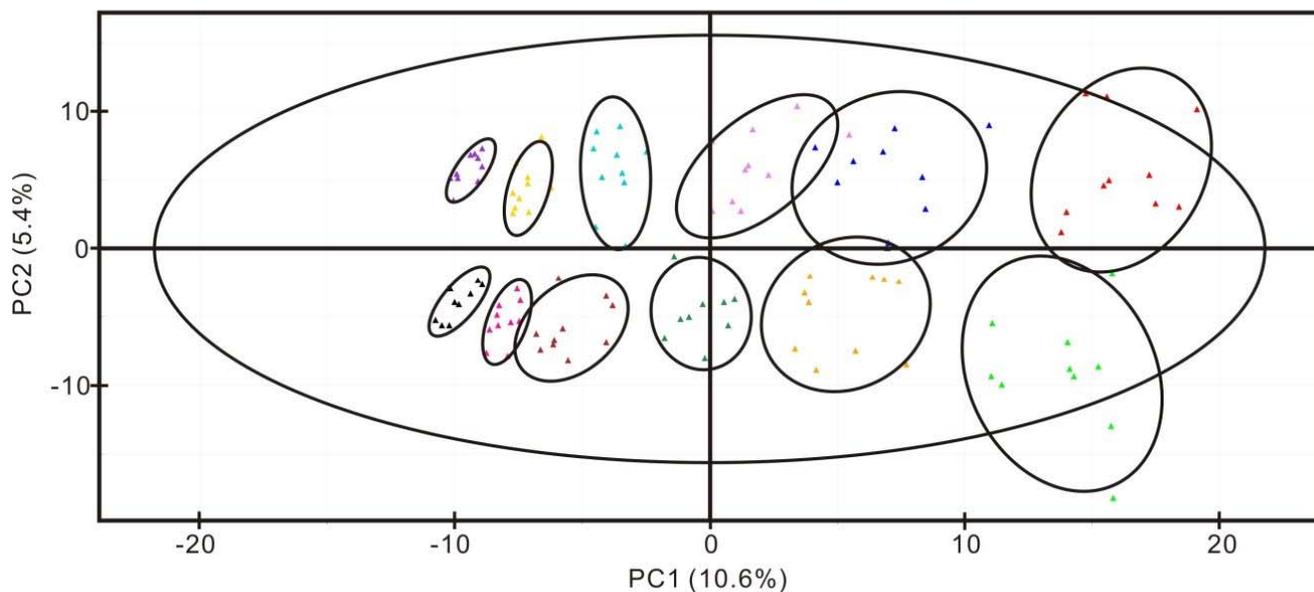
One-dimensional (1D) <sup>1</sup>H (proton) NMR is an unbiased, nonselective, and nondestructive approach that requires no modification of the samples, where the data can be collected in a high-throughput manner. A 1D <sup>1</sup>H NMR spectrum contains numerous proton signals generated from a complex metabolomics mixture, where the chemical shift of each signal describes the structural characteristic of a specific metabolite. Moreover, the peak intensities or volumes are directly proportional to the concentration of each metabolite. Quantification of metabolites can be achieved by using an internal standard with a known concentration, where we routinely use 50  $\mu$ M 3-(trimethylsilyl) propionic acid-2,2,3,3-d<sub>4</sub> (TMSP-d<sub>4</sub>, Sigma). Thus, 1D <sup>1</sup>H NMR experiments combined with multivariate statistics are commonly used for the global analysis of the metabolome.

Collecting 1D <sup>1</sup>H NMR data for metabolomics is fast and simple, and provides highly reproducible and accurate results. Importantly, the NMR experimental parameters need to be identical for each metabolomics sample in order to collect reliable metabolomics data. Any per sample variation will erroneously bias the resulting clustering patterns in the PCA and OPLS-DA scores plot. To avoid this and maintain sample consistency, we use a BACS-120 sample changer, Bruker ICON-NMR, an automatic tuning and matching

(ATM) unit, and autoshim to automate the NMR data collection. Nevertheless, instrument drift may still occur during the high-throughput metabolomics screen, so it is also important to randomize the samples during NMR data collection. If an NMR spectrum is collected first for all the control samples followed subsequently by each treatment class, there is a significant potential of inducing a biologically irrelevant bias into the analysis. The clustering pattern in the PCA and OPLS-DA scores plot may be dominated by the order of data collection instead of the expected biological differences.

In our laboratory, a typical 1D <sup>1</sup>H NMR spectrum is collected using 128 scans and 32k data points on a Bruker 500 MHz Avance DRX NMR spectrometer equipped with a triple-resonance, Z-axis gradient cryoprobe. The acquisition time is approximately 10 minutes per sample. The goal is to obtain optimal signal to noise while minimizing the total experimental time. We previously demonstrated that spectral noise is detrimental to the resulting PCA and OPLS-DA scores plot [55]. Random noise fluctuations results in large and irrelevant variations in the scores clustering. To avoid this problem, spectral noise needs to be removed prior to PCA and OPLS-DA. Correspondingly, the quality of the within class clustering in PCA and OPLS-DA scores plot is directly dependent on the spectral signal-to-noise (Figure 3). The within class variance decreases dramatically as the number of scans (signal-to-noise) is increased from right to left in the scores plot. Importantly, the spectral noise was still removed prior to PCA. Thus, the accuracy of identifying similarities or differences between multiple classes is dramatically improved by reducing within class variance, which is achieved by improving spectral sensitivity. Also, correctly identifying class differences improves with the number of replicates (Figure 4). The statistical significance of cluster separation as measured by *p*-value [67] is shown to decrease as both a function of group variance and the number of replicates. As a result, we prefer to use ten replicates per class and strive to achieve an average signal-to-noise ratio of > 100 to 200. This is achieved by simply increasing the number of scans or the number of cells, whichever is more practical. While signal-to-noise has a dramatic impact on scores clustering, PCA and OPLS-DA is indifferent to changes in spectral resolution unless the number of data points is  $\leq 2K$ .

A D<sub>2</sub>O phosphate buffer is the typical solvent of choice for aqueous metabolomics samples in order to efficiently remove residual water signals and avoid interference from buffer signals. Water and buffer signals are problematic since they can distort the NMR spectrum and may overlap and obscure important metabolite signals. Most NMR processing software can automatically remove the residual water peak, but extra data processing is required to correct for baseline distortions induced by the solvent. Unfortunately, simply applying a baseline correction changes the PCA and OPLS-DA clustering patterns [68]. Furthermore, different baseline correction protocols will induce variable changes into the scores plot. Also, removing the residual water peak may

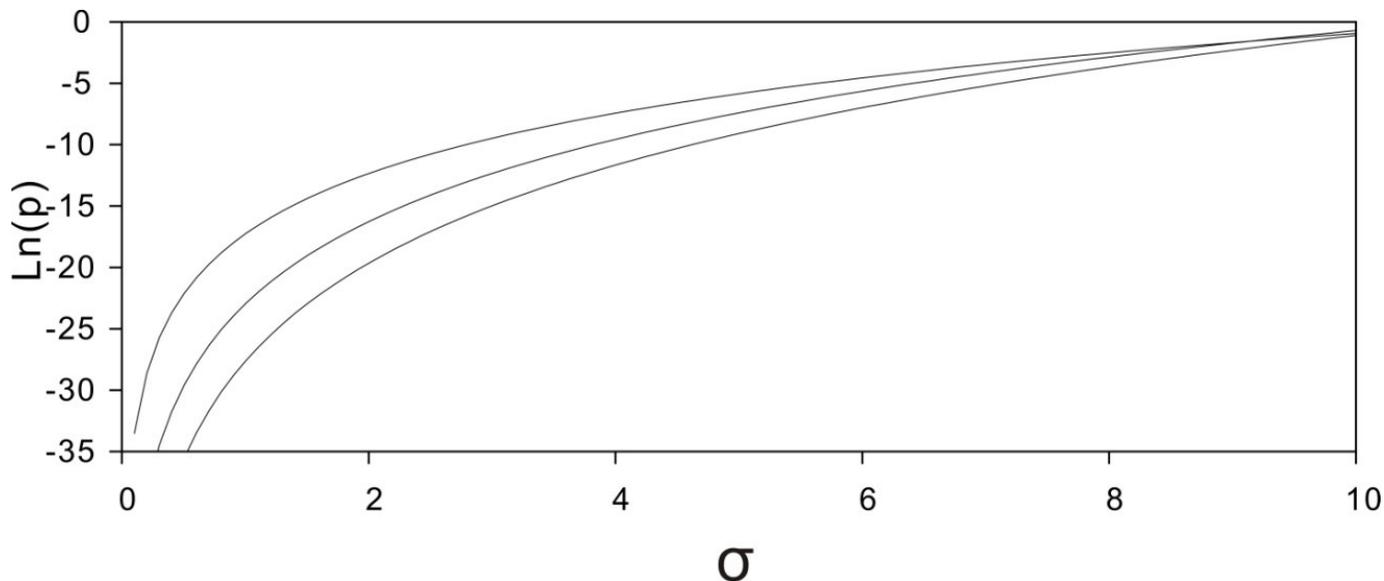


**Figure 3.** Illustration of the impact of the NMR signal-to-noise on within class variation in a PCA scores plot. From right to left, the 1D <sup>1</sup>H NMR spectra were collected with an increasing number of scans (1, 2, 4, 8, 16, and 32) resulting in a proportional increase in signal-to-noise. All other experimental parameters were kept constant.

result in a potential loss of information by also removing metabolite peaks near the water signal. Instead, a water suppression technique that experimentally removes the water peak without inducing baseline distortions is the preferred alternative.

There are a variety of NMR pulse sequences for water

suppression that are available to the metabolomics community, such as WATERGATE, water pre-saturation, WET, and PURGE [69-73]. Our preferred choice for a water suppression pulse sequence is Solvent-Optimized Gradient-Gradient Spectroscopy (SOGGY). SOGGY does an outstanding job in eliminating the water signal without



**Figure 4.** Illustration of the impact of within group variation and the number of replicates on the *p* values calculated between clusters in a simulated PCA scores plot. From top to bottom, *p* values from the simulated PCA scores plot were calculated with an increasing number of replicates (6, 8, 10) resulting in a proportional decrease in *p* values. Similarly, increasing the group variation by increasing the standard deviation ( $\sigma$ ) per cluster resulted in a significant increase in *p* values.

inducing any base line distortions (Figure 2) [73]. SOGGY is a variant of excitation sculpting that employs a pulsed field gradient with a simple phase-alternating composite pulse. SOGGY offers the flexibility to optimize the 180 degree hard pulse to achieve optimal excitation of the water signal, and adjusting the 180 degree soft pulse to optimize the range of the water frequency to be suppressed [73, 74]. As a result, SOGGY efficiently suppresses the water signal while removing any phase cycle artifacts. A flat baseline is obtained while also maintaining metabolite signals near the water signal [73]. SOGGY completely eliminates the need to apply any baseline correction.

#### 4.2 Two-dimensional $^1\text{H}$ - $^{13}\text{C}$ HSQC NMR methodology

The severe overlap of signals in a 1D  $^1\text{H}$  NMR spectrum is a challenge for metabolite identification. The difficulty arises because hundreds to thousands of peaks occupy a small chemical shift range ( $\sim 10$  ppm), where multiple metabolites share similar chemical shifts. Thus, we typically do not use 1D  $^1\text{H}$  NMR spectra to assign metabolites. Instead, we routinely use 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiments for metabolite assignments. The 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment is a more reliable approach for metabolite identification because of the significantly higher resolution and the correlation between  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts for each C-H pair in a molecule [75, 76]. Also, the 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment simplifies the analysis of the metabolome because only compounds containing a  $^{13}\text{C}$ -carbon derived from the  $^{13}\text{C}$ -labeled metabolite added to the media will be detected.

In our laboratory, we use a standard Bruker 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC pulse sequence on our 500 MHz spectrometer, where an acceptable signal-to-noise is achievable using 64 scans. Similarly, a reasonable digital resolution is achieved by collecting 2K and 128 data points in the direct and indirect direction, respectively, with a corresponding spectral width of 10 ppm and 140 ppm along the  $^1\text{H}$  and  $^{13}\text{C}$  axis, respectively. Since some aromatic C-H pairs have a  $^{13}\text{C}$  chemical shift greater than 140 ppm, the spectrum will contain folded peaks, but the folded peaks will not interfere with or overlap with other metabolite peaks due to their unique position along the  $^1\text{H}$  axis ( $\sim 7.0$  ppm). This folding technique allows for an increase in the digital resolution without incurring an increase in acquisition time. In general, the 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment requires approximately 4 hours per sample on our system.

A conventional 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC spectrum is useful for detecting metabolite changes by overlaying multiple spectra to identify missing peaks or peaks with significant intensity changes. Unlike 1D  $^1\text{H}$  NMR spectra, obtaining metabolite concentrations is more difficult because peak intensities are dependent on J-couplings, dynamics and relaxation, in addition to metabolite concentrations [77, 78]. To quantify absolute metabolite concentrations, we use the Time-Zero HSQC (HSQC<sub>0</sub>) experiment [77]. This approach requires collecting a series of three HSQCs spectra (HSQC<sub>1</sub>, HSQC<sub>2</sub>,

HSQC<sub>3</sub>) with an increased number of pulse sequence repetitions. A natural log plot of peak areas or intensities versus the increment number (1, 2, 3) allows for an extrapolation back to increment 0 or zero-time. The experimental parameters used in the HSQC<sub>0</sub> experiment is similar to the conventional method, but with some minor variations. The number of scans is increased to 128 due to the decrease in signal-to-noise in HSQC<sub>2</sub> and HSQC<sub>3</sub>. To partially compensate for the increase in experimental time, the number of data points in the indirect dimension is reduced to 64. In general, the HSQC<sub>0</sub> set of experiment requires approximately 6 hours per sample on our system.

## 5. Data analysis

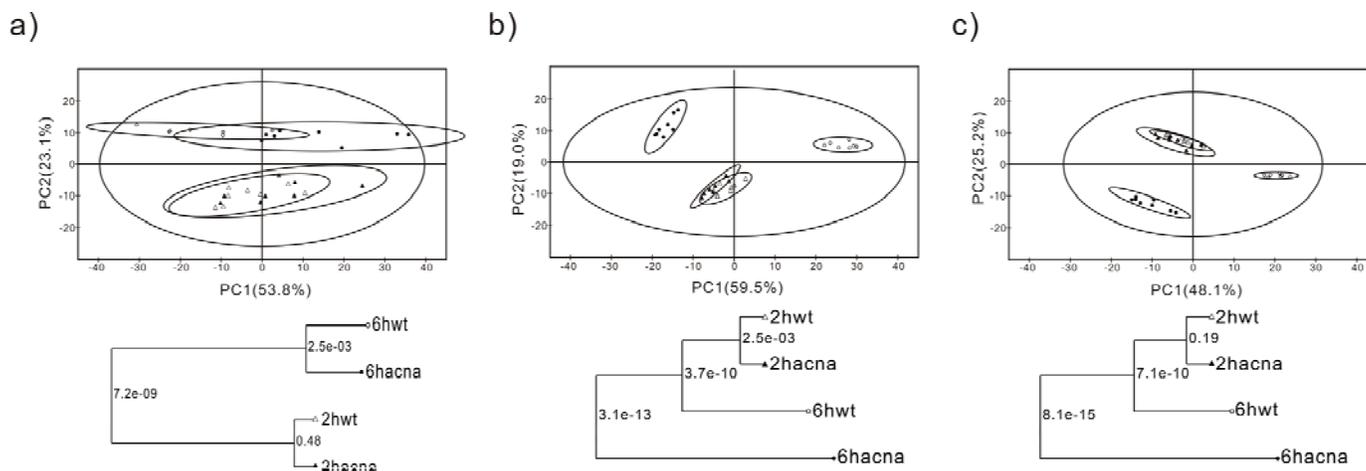
### 5.1 Preprocessing of 1D $^1\text{H}$ NMR data

The 1D  $^1\text{H}$  NMR spectra are minimally processed (Fourier transformed and phase corrected) using ACD/1D NMR Manager (Advanced Chemistry Development). Each NMR metabolomics sample contains 50  $\mu\text{M}$  of TMSP- $\text{d}_4$  as an internal standard, where each NMR spectrum is referenced to the TMSP- $\text{d}_4$  peak and uniformly aligned to 0.00 ppm. Also, all peak heights are normalized to the intensity of the TMSP- $\text{d}_4$  peak. Intelligent bucketing within the ACD/1D NMR Manager is then used to integrate each spectral region with a bin size of 0.025 ppm. The spectra are normalized; noise regions and residual solvent and buffer resonances are removed, and then the remaining bins are scaled prior to PCA and OPLS-DA using the commercial SIMCA12.0+ (UMETRICS) statistical package (<http://www.umetrics.com/>).

The need for data normalization and scaling prior to multivariate statistical analysis has been extensively discussed in the literature [79, 80]. Normalization adjusts for experimental variations between replicates, different number of cells, varying signal-to-noise, etc., and minimizes these contributions to the clustering patterns in PCA and OPLS-DA scores plot. We have encountered significant success in using a Z-score or center averaging the spectrum:

$$Z = \frac{X_i - \bar{X}}{\sigma} \quad (1)$$

where  $\bar{X}$  is the average signal intensity in a given spectrum,  $\sigma$  is the standard deviation in the signal intensity, and  $X_i$  is the signal intensity within bin  $i$  (Figure 5a). After normalization, all the noise bins are uniformly removed. This was initially accomplished by manually identifying a “reference” noise region above 10 ppm and below 0 ppm; and calculating an average noise value. If a bin across all replicates had an integral value of less than twice the average noise, it was also identified as noise and removed (Figure 5b). The protocol for identifying noise regions has been recently improved upon and results in smaller within class variations (Figure 5c). This also results in an improved separation between truly distinct classes and removed erroneous separations. For



**Figure 5.** Illustration of the impact of NMR preprocessing on within and between class variations in a PCA scores plot. (a) The 1D <sup>1</sup>H NMR spectra was not properly preprocessed. The spectra were not normalized and the noise was not removed. The spectra were only Fourier transformed, phased corrected, and the residual H<sub>2</sub>O resonance was removed. (b) The 1D <sup>1</sup>H NMR spectra were processed as in (a) with the addition of normalization using center averaging, but without noise removal. (c) The 1D <sup>1</sup>H NMR spectra were processed as in (b) with the addition of noise removal. Each spectrum was binned using intelligent bucketing with a bin size of 0.025 ppm. The ellipses correspond to the 95% confidence limits from a normal distribution for each cluster. The PCA scores plots compare the metabolomes of *S. aureus* wild-type (wt) strain SA564 with an aconitase mutant (acna) strain SA564-*acnA::tetM* at either two hours (2h) or six hours (6h) of cell growth. Below each PCA scores plot is a corresponding dendrogram generated from the scores using Mahalanobis distances, with *p* values for the null hypothesis reported at each branch.

example, the statistical significance between clusters 6hwt and 6hacna improved from a *p*-value of  $3.1 \times 10^{-13}$  to  $8.1 \times 10^{-15}$ , while the small, but biologically irrelevant, separation between clusters 2hwt and 2hacna (*p*-value  $2.5 \times 10^{-3}$ ) was removed (Figure 5). Instead of using an average minimal signal intensity to define noise, we now define noise based on a relative standard deviation. This is based on the expectation that real NMR peaks from metabolites will have a higher intrinsic variability compared to the noise because of biological variations that naturally occur even between within class replicates. Conversely, the variability of the noise should be effectively constant for a given spectrometer operating within normal parameters. Simply, the standard deviation and average is calculated for each bin, where the standard deviation is normalized by the average peak intensity. This avoids eliminating weak peaks with a relatively small standard deviation. The same is done for the reference noise region, which is then used to define noise:

$$\text{Noise: } \sigma_i \leq \sigma_n \quad (2)$$

$$\text{Cutoff: } \sigma_0 = \text{avg}(\sigma_n) + 2 * \text{sd}(\sigma_n) \quad (3)$$

where  $\sigma_i, \sigma_n$  are the relative standard deviations (absolute standard deviation divided by the mean) for the *i*th bin in the spectral region and *n*th bin in the reference noise region, respectively, and  $\text{avg}(\sigma_n)$  and  $\text{sd}(\sigma_n)$  are the mean and standard deviation of  $\sigma_n$  respectively. In effect, any peak that falls within the normal distribution of the reference noise region is defined as a noise bin. This approach is better at defining noise peaks in crowded and overlapping regions of the NMR spectra.

In addition to normalization, each bin or column in the data matrix also needs to be scaled to account for the large dynamic range in peak intensities. PCA and OPLS-DA emphasizes the absolute variation in bins between classes. Correspondingly, the relative variation of an intense peak may be insignificant compared to a weak peak, but the absolute changes in its intensity may completely mask biologically relevant changes in a small peak. Scaling increases the weight of the low intensity peaks so strong peaks do not dominate in PCA and OPLS-DA [79, 80]. In our experience, unit variance scaling, also known as autoscaling or a Z-score (see eqn. 1), has been shown to be effective in generating reliable clusters with the correct separation based on biologically relevant class distinctions. Also, within class variance is reduced using autoscaling, which is our default scaling method.

### 5.2 Multivariate statistical analysis of 1D <sup>1</sup>H NMR data

We routinely apply PCA, a non-supervised technique, to determine if the 1D <sup>1</sup>H NMR data can easily distinguish between the various test classes. PCA provides an unbiased view of group clustering in the resulting 2D scores plot. We only use a three-dimensional (3D) scores plot if class separation in a 2D scores plot is insufficient and the PC<sub>3</sub> contribution is significant (> 5 to 10%). OPLS-DA is only used if class separation is observed in the PCA scores plot. OPLS-DA is a supervised technique and assesses a relationship between the NMR data class designations. We limit OPLS-DA to only two class designations that

differentiate between the single control group (0) and the entire treatment group (1). As a supervised technique, OPLS-DA maximizes a separation between these two classified groups, while minimizing within class variations [39]. Thus, OPLS-DA identifies the important spectral features (metabolites) that primarily contribute to class separation. We routinely use an OPLS-DA S-plot or loading plot (Figure 1) to readily identify the key metabolites that contribute to class separation. Since OPLS-DA is a supervised technique and can generate a class separation even for random data [81], it is essential to verify the model [46]. But this is also an advantage of OPLS-DA over PCA; the statistical significance of the model is quantified. We cross-validate OPLS-DA models using a modified leave-one-out method [82, 83] and CV-ANOVA [84]. The modified leave-one-out method provides a quality assessment score ( $Q^2$ ) and  $R^2$  values, where CV-ANOVA provides a standard  $p$ -value. The theoretical maximum for  $Q^2$  is 1, where a value of  $\geq 0.4$  is an empirically acceptable value for biological samples [85], but  $Q^2$  does not have a critical value for inferring significance. It is still possible for an invalid model to produce a large  $Q^2$  value. Similarly, the  $R^2$  values only provide a measure of the fit of the data to the model. But large differences between  $Q^2$  and  $R^2$  ( $R^2 \gg Q^2$ ) does suggest an over-fit model. Conversely, a  $p$ -value  $\ll 0.05$  from CV-ANOVA provides clear validation of the OPLS-DA model.

In addition to validating the OPLS-DA model, it is also extremely important to verify the statistical significance of the clustering patterns in the PCA and OPLS-DA scores plot. Is the between group difference larger than the within group variations? One key factor is the number of replicate samples. We have previously shown that increasing the number of replicates improves the statistical significance of cluster separation [86]. This finding is also supported by the increase in  $p$ -values seen with an increase in within class variations (Figure 4). Again, increasing the number of replicates improves the statistical significance of the class separation (lower  $p$ -value) even when within class variation increases. Correspondingly, we routinely use 10 replicates per group in our metabolomics study to improve the likelihood of observing statistical significant class separations.

It is also important to visually define each group or class within the PCA and OPLS-DA scores plot and to classify the statistical significance of the class separation. We developed a free PCA and OPLS-DA utilities software package (<http://bionmr.unl.edu/pca-utils.php>) [67] that draws ellipses or ellipsoids around each group cluster in a scores plot, where the ellipse corresponds to the 95% confidence limits from a normal distribution for each cluster. Visual separation of the ellipses infers a class separation. The same software package is also used to generate a metabolomics tree diagram based on the group clusters in the scores plot [67, 86]. Simply, a centroid from each cluster is used to calculate a Mahalanobis distance between clusters, where dendrograms are then generated from the resulting distance matrix. The

significance of each node (cluster separation) is determined by using standard bootstrapping techniques and returning a bootstrap number [87], where a value above 50 infers a significant separation; or from Hotelling's  $T^2$  and  $F$ -distributions that returns a  $p$ -value, where a number  $\ll 0.05$  infers a statistically significant separation.

Observing a statistically significant difference in the global metabolome between two or more bacterial samples is typically the first objective of a metabolomics investigation. While this difference may infer some biological significance, the ultimate goal is to identify the underlying metabolites and associated pathways that are the primary contributors to the observed class separation in the PCA and OPLS-DA scores plot. One approach is to generate an S-plot (Figure 1) from the resulting OPLS-DA analysis. The S-plot identifies the key bins or  $^1\text{H}$  chemical shifts that are correlated or anti-correlated with the separation between the two classes in an OPLS-DA scores plot. The  $^1\text{H}$  chemical shifts can then be compared against a number of online NMR metabolomics databases [88-92] to assign the metabolites. Unfortunately, an unambiguous assignment is rarely possible because of the low chemical shift dispersion and the large number of potential metabolites. Instead, 2D NMR experiments combined with the biological knowledge of the system under investigation are required to improve the accuracy of metabolite identification.

### 5.3 Metabolite Identification

#### 5.3.1 Automated peak picking of 2D NMR data

2D  $^1\text{H},^{13}\text{C}$  HSQC and  $^1\text{H},^1\text{H}$  TOCSY spectra are commonly used for metabolite identification because of the increase in chemical shift resolution achieved by spreading the information out into two-dimensions. Also, the correlation between  $^1\text{H}$  chemical shifts for each J-coupled H pair; and the correlation between  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts for each C-H pair significantly reduces the assignment ambiguity. This occurs because both chemical shifts have to match a single metabolite in the database to make an assignment. Despite the advantages, peak picking and organizing a table of intensities from a 2D NMR experiment is a time consuming process, especially when multiple spectra are involved. Numerous software packages are available to automate the peak picking of 2D NMR spectra, however; it is extremely difficult, if not impossible, to align and match multiple sets of spectra with different peak patterns due to unique metabolomes.

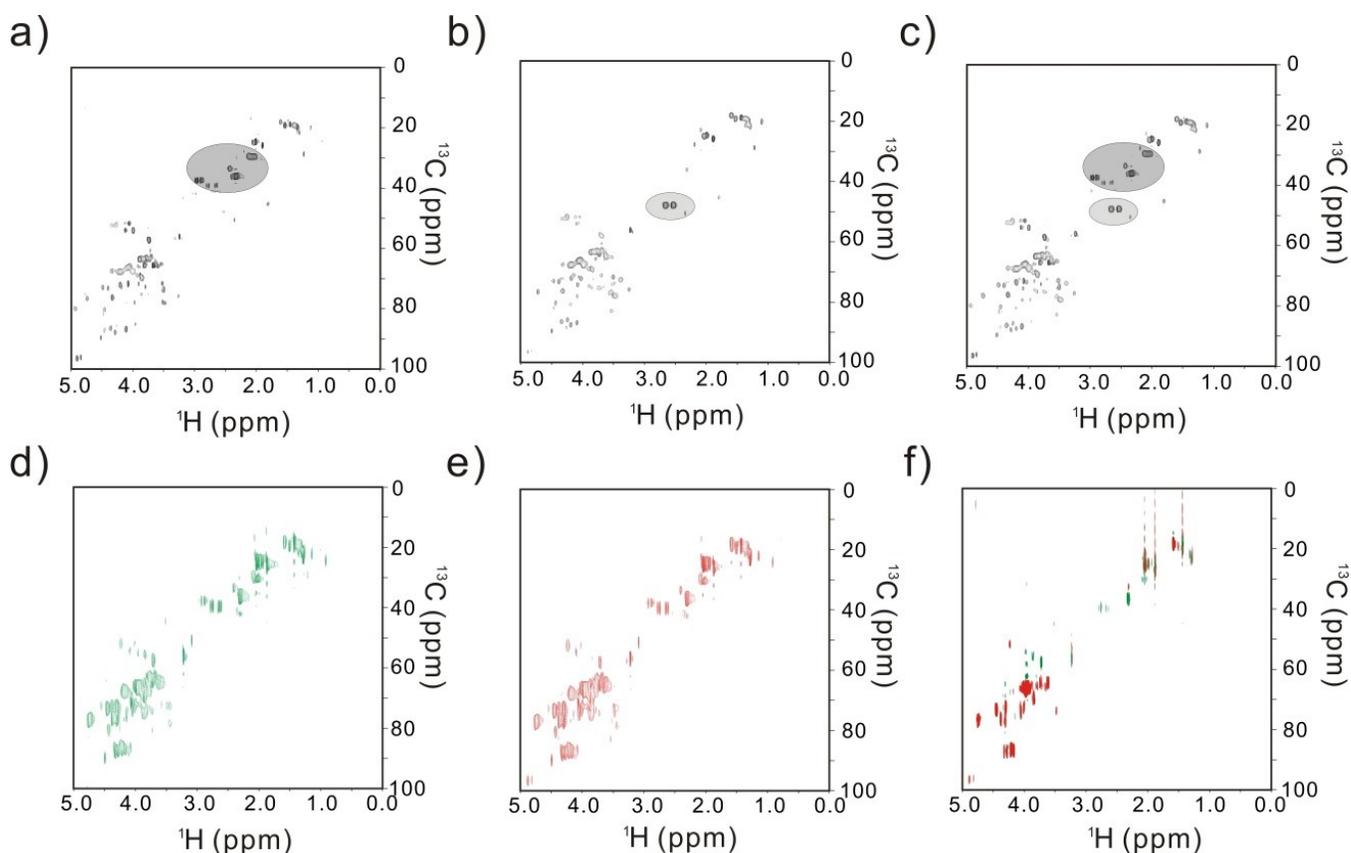
For example, three different sets of cell cultures (different cell types, treatments or environmental conditions, etc.) will each exhibit a distinct set of peaks in the NMR spectrum due to the presence of unique metabolites. These unique peaks will be mixed with other peaks common to all three groups, but the relative peak intensities are likely to vary due to different metabolite concentrations. Thus, if the control group is designated as the reference spectrum for automated

peak picking, a peak list will be generated that only contains peaks observed in the control spectrum that are above the designated noise threshold. Correspondingly, peaks unique to the other two groups will be missed when this peak list is used to peak pick their spectrum. In addition, weak peaks may also be missed due to different noise levels between the spectra and a corresponding difference in the threshold setting for peak picking. Instead, a composite reference spectrum for automated peak picking needs to be generated that captures *all* the peaks present in the three separate groups. We accomplish this task by using the addNMR function in the free NMRpipe software package (<http://spin.niddk.nih.gov/NMRPipe/>) [93]. As the name implies, addNMR mathematically sums all spectra together from the three groups to make a single spectrum. This resulting “master spectrum” contains all the peaks observed throughout the set of 2D experiments and is used to generate a peak list for automated peak picking of each individual spectrum. Critically, the 2D NMR spectra need to be collected and processed using identical experimental

parameters (spectral width, data points, zero-filling, etc.) and needs to be aligned to an internal reference (TMSP-d<sub>4</sub>). In our experience, all the peaks from the complete set of NMR spectra are routinely matched to the reference list by using a chemical shift error-tolerance of 0.04 ppm and 0.25 ppm in the <sup>1</sup>H and <sup>13</sup>C dimensions, respectively. This approach has greatly simplified and increased the efficiency of a previously laborious procedure. The addNMR command can also be used to generate a difference spectrum that clearly highlights the major spectral changes between two classes (Figure 6).

### 5.3.2 Assignment of an NMR peak to a metabolite

Metabolite identification is an extremely important component of the metabolomics process because it enables the determination of the key metabolites perturbed by the treatment or the metabolites primarily contributing to class distinction. This includes the discovery of important biomarkers associated with drug efficacy or drug resistance. Also, metabolite identification is important to the drug



**Figure 6.** (a-c) Illustration of the procedure to generate a “master spectrum” and facilitate automated peak picking by creating a complete peak list. (a-b) Representative 2D <sup>1</sup>H-<sup>13</sup>C HSQC spectra obtained from two distinct bacterial cultures, where some major spectral differences are highlighted. (c) The two 2D <sup>1</sup>H-<sup>13</sup>C HSQC spectra from (a-b) were added to yield a master spectrum that contains all the observed NMR peaks. (d-f) Illustration of the procedure to generate a “difference spectrum” to facilitate metabolite identification by creating a signed (+, -, null) peak list. (d-e) Representative 2D <sup>1</sup>H-<sup>13</sup>C HSQC spectra obtained from two distinct bacterial cultures. (f) The two 2D <sup>1</sup>H-<sup>13</sup>C HSQC spectra from (d-e) were subtracted to yield a difference spectrum that identifies the NMR peaks, and correspondingly metabolites, that differ between the two bacterial cell cultures. Positive peaks, increased metabolite concentration, are colored green and negative peaks, decreased metabolite concentration, are colored red.

discovery process by either identifying metabolic pathways affected by a drug to evaluate efficacy or potential toxicity; or by identifying potentially new therapeutic targets. Nevertheless, accurate metabolite identification is very difficult and labor-intensive. The success of metabolite spectral assignment relies largely on the completeness of metabolomics databases. We routinely use a combination of the following freely-accessible databases: Human Metabolome Database (<http://www.hmdb.ca/>) [88], Madison Metabolomics Consortium Database (<http://mmcd.nmrfa.wisc.edu/>) [89], Platform for RIKEN Metabolomics (PRIME) (<http://prime.psc.riken.jp/>) [90], BioMagResBank (<http://www.bmrw.wisc.edu/>) [91], and Metabominer (<http://wishart.biology.ualberta.ca/metabominer/>) [92], which provide both redundant and complementary NMR spectral data. Importantly, the reference NMR spectra in the various databases were obtained under different buffer conditions and use different internal standards. This results in a range of potential chemical shifts for a given metabolite. Thus, the database with sample conditions that closely match our experimental conditions are used for chemical shift matching. The overall goal is to identify a complete set of metabolites as quickly and accurately as possible without any bias, by matching the experimental chemical shifts from the 2D NMR spectra with the values in the database.

For a 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment, it is important to realize that metabolites may be heterogeneously labeled by the carbon-13 source present in the growth media. Correspondingly, all the peaks for a specific metabolite may not be detectable in the 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment. Also, a reference spectrum for the metabolite may not be present in any of the available databases. The assignment of a particular peak might still be ambiguous because multiple metabolites may contain the same chemical shift or contain an identical substructure (*e.g.*, ATP, ADP, AMP or NAD, NADPH). Therefore, a few automated filters are applied to overcome some of these ambiguities during the peak assignment process.

The first filter is to verify that the bacteria can actually produce the proposed metabolite. This is routinely accomplished by searching the freely-accessible Biocyc (<http://biocyc.org/>) [94] and KEGG (<http://www.genome.jp/kegg/>) [95] database for metabolites known to exist in the bacteria under investigation. The second filter is based on a differential peak list. All the NMR peaks potentially assigned to a specific metabolite should have the same trend relative to the control. Obviously, the metabolite can only have one concentration and all the NMR peaks need to be consistent with this single concentration. Correspondingly, all the peaks have to be increased, decreased or the same relative to the same peaks in the control spectrum. This is easily and quickly visualized by subtracting the two sets of spectra and generating a signed (+, -, null) peak list. Peaks assigned to the same metabolite have to have the same sign. The third filter is based on a biological relationship with other

metabolites. Simply, the likelihood of a correct assignment increases if other metabolites in a specific metabolic pathway have also been assigned. It is more likely to observe multiple metabolites from the same pathway than various metabolites from unrelated pathways. Similarly, if there is a direct metabolic path between two or more metabolites, then their assignments are more likely to be correct. The final filter is the application of our biological knowledge of the bacterial system under investigation. The pathways or metabolites that are expected to be perturbed by the treatment would be given precedent in the assignment process. As a simple example, a comparison between wild-type and mutant bacterial strains where aconitase has been inactivated would reasonably be expected to lead to changes in metabolites associate with the TCA cycle. Likewise, a comparison between untreated and drug-treated cells would be expected to lead to changes in metabolic pathways inhibited by the drug.

### 5.3.3 Statistical analysis of the 2D $^1\text{H}$ - $^{13}\text{C}$ HSQC data.

After assigning the 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC spectra to a set of metabolites, the next goal is to determine metabolite concentration differences between the various bacterial culture conditions under investigation. Unfortunately, peak intensities in a standard 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment are dependent on multiple parameters [77, 78], so only a relative percent change in a metabolite concentration can be determined [5]. Alternatively, an absolute concentration can be determined using HSQC<sub>0</sub>, which requires a set of three HSQC experiments per sample. We routinely employ both approaches [77].

A relative difference in peak intensities is determined by using a triplicate set of a conventional 2D  $^1\text{H}$ - $^{13}\text{C}$  HSQC experiment for each bacterial culture condition. Prior to calculating a relative percent change in peak intensities, a detailed normalization process is required, which was previously described in detail [5]. First, the peak intensities within each spectrum are normalized by dividing each peak by the internal standard, the intensity of the TMS- $\text{d}_4$  peak. Each peak pertaining to a specific chemical shift across each triplicate data set is then normalized by the most intense peak in the set of three peaks. Specifically, the maximal intensity for each peak across all data sets would be set to 100 and all other intensities are scaled relative to this peak intensity. Then all the normalized intensities for a given metabolite for each triplicate set are averaged together, and a relative percent error can be calculated between different cultures. A Student's t-test or ANOVA is then used to determine if the relative change in peak intensities is statistically significant at the 95% confidence limit. Calculating a relative difference in metabolite concentrations can be beneficial to understanding broader changes to the system, especially when a cluster of metabolites in a specific pathway exhibit a similar trend in concentration changes inferring an important role for the metabolic pathway.

Nevertheless, this approach is rather cumbersome and does not allow for a direct comparison between different metabolites.

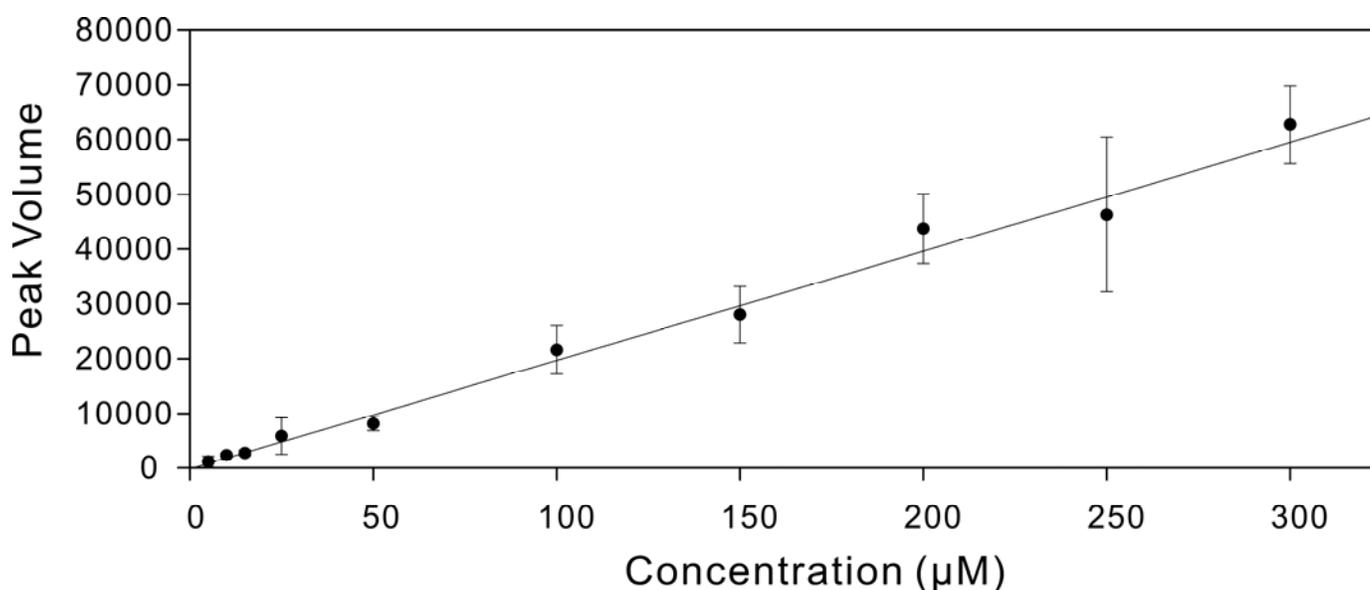
Alternatively, we routinely use the HSQC<sub>0</sub> experiment to determine absolute metabolite concentrations. The overall protocol for the extrapolation of peak intensities to time-zero and the determination of the associated concentration has been previously described in detail [77, 78]. A distinct advantage of this method is that a single calibration curve can be made using multiple compounds with known concentrations to correlate the time-zero peak intensity with a concentration. Figure 7 illustrates such a calibration curve using 5 different mixtures, each consisting of 9 different <sup>13</sup>C-labeled metabolites ranging in concentrations from 5 to 300 μM. Also, the concentration for each metabolite was randomized in each mixture. For example the concentration of <sup>13</sup>C-D-alanine in the 5 mixtures is 300, 10, 25, 5, and 100 μM, respectively. The data was fitted using a weighted linear least squares calculation. Notably, the best-fit line ( $R^2$  0.997) has a y-intercept close to zero as expected for a concentration of zero. Also, the correlation between peak volume and concentration is independent of the metabolite. Importantly, the accurate application of the calibration curve requires collecting and processing HSQC<sub>0</sub> spectra *identical* to the parameters used to obtain the original calibration curve. Critically, the receiver gain must be the same for all samples, because any change in the receiver gain influences the slope of the calibration curve. Also, the addition of 500 μM TMSP-d<sub>4</sub> as an internal standard is crucial, because both the calibration samples and experimental samples must both be normalized to the TMSP-d<sub>4</sub> peak. As an example, if the

TMSP-d<sub>4</sub> peak volumes for the calibration mixtures are 1000, 500, and 250 for each HSQC<sub>i</sub> ( $i = 1, 2, 3$ ) spectrum, respectively, then the experimental results for all *in vivo* metabolite extracts must be normalized so that the internal standard (TMSP-d<sub>4</sub>) peak volumes are also 1000, 500, and 250. The concentrations are measured in triplicate, where a Student's t-test or ANOVA is used to determine if the concentration changes are statistically significant at the 95% confidence limit.

#### 5.4 Metabolomics Network Map

Metabolites are highly interconnected through numerous metabolic pathways that form an extremely complex network [96]. Correspondingly, it is not uncommon to observe correlated changes between distantly connected metabolites. In effect, metabolomics depends on these complex interactions to understand the phenotype of a bacterial cell. Thus, a metabolomics network map provides an efficient approach to visualize and summarize the overall changes to the metabolome, to validate metabolite assignments based on clear connections to other metabolites, and the identification of key metabolic pathways.

We have routinely used Cytoscape (<http://www.cytoscape.org/>) to easily and quickly generate metabolomics network maps [97-100]. Cytoscape is a free, user-friendly software package with plug-ins related to metabolomics. Cytoscape simply requires a list of the metabolites and their associated concentration changes as input. The connections between nodes (metabolites) in the map are based on metabolic pathways from the freely-



**Figure 7.** A strong correlation between NMR peak volumes and metabolite concentrations ( $R^2$  0.997) is demonstrated by linear regression plot generated from HSQC<sub>0</sub> data. HSQC<sub>0</sub> NMR spectra were collected for five different metabolite mixtures containing nine <sup>13</sup>C-labeled compounds with concentrations ranging from 5 μM to 300 μM. The relationship between peak volume and metabolite concentration is independent of the metabolite.

accessible MetaCyc database (<http://metacyc.org/>) [101]. An example of a typical Cytoscape map summarizing the observed changes in the *S. epidermidis* metabolome caused by environmental stimuli associated with biofilm formation is shown in Figure 1. The metabolomics network map can be easily modified to highlight specific features of the metabolome. Edges can be broadened to highlight specific pathways; and the color and size of nodes can be adjusted to reflect the direction and magnitude of the concentration changes, respectively [102]. Cytoscape also provides a range of map design choices. Unfortunately, the resulting network maps (Figure 1) do not resemble standard metabolic pathways. Thus, Cytoscape maps are simply used as a template to manually draw more traditional looking metabolic pathways. Since Cytoscape maps are so easily generated, we also use the software to assist in metabolite assignments. Potential lists of metabolite assignments are input into Cytoscape to identify metabolites that are isolated nodes excluded from the main network map. These metabolites are likely misassigned and are reevaluated. In addition to Cytoscape, we also use the free R statistics package (<http://www.r-project.org/>) [103] to create heat maps from absolute metabolite concentrations or percent relative concentration changes.

## 6. Conclusion

NMR metabolomics is an invaluable tool for systems biology and its application is rapidly expanding. Global changes in the metabolic state of bacterial cells occur as a result of environmental stressors, genetic modifications, drug treatments, or numerous other factors. A detailed analysis of the differences in the NMR spectra is commonly used to identify the key metabolite changes that differentiate between these bacterial classes (*e.g.*, controls versus treated). In addition, metabolite identification by NMR allows for the subsequent identification of the important metabolic pathways that are affected by the treatment, providing further insight into the underlying biological process. The appeal of NMR metabolomics is its simplicity, but unfortunately it is also easy to obtain unreliable results. The observed changes in the metabolome should be biologically relevant, but because the metabolome is so sensitive to any environmental change; it is also easily perturbed by the experimental protocol. This is clearly an undesirable outcome. To address this issue, we described in detail our optimized protocols for the NMR analysis of bacterial metabolomes. We also highlighted common problems and potential sources of mistakes. We discussed the entire process that includes growing and harvesting bacterial cells, extracting the metabolome, NMR data collection, processing and analysis, statistical analysis, metabolite and network identification. The protocols described have been successfully applied to a number of systems biology projects [5, 49, 50, 52, 104-106].

## Acknowledgements

This manuscript was supported in part by funds from the National Institute of Health (AI087668, P20 RR-17675, P30 GM103335), the American Heart Association (0860033Z), and the Nebraska Research Council. S. Halouska and R. J. Fenton were partially supported by O. Chacon's NIH grant (R21 AI087561) to standardize NMR techniques included in this publication. The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01).

## References

- [1] L.F. Shyur, N.S. Yang, *Curr. Opin. Cell Biol.*, 12 (2008) 66-71. 10.1016/j.cbpa.2008.01.032.
- [2] J.L. Spratlin, N.J. Serkova, S.G. Eckhardt, *Clin Cancer Res*, 15 (2009) 431-440. 10.1158/1078-0432.CCR-08-1059.
- [3] S. Rochford, *J. Nat. Prod.*, 68 (2005) 1813-1820. 10.1021/np050255w.
- [4] L.M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M.C. Walsh, J.A. Berden, K.M. Brindle, D.B. Kell, J.J. Rowland, H.V. Westerhoff, K. van Dam, S.G. Oliver, *Nat. Biotechnol.*, 19 (2001) 45-50. 10.1038/83496.
- [5] B. Zhang, S. Halouska, C.E. Schiaffo, M.R. Sadykov, G.A. Somerville, R. Powers, *J Proteome Res*, 10 (2011) 3743-3754. 10.1021/pr200360w.
- [6] J.C. Ewald, T. Matt, N. Zamboni, *Mol. Biosyst.*, 9 (2013) 440-446. 10.1039/c2mb25423a.
- [7] O. Fiehn, *Plant Mol. Biol.*, 48 (2002) 155-171.
- [8] R. Kaddurah-Daouk, B.S. Kristal, R.M. Weinshilboum, *Annu. Rev. Pharmacol. Toxicol.*, 48 (2008) 653-683. 10.1146/annurev.pharmtox.48.113006.094715.
- [9] H.C. Keun, *Pharmacol. Ther.*, 109 (2006) 92-106. 10.1016/j.pharmthera.2005.06.008.
- [10] W. Zhang, F. Li, L. Nie, *Microbiology*, 156 (2010) 287-301. 10.1099/mic.0.034793-0.
- [11] T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, L. Hood, *Science*, 292 (2001) 929-934. 10.1126/science.292.5518.929.
- [12] P. Pir, B. Kirdar, A. Hayes, Z.Y. Onsan, K.O. Ulgen, S.G. Oliver, *BMC Bioinf.*, 7 (2006) 203. 10.1186/1471-2105-7-203.
- [13] A.C. Mosier, N.B. Justice, B.P. Bowen, R. Baran, B.C. Thomas, T.R. Northen, J.F. Banfield, *mBio*, 4 (2013). 10.1128/mBio.00484-12.
- [14] M.R. Sadykov, B. Zhang, S. Halouska, J.L. Nelson, L.W. Kreimer, Y.F. Zhu, R. Powers, G.A. Somerville, *J Biol Chem*, 285 (2010) 36616-36624. DOI 10.1074/jbc.M110.152843.
- [15] S.H. Yoon, M.J. Han, H. Jeong, C.H. Lee, X.X. Xia, D.H. Lee, J.H. Shim, S.Y. Lee, T.K. Oh, J.F. Kim, *Genome Biol.*, 13 (2012) R37. 10.1186/gb-2012-13-5-r37.
- [16] R. Yang, Z. Du, Y. Han, L. Zhou, Y. Song, D. Zhou, Y. Cui, *Front. Cell Infect. Microbiol.*, 2 (2012) 157. 10.3389/fcimb.2012.00157.
- [17] G.P. Bisson, C. Mehaffy, C. Broeckling, J. Prenni, D. Rifat, D.S. Lun, M. Burgos, D. Weissman, P.C. Karakousis, K. Dobos, *J. Bacteriol.*, 194 (2012) 6441-6452. 10.1128/JB.01013-12.

- [18] I. Nobeli, H. Ponstingl, E.B. Krissinel, J.M. Thornton, *J. Mol. Biol.*, 334 (2003) 697-719. 10.1016/j.jmb.2003.10.008.
- [19] W. Weckwerth, *Bioanalysis*, 2 (2010) 829-836. 10.4155/bio.09.192.
- [20] R. Powers, *Magn Reson Chem*, 47 Suppl 1 (2009) S2-11. 10.1002/mrc.2461.
- [21] S. Moco, R.J. Bino, R.C.H. De Vos, J. Vervoort, *Trac-Trends in Anal. Chem.*, 26 (2007) 855-866. DOI 10.1016/j.trac.2007.08.003.
- [22] S.G. Villas-Boas, S. Mas, M. Aakesson, J. Smedsgaard, J. Nielsen, *Mass Spectrom. Rev.*, 24 (2005) 613-646. 10.1002/mas.20032.
- [23] I.D. Wilson, R. Plumb, J. Granger, H. Major, R. Williams, E.M. Lenz, *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.*, 817 (2005) 67-76. 10.1016/j.jchromb.2004.07.045.
- [24] N.V. Reo, *Drug Chem. Toxicol.*, 25 (2002) 375-382. 10.1081/dct-120014789.
- [25] T.W.M. Fan, A.N. Lane, *J. Biomol. NMR*, 49 (2011) 267-280. 10.1007/s10858-011-9484-6.
- [26] I.A. Lewis, R.H. Karsten, M.E. Norton, M. Tonelli, W.M. Westler, J.L. Markley, *Anal. Chem.*, 82 (2010) 4558-4563. 10.1021/ac100565b.
- [27] D.B. Kell, *Curr. Opin. Microbiol.*, 7 (2004) 296-307. 10.1016/j.mib.2004.04.012.
- [28] T.O. Metz, J.S. Page, E.S. Baker, K. Tang, J. Ding, Y. Shen, R.D. Smith, *TrAC, Trends Anal. Chem.*, 27 (2008) 205-214. 10.1016/j.trac.2007.11.003.
- [29] Z. Pan, D. Raftery, *Anal. Bioanal. Chem.*, 387 (2007) 525-527. 10.1007/s00216-006-0687-8.
- [30] M.R. Viant, E.S. Rosenblum, R.S. Tjeerdema, *Environ. Sci. Technol.*, 37 (2003) 4982-4989. 10.1021/es034281x.
- [31] O. Beckonert, H.C. Keun, T.M.D. Ebbels, J. Bundy, E. Holmes, J.C. Lindon, J.K. Nicholson, *Nat. Protoc.*, 2 (2007) 2692-2703. 10.1038/nprot.2007.376.
- [32] S. Dietmair, N.E. Timmins, P.P. Gray, L.K. Nielsen, J.O. Krömer, *Anal. Biochem.*, 404 (2010) 155-164. 10.1016/j.ab.2010.04.031.
- [33] H.K. Kim, Y.H. Choi, R. Verpoorte, *Nat. Protoc.*, 5 (2010) 536-549. 10.1038/nprot.2009.237.
- [34] R. Marcinowska, J. Trygg, H. Wolf-Watz, T. Mortiz, I. Surowiec, *J. Microbiol. Methods*, 87 (2011) 24-31. 10.1016/j.mimet.2011.07.001.
- [35] M. Mashego, K. Rumbold, M. De Mey, E. Vandamme, W. Soetaert, J. Heijnen, *Biotechnol. Lett.*, 29 (2007) 1-16. 10.1007/s10529-006-9218-0.
- [36] X.H. Wu, H.L. Yu, Z.Y. Ba, J.Y. Chen, H.G. Sun, B.Z. Han, *Biotechnol J*, 5 (2010) 75-84. 10.1002/biot.200900038.
- [37] M. Cuperlovic-Culf, D.A. Barnett, A.S. Culf, I. Chute, *Drug Discov. Today*, 15 (2010) 610-621. 10.1016/j.drudis.2010.06.012.
- [38] Z. Ramadan, D. Jacobs, M. Grigorov, S. Kochhar, *Talanta*, 68 (2006) 1683-1691. 10.1016/j.talanta.2005.08.042.
- [39] M. Bylesjo, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, *J. Trygg, J. Chemom.*, 20 (2006) 341-351. Doi 10.1002/Cem.1006.
- [40] A.R. Fernie, R.N. Trethewey, A.J. Krotzky, L. Willmitzer, *Nat. Rev. Mol. Cell Biol.*, 5 (2004) 763-769. Doi 10.1038/Nrm1451.
- [41] C.J. Clarke, J.N. Haselden, *Toxicol. Pathol.*, 36 (2008) 140-147. 10.1177/0192623307310947.
- [42] M. Faijes, A.E. Mars, E.J. Smid, *Microb. Cell Fact.*, 6 (2007) 27. 10.1186/1475-2859-6-27.
- [43] M.F. Van Batenburg, L. Coulier, F. van Eeuwijk, A.K. Smilde, J.A. Westerhuis, *Anal. Chem.*, 83 (2011) 3267-3274. 10.1021/ac102374c.
- [44] H. Kanani, P.K. Chrysanthopoulos, M.I. Klapa, *J. Chromatogr. B*, 871 (2008) 191-201. 10.1016/j.jchromb.2008.04.049.
- [45] O. Teahan, S. Gamble, E. Holmes, J. Waxman, J.K. Nicholson, C. Bevan, H.C. Keun, *Anal. Chem.*, 78 (2006) 4307-4318. 10.1021/ac051972y.
- [46] B. Worley, R. Powers, *Current Metabolomics*, 1 (2013) 92-107. 10.2174/2213235X130108.
- [47] C. Birkemeyer, A. Luedemann, C. Wagner, A. Erban, J. Kopka, *Trends Biotechnol.*, 23 (2005) 28-33. 10.1016/j.tibtech.2004.12.001.
- [48] T.M. Fan, L. Bandura, R. Higashi, A. Lane, *Metabolomics*, 1 (2005) 325-339. 10.1007/s11306-005-0012-0.
- [49] S. Halouska, R.J. Fenton, R.G. Barletta, R. Powers, *ACS Chem. Biol.*, 7 (2012) 166-171. 10.1021/Cb200348m.
- [50] S. Halouska, O. Chacon, R.J. Fenton, D.K. Zinniel, R.G. Barletta, R. Powers, *J. Proteome Res.*, 6 (2007) 4608-4614. 10.1021/Pr0704332.
- [51] P. Fargue, S. Halouska, M. Werth, K.M. Xu, S. Harris, R. Powers, *J. Proteome Res.*, 5 (2006) 1916-1923. 10.1021/Pr060114v.
- [52] M.R. Sadykov, M.E. Olson, S. Halouska, Y. Zhu, P.D. Fey, R. Powers, G.A. Somerville, *J. Bacteriol.*, 130 (2008) 7621-7632. 10.1128/JB.00806-08.
- [53] K. Suhre, C. Gieger, *Nat. Rev. Genet.*, 13 (2012) 759-769. 10.1038/Nrg3314.
- [54] D.K. Zinniel, R.J. Fenton, S. Halouska, R. Powers, R.G. Barletta, *J. Vis. Exp.*, 67 (2012) e3673.
- [55] S. Halouska, R. Powers, *J. Magn. Reson.*, 178 (2006) 88-95. 10.1016/j.jmr.2005.08.016.
- [56] C.J. Bolten, P. Kiefer, F. Letisse, J.C. Portais, C. Wittmann, *Anal. Chem.*, 79 (2007) 3843-3849. 10.1021/ac0623888.
- [57] N.J.C. Bailey, M. Oven, E. Holmes, M.H. Zenk, J.K. Nicholson, *Spectrosc. Int. J.*, 18 (2004) 279-287. 10.1155/2004/862164.
- [58] M. Defernez, I.J. Colquhoun, *Phytochemistry*, 62 (2003) 1009-1017. 10.1016/S0031-9422(02)00704-5.
- [59] D.I. Broadhurst, D.B. Kell, *Metabolomics*, 2 (2006) 171-196. 10.1007/s11306-006-0037-z.
- [60] R.G. Brereton, *Trac-Trends in Anal. Chem.*, 25 (2006) 1103-1111. 10.1016/j.trac.2006.10.005.
- [61] A. Canelas, C. Ras, A. ten Pierick, J. van Dam, J. Heijnen, W. van Gulik, *Metabolomics*, 4 (2008) 226-239. 10.1007/s11306-008-0116-4.
- [62] E.J. Saude, B.D. Sykes, *Metabolomics*, 3 (2007) 19-27. 10.1007/s11306-006-0042-2.
- [63] A.M. Giuliodori, C.O. Gualerzi, S. Soto, J. Vila, M.M. Tavio, *Ann. N. Y. Acad. Sci.*, 1113 (2007) 95-104. 10.1196/annals.1391.008.
- [64] F. Schadel, F. David, E. Franco-Lara, *Appl. Microbiol. Biotechnol.*, 92 (2011) 1261-1274. DOI 10.1007/s00253-011-3377-1.
- [65] M. Wellerdiek, D. Winterhoff, W. Reule, J. Brandner, M. Oldiges, *Bioprocess Biosyst. Eng.*, 32 (2009) 581-592. DOI 10.1007/s00449-008-0280-y.
- [66] C. Wittmann, J.O. Kromer, P. Kiefer, T. Binz, E. Heinzle, *Anal. Biochem.*, 327 (2004) 135-139. 10.1016/J.Ab.2004.01.002.
- [67] B. Worley, S. Halouska, R. Powers, *Anal. Biochem.*, 433 (2013)

- 102-104. 10.1016/j.ab.2012.10.011.
- [68] B.C.M. Potts, A.J. Deese, G.J. Stevens, M.D. Reily, D.G. Robertson, J. Theiss, *J. Pharm. Biomed. Anal.*, 26 (2001) 463-476. 10.1016/S0731-7085(01)00430-7.
- [69] D.I. Hoult, *J. Magn. Reson.*, 21 (1976) 337-347. 10.1016/0022-2364(76)90081-0.
- [70] V. Sklenar, M. Piotto, R. Leppik, V. Saudek, *J. Magn. Reson. Series A*, 102 (1993) 241-245. 10.1006/jmra.1993.1098.
- [71] M.L. Liu, X.A. Mao, C.H. Ye, H. Huang, J.K. Nicholson, J.C. Lindon, *J. Magn. Reson.*, 132 (1998) 125-129. 10.1006/jmre.1998.1405.
- [72] R.J. Ogg, P.B. Kingsley, J.S. Taylor, *J. Magn. Reson. Series B*, 104 (1994) 1-10. 10.1006/jmrb.1994.1048.
- [73] A.J. Simpson, S.A. Brown, *J. Magn. Reson.*, 175 (2005) 340-346. 10.1016/j.jmr.2005.05.008.
- [74] T.L. Hwang, A.J. Shaka, *J. Magn. Reson. Series A*, 112 (1995) 275-279. 10.1006/jmra.1995.1047.
- [75] J. McKenzie, A. Charlton, J. Donarski, A. MacNicoll, J. Wilson, *Metabolomics*, 6 (2010) 574-582. 10.1007/s11306-010-0226-7.
- [76] Y. Xi, J.S. de Ropp, M.R. Viant, D.L. Woodruff, P. Yu, *Anal. Chim. Acta*, 614 (2008) 127-133. 10.1016/j.aca.2008.03.024.
- [77] K.F. Hu, W.M. Westler, J.L. Markley, *J. Amer. Chem. Soc.*, 133 (2011) 1662-1665. 10.1021/Ja1095304.
- [78] K.F. Hu, J.J. Ellinger, R.A. Chylla, J.L. Markley, *Anal. Chem.*, 83 (2011) 9352-9360. 10.1021/Ac201948f.
- [79] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, *BMC genomics*, 7 (2006) 142. 10.1186/1471-2164-7-142.
- [80] A. Craig, O. Cloareo, E. Holmes, J.K. Nicholson, J.C. Lindon, *Anal. Chem.*, 78 (2006) 2262-2267. 10.1021/Ac0519312.
- [81] K. Kjeldahl, R. Bro, *J. Chemom.*, 24 (2010) 558-564. 10.1002/cem.1346.
- [82] J. Shao, *J. Am. Stat. Assoc.*, 88 (1993) 486-494. 10.1080/01621459.1993.10476299.
- [83] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell.*, 20 (2002) 269-276. 10.1016/S1093-3263(01)00123-1.
- [84] L. Eriksson, J. Trygg, S. Wold, *J. Chemom.*, 22 (2008) 594-600. 10.1002/cem.1187.
- [85] J. Westerhuis, H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven, F. van Dorsten, *Metabolomics*, 4 (2008) 81-89. 10.1007/s11306-007-0099-6.
- [86] M.T. Werth, S. Halouska, M.D. Shortridge, B. Zhang, R. Powers, *Anal. Biochem.*, 399 (2010) 58-63. 10.1016/j.ab.2009.12.022.
- [87] A.R. Henderson, *Clin. Chim. Acta*, 359 (2005) 1-26. 10.1016/j.cccn.2005.04.002.
- [88] D.S. Wishart, C. Knox, A.C. Guo, R. Eisner, N. Young, B. Gautam, D.D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J.G. Xia, L. Jia, J.A. Cruz, E. Lim, C.A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y.P. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhtudinov, L. Li, H.J. Vogel, I. Forsythe, *Nucleic Acids Res.*, 37 (2009) D603-D610. 10.1093/Nar/Gkn810.
- [89] Q. Cui, I.A. Lewis, A.D. Hegeman, M.E. Anderson, J. Li, C.F. Schulte, W.M. Westler, H.R. Eghbalnia, M.R. Sussman, J.L. Markley, *Nat. Biotechnol.*, 26 (2008) 162-164. 10.1038/Nbt0208-162.
- [90] K. Akiyama, E. Chikayama, H. Yuasa, Y. Shimada, T. Tohge, K. Shinozaki, M.Y. Hirai, T. Sakurai, J. Kikuchi, K. Saito, *In silico biology*, 8 (2008) 339-345.
- [91] E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R. Kent Wenger, H. Yao, J.L. Markley, *Nucleic Acids Res.*, 36 (2008) D402-408. 10.1093/nar/gkm957.
- [92] J. Xia, T.C. Bjorndahl, P. Tang, D.S. Wishart, *BMC Bioinf.*, 9 (2008) 507. 10.1186/1471-2105-9-507.
- [93] F. Delaglio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J. Biomol. NMR*, 6 (1995) 277-293. 10.1007/BF00197809.
- [94] R. Caspi, T. Altman, K. Dreher, C.A. Fulcher, P. Subhraveti, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, A. Pujar, A.G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, P.D. Karp, *Nucleic Acids Res.*, 40 (2012) D742-753. 10.1093/nar/gkr1014.
- [95] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi, *Nucleic Acids Res.*, 36 (2008) D480-D484. 10.1093/nar/gkm882.
- [96] M. Kohlstedt, J. Becker, C. Wittmann, *Appl. Microbiol. Biotechnol.*, 88 (2010) 1065-1075. 10.1007/s00253-010-2854-2.
- [97] J.J. Bot, M.J. Reinders, *Bioinformatics*, 27 (2011) 2451-2452. 10.1093/bioinformatics/btr388.
- [98] J. Gao, V.G. Tarcea, A. Karnovsky, B.R. Mirel, T.E. Weymouth, C.W. Beecher, J.D. Cavalcoli, B.D. Athey, G.S. Omenn, C.F. Burant, H.V. Jagadish, *Bioinformatics*, 26 (2010) 971-973. 10.1093/bioinformatics/btq048.
- [99] M. Kohl, S. Wiese, B. Warscheid, *Methods Mol. Biol.*, 696 (2011) 291-303. 10.1007/978-1-60761-987-1\_18.
- [100] M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, T. Ideker, *Bioinformatics*, 27 (2011) 431-432. 10.1093/bioinformatics/btq675.
- [101] R. Caspi, H. Foerster, C.A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S.Y. Rhee, A.G. Shearer, C. Tissier, T.C. Walk, P. Zhang, P.D. Karp, *Nucleic Acids Res.*, 36 (2008) D623-D631. 10.1093/nar/gkm900.
- [102] N.V. Chaika, T. Gebregiworgis, M.E. Lewallen, V. Purohit, P. Radhakrishnan, X. Liu, B. Zhang, K. Mehla, R.B. Brown, T. Caffrey, F. Yu, K.R. Johnson, R. Powers, M.A. Hollingsworth, P.K. Singh, *Proc Natl Acad Sci U S A*, 109 (2012) 13787-13792. 10.1073/Proc.Natl.Acad.Sci.U.S.A..1203339109.
- [103] R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2012.
- [104] A.S. Nuxoll, S. Halouska, M. Sadykov, G.A. Somerville, K.W. Bayles, T. Kielian, R. Powers, P.D. Fey, *PLoS Pathogens*, 8 (2012) e1003033. 10.1371/journal.ppat.1003033.
- [105] M.R. Sadykov, V.C. Thomas, D.D. Marshall, C.J. Wenstrom, D.E. Moormeier, T.J. Widhelm, A.S. Nuxoll, R. Powers, K.W. Bayles, *J Bacteriol*, 195 (2013) 3035-3044. 10.1128/JB.00042-13.
- [106] M.R. Sadykov, B. Zhang, S. Halouska, J.L. Nelson, L.W. Kreimer, Y. Zhu, R. Powers, G.A. Somerville, *J. Biol. Chem.*, 285 (2010) 36616-36624. 10.1074/jbc.M110.152843.